Testing the potential of Twitter mining methods for data acquisition: Evaluating novel opportunities for ecological research in multiple taxa

Adam G. Hart[1], William S. Carpenter, Estelle Hlustik-Smith, Matt Reed, Anne E. Goodenough

[1] School of Natural and Social Sciences, University of Gloucestershire, Cheltenham, UK.
ahart@glos.ac.uk

Abstract

1. Social media provides unique opportunities for data collection. Retrospective analysis of social media posts has been used in seismology, political science and public risk perception studies but has not been used extensively in ecological research. There is currently no assessment of whether such data are valid and robust in ecological contexts.

2. We used "Twitter mining" methods to search Twitter (a microblogging site) for terms relevant to three nationwide UK ecological phenomena: winged ant emergence; autumnal house spider sightings; and starling murmurations. To determine the extent to which Twitter-mined data were reliable and suitable for answering specific ecological questions the data so gathered were analysed and the results directly compared to the findings of three published studies based on primary data collected by citizen scientists during the same time period.

3. Twitter-mined data proved robust for quantifying temporal ecological patterns. There was striking similarity in the temporal patterns of winged ant emergence between previously published work and our analysis of Twitter-mined data at national scales; this was also the case for house spider sightings. Spatial data were less available but analysis of Twitter-mined data was able to replicate most spatial findings from all three studies. Baseline ecological findings, such as the sex ratio of house spider sightings, could also be replicated. Where Twitter mining was less successful was answering specific questions and testing hypotheses. Thus, we were unable to determine the influence of microhabitat on winged ants or test predation and weather hypotheses for initiation of murmuration behaviour.

4. Twitter mining clearly has great potential to generate spatiotemporal ecological data and to answer specific ecological questions. However, we found that the types and usefulness of data differed substantially between the three phenomena. Consequently, we suggest that understanding users' behaviour when posting on ecological topics would be useful if using social media is to generate ecological data.

Keywords

House spiders, Phenology, Social media, Spatial ecology, Starling mumurations, Twitter mining, Winged ants

1. Introduction

Public participation in scientific research, especially when members of the public directly assist scientists in the collection or processing of data, has become known as citizen science (Bonney et al., 2009). Citizen science is now an established research method that is increasingly well used (e.g. Cooper, Shirk, & Zuckerberg, 2014; McKinley et al., 2017; Newson et al., 2016; Pescott et al., 2015; Theobald et al., 2015). Instrumental in the increase of citizen science research has been the development of web-enabled mobile devices, especially smart phones. Such technology has allowed people to participate in projects locally, nationally and internationally in fields from astronomy (Kuchner et al., 2017) to public health (Rowbotham, McKinnon, Leach, Lamberts, & Hawe, 2017).

One field that has been particularly successful in using citizen science is ecology, especially for studies on spatiotemporal distribution where the ubiquity of the public is a major advantage (e.g. Hart, Hesselberg, Nesbit, & Goodenough, 2017). There are some challenges with using citizen science data (Lukyanenko, Parsons, & Wiersma, 2016) including the fact that: (1) data often cannot be validated; (2) data on rare species or complex ecological phenomena can be hard to obtain; (3) data can be of lower quality and consistency than those collected by experts (but see Willis et al., 2017); and (4) recording frequency can be biased towards highly populated areas or times such as weekends and holidays. Despite these issues, large datasets can be assembled across larger spatial scales and longer time periods than would otherwise be possible. Thus, citizen science is now widely used to understand species' distributions (Fournier et al., 2017), record the spread of non-native species, pests, or diseases (Parr & Sewell, 2017), monitor population trends (Dennis, Morgan, Brereton, Roy, & Fox, 2017), quantify seasonal phenological patterns (Méndez, de Jaime, & Alcántara, 2017), and better understand behaviour (Goodenough, Little, Carpenter, & Hart, 2017).

The same technological developments that have facilitated the rise of citizen science have also led to the rise of social media applications including Facebook, Twitter, Flikr, Instagram and Snapchat. The willingness with which people post on public-facing social media, and the sheer volume of such posts, makes these platforms a potential source of useful scientific data (e.g. Tufekci, 2014). Such data still make use of citizens for scientific research (and is therefore still "citizen science"), but are gathered post-hoc and indirectly, with citizens contributing, usually unknowingly, as an incidental consequence of social media activity.

Twitter, a microblogging application, is a particularly valuable tool for researchers seeking to make use of information contained in, or derived from, social media (Kumar, Morstatter, & Liu, 2014). Users of Twitter ("tweeters") can post short messages (140 characters until 2017, thence 280), termed "tweets" and reply to other users' messages with tweets being visible to users and non-users

via search engines. Each tweet has date- and time-stamps; some users also indicate location. The immediacy with which users can tweet, the apparent desire to share information, and the ability to search tweets for text or hashtags (keywords related to a topic that are proceeded by #) enables researchers to examine Twitter archives to extract information. This process is known as "Twitter mining" and has been used in a number of disciplines including seismology (Crooks, Croitoru, Stefanidis, & Radzikowski, 2013), coordinating disaster relief (Purohit et al., 2013), studying voter behaviour (Grover, Kar, Dwivedi, & Janssen, 2017), quantifying the timing of crop planting (Zipper, 2018) and evaluating public perception of ecological risks (Fellenor et al., 2017).

The immediacy of Twitter gives Twitter mining great potential for studying memorable or significant ecological phenomena. However, although this has been suggested, for example in monitoring invasive species (Daume, 2016) and studying phenology (Catlin-Groves, 2012), few ecological studies have used the approach (an exception being Fuka, Osborne-Gowey, and Fuka (2013) to map range shifts). Before widespread use of Twitter mining for ecological research is recommended, it is necessary to be confident that ecological data in this way reflect ecology rather than patterns of social media use. This validation can only be achieved if Twitter-derived data can be directly compared to data gathered on the same phenomena by other means.

In this study, we compare Twitter-derived data on ecological phenomena with primary data from published ecological studies in three taxa. The primary studies used citizen science to gather data across the UK to answer different ecological questions. The first study quantified spatiotemporal distribution and environmental triggers of ant mating flights (primarily of *Lasius niger*) (Hart et al., 2017). It included analyses of synchronicity and spatial concurrence of winged ant emergences ("flying ants") at national and regional scales. The second study assessed geographical patterns, seasonal peaks, daily rhythms and location of spiders (primarily those in the genera *Tengenaria* and *Eratigena*) within houses during the autumn (Hart, Nesbit, & Goodenough, 2018). The third study examined starling (*Sturnus vulgaris*) murmuration behaviour to assess the effect of predators and temperature (Goodenough et al., 2017). Here, for each study subject, we compare Twitter-mined data with published data to determine: (1) whether the research questions in each study could be answered robustly through Twitter mining (i.e. to determine whether tweets contain relevant information, provide a sufficient sample size and contain the necessary spatiotemporal data); and (2) for those research questions that can be answered using Twitter data, whether the findings replicate the published results. We draw our findings together to highlight the opportunities and challenges presented by Twitter mining, and offer suggestions on the use of this approach in future ecological projects.

2. Materials and methods

2.1 Ecological studies

Three published datasets based on ecological studies conducted in the UK were used for comparison purposes. The winged ant study (Hart et al., 2017) ran for three UK summers from 1st June (day 1) to 4th September (day 96) in 2012, 2013 and 2014. To find out more about the species involved an additional study in 2013 where the public also submitted samples of the winged ants they recorded (N = 436). The house spider study (Hart et al., 2018) ran across autumn and winter 2013/14 from 1st August 2013 (day 1) to 28th January 2014 (day 181). The starling murmuration study (Goodenough et al., 2017) ran across autumn and winter in 2014/15 and 2015/16 covering the period 1st October (day 1) to 31st March (day 183). See above and individual publications for more details.

2.2 Twitter mining

Data were mined from Twitter's Application Program Interface (API), which allows access to Twitter's raw data, using proprietary code commercially developed by the eponymous company "FollowtheHashtag.com". The data provided included measures of influence and suggestion of the likely gender of the Twitter user, using algorithms developed by the company. These secondary analyses were not relevant to our research and only the basic components, as described below, were used. For tweets about ants and starlings, use of hashtags generated extensive datasets. For ants, the search hashtags were #flyingants and #flyingantday; for starlings the search hashtags were #murmuration, #murmurations, #murmurating and #starlingsurvey. For house spiders, the planned hashtags (#spider, #spiders, #housespider and #housespiders) generated just 38 tweets over 2 years combined. Accordingly, the search was changed to also include any tweet including "house spider*". For each tweet, information was available on: (1) tweet content, (2) associated media (photographs, gifs or video), (3) date, (4) time, (5) twitter username, (6) tweeter biography (if available) and (7) self-declared tweeter location when available. These data were automatically parsed to create a comma separated variable file that could be loaded into Excel. Data included all tweets within the period covered by the relevant study and, for ants and spiders, the corresponding period from the preceding year (ants: 1st June–4th September 2011; spiders: 1st August 2012–28th January 2013). This enabled us to identify whether publicity surrounding the studies influenced Twitter activity

2.3 Tweet cleaning and processing

Tweets were manually processed before analysis on a tweet-bytweet basis. All tweets that did not relate to the phenomenon under consideration were removed, including non-relevant tweets (e.g. those about "The Flying Ants" band or "Murmuration Theatre"), tweets about flying ants, house spiders or starling murmurations not reporting a primary sighting, and non-UK tweets. Also, all

retweets were removed, again manually by searching for "RT" within the tweet contents. Finally, tweets from the organizations and individuals running the initially surveys (@uglosbioscience, @society_biology, @ RoyalSocBio, @AdamHartScience, @Dockling83, @RebeccaNesbit, @Natwittle) were deleted. This reduced the number of tweets from 3,009 to 2,345 for ants (77.9% retention), 11,227 to 6,218 for spiders (55.1% retention), and for starlings 1,520 to 135 (8.8% retention) (Table 1).

There was no automated geotagging information for tweets but locational data could be determined for some based on content or tweeter biographies on a tweet-by-tweet basis. If a location was specified in the tweet (e.g. "#flyingants in Cardiff") this could be manually transformed into latitude and longitude. The second method used the "home location" information present in around a third of Twitter biographies. Because there was no reason to suppose that the tweeter's home location was the location of the ecological record, this approach was only used when additional confirmatory information was given in the tweet (e.g. "amazing #murmuration from my garden" or "my house is being overrun with #housespiders"). Again, location was transformed into latitude and longitude. Table 1 details the relative use of different methods; all work was done manually.

Because tweeting is often undertaken on personal mobile devices as a rapid reaction to an ephemeral event it was assumed that tweet date (and in the case of spiders, tweet time) was closely related to that of the initial observation. Where there was evidence that the observation occurred on a different date (e.g. "fantastic #murmuration last night") an amended date was generated and used in subsequent analyses. This happened for <1% of records across the three datasets.

Data analysis was undertaken in SPSS version 24 (IBM) and Oriana Circular Statistics for Windows version 4 (Kovach Computing, Wales).

3. Results

3.1 Winged ants

A mean of 807 ± 35.1 SD tweets per year referenced winged ants mass emergences. This was reduced to 609 ± 22.9 following processing (Table 1).

3.1.1 Species identification

Species identification was determined by Hart et al. (2017) for 436 ants from samples in 2013, with 88.5% identified as *Lasius niger*.

Table 1 Sample sizes of Citizen Science (CS) studies and Twitter analyses (N= number of individual tweets)

	Flying ants		House spiders		Starling murmurations	
	CS	**Twitter**	**CS**	**Twitter**	**CS**	**Twitter**
Submitted (raw data)						
Records year -1	-	577	-	6,384	-	-
Records year 1	6,034	806	10,268	4,893	1,644	312
Records year 2	5,023	850	-	-	1,567	1194
Records year 3	4,982	766	-	-	-	-
Processed data (postcleaning)						
Records year -1	-	533	-	3,320	-	-
Records year 1	5,073	642	9,905	2,898	553	31
Records year 2	4,074	598	-	-	513	104
Records year 3	4,247	572	-	-	-	-
Location derivable postcleaning						
Records year -1	-	107 (64+43)[a]	-	697 (0+697)	-	-
Records year 1	All	85 (68+17)[a]	All	909 (0+909)	All	28 (28+0)[a]
Records year 2	All	162 (82+21)[a]	-	-	All	84 (84+0)[a]
Records year 3	All	138 (89+49)[a]	-	-	-	-

Flying ants, year 1 = 2012, House spiders year 1 = 2013 and Starling murmurations year 1 = 2014. a Numbers in parentheses denote method of establishing location (location specified in tweet content + location specified in the biography of tweeter coupled with information within the content of the tweet specifying that they were at home or close by).

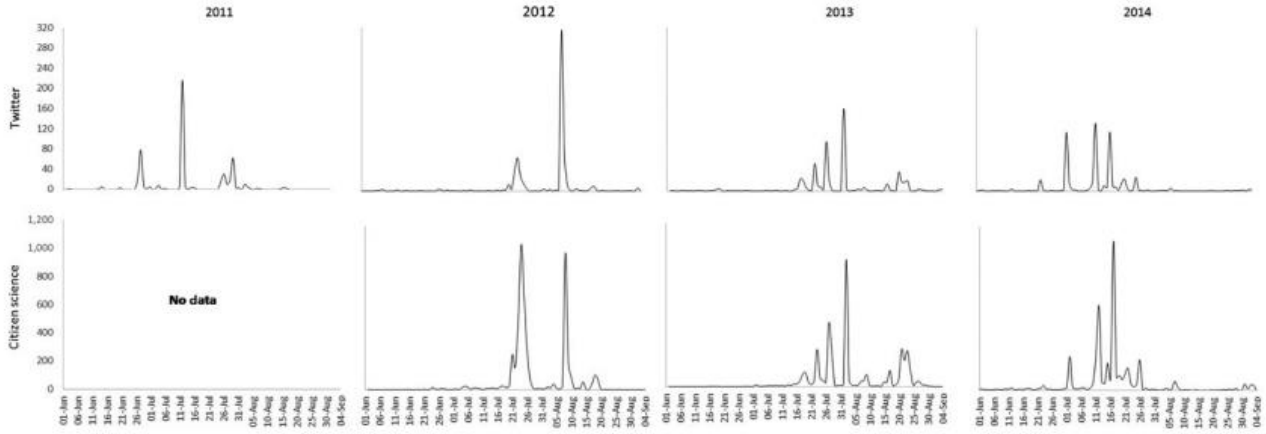

Figure 1 Data from UK-wide study of winged ant emergences derived from Twitter mentions (top row) and a citizen science study (Hart et al., 2017) (bottom row) clearly show the close similarity between methods in temporal pattern and relative number, with no significant differences in distribution for any year (see Results). The number of tweets reporting ant emergences (See Methods) per week across the UK summer is shown for years −1 to 3 (2011–2014), whereas ant emergences reported as part of the flying ant citizen science study start in year 1 (2012)

Using Twitter data from the same year, only 5 of 597 (0.8%) tweets contained video or photos that were unambiguously of winged ants and clear enough for identification to genus (*Lasius* in all cases).

3.1.2 Temporal patterns

The temporal pattern of winged ant emergences was markedly different between years (Hart et al., 2017) and this was replicated in the Twitter-derived data in this study (Figure 1). There was broad agreement between the datasets at monthly level, with 97% of sightings occurring in July and August in Hart et al. (2017) compared to 95% in the same period for Twitter-derived data. However, more striking is the remarkable agreement in the national-scale temporal patterns described in Hart et al. (2017) and those derived from corresponding Twitter data for each of the three study years (Figure 1). National-scale Twitter-derived data for 2011 (the year preceding the start of the study by Hart et al. (2017)) are generally comparable in terms of patterns and amplitude, strongly suggesting that Twitter-derived data for 2012–2014 (Years 1–3 of Hart et al. (2017)) were not confounded by social media activity related to the original study.

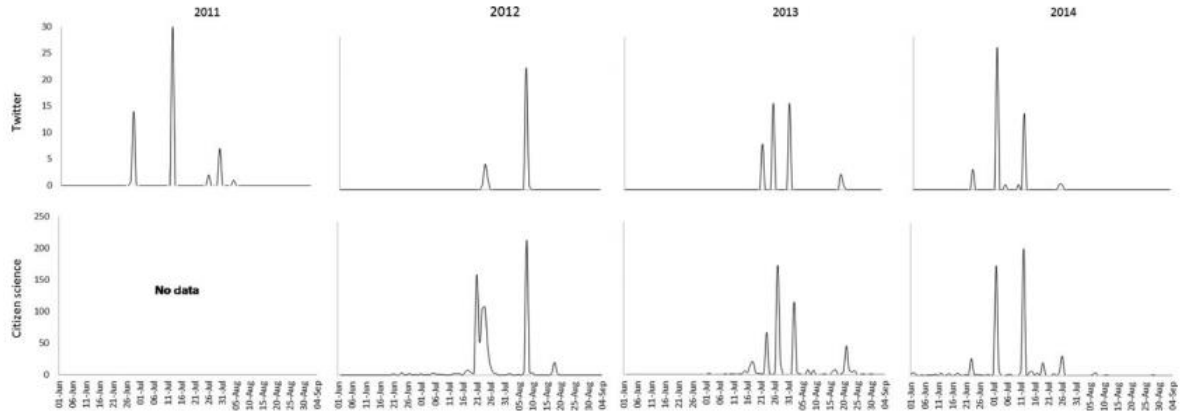

Figure 2 A subset of data taken from the Greater London region for weekly winged ant emergences derived from Twitter mentions (top row) and a UK-wide citizen science study (Hart et al., 2017) (bottom row). The number of tweets reporting ant emergences (See Methods) per week across the UK summer is shown for years −1 to 3 (2011–2014) whereas ant emergences reported as part of the flying ant citizen science study start in year 1 (2012). As with the national data (Figure 1), the close similarity between methods in temporal pattern and relative number is clearly apparent

The relatively high density of records from Greater London allowed Hart et al. (2017) to use London as a subset of data. London was a consistent subset in Hart et al. (2017) and in tweets response (Hart et al., 2017: 16.1%, 13.8% and 13.4% of all records in 2012, 2013 and 2014, respectively versus 5.1%, 7.9% and 8.7% of all tweets in the same 3 years) (overall: 1,944/13,394 = 14.5% vs. 131/1827 = 7.2%). Hart et al. (2017) analysed emergences using the subset of data relating to London, and once

again the patterns found in that study clearly agree with the patterns obtained using Twitter-derived data (Figure 2).

A two-sample Kolmogorov–Smirnov test was used to compare temporal patterns of national ant sightings based on Twitter-derived and published data for each of the three survey years. To ensure that annual data could be compared directly (the samples sizes from Twitter differed by almost an order of magnitude from published data), data were transformed to give, for each dataset, the percentage of sightings reported each week that the survey was open. There was no significant difference in the temporal patterns shown by Hart et al. (2017) and Twitter-derived data for any year for either the national-scale data (2012: $Z = 0.844$, $n_1 = 14$, $n_2 = 14$, $p = 0.415$; 2013: $Z = 0.530$, $n_1 = 14$, $n_2 = 14$, $p = 0.941$; 2014: $Z = 0.705$, $n_1 = 14$, $n_2 = 14$, $p = 0.730$) or the regional London subset (2012: $Z = 1.095$, $n_1 = 14$, $n_2 = 14$, $p = 0.181$; 2013: $Z = 0.1.278$, $n_1 = 14$, $n_2 = 14$, $p = 0.076$; 2014: $Z = 0.1.278$, $n_1 = 14$, $n_2 = 14$, $p = 0.076$).

3.1.3 Location: latitude and longitude

Hart et al. (2017) found a weak but significant northwards and westwards movement in winged ant sightings as summer progressed (i.e. latitude was positively related and longitude was negatively related to date). Year was factored into their model but separate annual analyses were not reported. For comparison purposes, we have performed annual correlations in the original dataset (Table 2) showing that latitude was significantly positively related to date for all 3 years and that longitude was significantly negatively related to date every year. An analysis of Twitter-derived data showed latitude was likewise significantly positively related to date for all 3 years, but that longitude was only significantly negatively related to date in 2014 (Table 2), the year where the effect size in the original data was the largest.

3.1.4 Spatial co-occurrence

Hart et al. (2017) demonstrated that observations of winged ants were not significantly clustered at national, regional or local (Meteorological Office weather station) scales. They did this by calculating Euclidean distances between observations on a given day and using bootstrapping to compare these distances to the mean Euclidean distance for the same number of samples randomly drawn from the relevant full dataset nationally (UK), regional (London subset) or locally (closest Meteorological Office weather station) for that year (Hart et al. (2017)). The distance between observation locations at any scale on specific days was no lower than the distance for the related comparison data points. It would be possible to undertake the same analyses on Twitter-derived data but there were insufficient data: 128 tweets per year, on average, had reliable locational data, only four more data

points than the number of weather stations (*N* = 124) used for analysis of local spatial synchrony by Hart et al. (2017).

3.1.5 Environmental triggers of ant flights

Hart et al. (2017) performed detailed analyses of the environmental triggers of ant flights by comparing weather (specifically wind, temperature and pressure) on flight days with non-flight days (3 days before and after the focal day) at the same location. As with spatial synchrony, the density of tweets with suitable locational information was too low for such analysis. Theoretically it would be possible to examine tweets for weather information but only 124 of 1827 (6.8%) tweets across the 3 years directly mentioned weather (hot* *n* = 49, warm* *n* = 8, sun* = 45, heatwave *n* = 5, humid *n* = 5, heat *n* = 12). No tweets mentioned any antonyms of weather conditions found by Hart et al. (2017) to be favourable for ant flight (cold*, cool, cloudy, windy, rain*).

Table 2 Spatial patterns in winged ant emergence in relation to latitude and longitude in the original dataset and the Twitter-derived data from the same year

		Original data					**Twitter-derived data**				
	Year	***F***	***df***	***p***	**Dir**	R^2	***F***	***df***	***p***	**Dir**	R^2
Latitude	2012	13.960	1,5071	<0.001	+	0.003	4.888	1,83	0.030	+	0.056
Latitude	2013	5.320	1,4072	0.021	+	0.001	0.186	1,160	0.018	+	0.035
Latitude	2014	243.082	1,4245	<0.001	+	0.054	11.081	1,136	0.001	+	0.075
Longitude	2012	37.711	1,5071	<0.001	-	0.007	0.090	1,83	0.764	N/A	0.001
Longitude	2013	12.759	1,4072	<0.001	-	0.003	0.879	1,160	0.350	N/A	0.005
Longitude	2014	253.746	1,4245	<0.001	-	0.056	5.841	1,136	0.017	-	0.041

3.1.6 Urban versus rural and heat-retaining structures

By asking specific questions in their survey, Hart et al. (2017) showed that urban ant nests emerged 3 days earlier than rural nests (26 July (N-7036) vs. 29 July (*N* = 5286) respectively). There was a similar difference between nests associated with heat-retaining structures such as patios and greenhouses and those that were not (25 July (*N* = 6543) versus 29 July (*N* = 5779), respectively). We were unable to investigate the urban–rural finding because there were insufficient tweets with detailed location information. We searched tweets for indications of heat-retaining structures, specifically path, patio, greenhouse, wall, deck, paving, pavement and compost. Only six tweets mentioned any of these terms compared to 6,543 records in Hart et al. (2017) that definitely mentioned the presence of such structures and 5,579 records that confirmed the absence of such structures.

3.2 Spiders

3.2.1 Temporal pattern

A two-sample Kolmogorov–Smirnov test was used to compare the temporal pattern of house spider tweets with the temporal pattern of house spider records from Hart et al. (2018). No rescaling of the house spider data to percentages was necessary as the sample sizes of the two datasets were approximately equal. There was no significant difference between the temporal distribution of recorded sightings and sightings derived from tweets (two-sample Kolmogorov–Smirnov test: Z = 1.248, n1 = 26, n2 = 26, p = 0.089) (Figure 3).

3.2.2 Time of day

Sighting times reported by Hart et al. (2018) were significantly unimodal with a pronounced peak in early evening (mean 19:35 GMT (19:25–19:45 95% CI); Rayleigh's test: $Z = 981.6$, $N = 9{,}807$, $p < 0.001$). Sighting times derived from the time that a tweet was posted and were again significantly unimodal with a pronounced peak in early evening (mean 21:02 GMT (20:47–21:17 95% CI); Rayleigh's test: $Z = 410.7$, $N = 2{,}898$ $p < 0.001$) (Figure 4). There was a statistically significant difference in the circular (time) distributions between the datasets (Mardia–Watson–Wheeler test: $W = 97.687$, $n_1 = 9807$, $n_2 = 2{,}898$, $p < 0.001$) with the Twitter data yielding a significantly later time.

3.2.3 Latitude and longitude

Hart et al. (2018) found a statistically significant but weak effect of latitude and longitude on spider phenology with sightings moving northwards and westwards through the autumn; a similar effect was found here using Twitter-derived data (Spearman rank correlation for latitude: $r_s = 0.067$, $n = 1{,}606$, $p = 0.008$; longitude: $r_s = -0.070$, $n = 1{,}606$, $p = 0.005$). The r^2 values estimated from Pearson were 0.004 and 0.002 for latitude and longitude, respectively, versus 0.076 and 0.027 in Hart et al. (2018).

3.2.4 Sex ratio

The relative ease with which larger spiders can be sexed allowed Hart et al. (2018) to ask respondents specifically to provide sex information for recorded sightings. They found 3,795 male spiders (82.3%) and 818 female spiders (17.7%), giving a highly significant male-skewed sex ratio for spiders recorded in residential homes. Of tweets that reported sex of observed spiders, 43 reported males (75.4%) and 14 reported females (24.6%), again giving a significant male-bias (chi-square goodness-of-fit test: $\chi^2 = 14.754$, $df = 1$, $p = 0.0001$). A chi-square test for association between Hart et al. (2018) data and Twitter-derived data was not significant ($\chi^2 = 1.793$ $df = 1$, $p = 0.181$), and thus the sex ratio does not differ between the datasets.

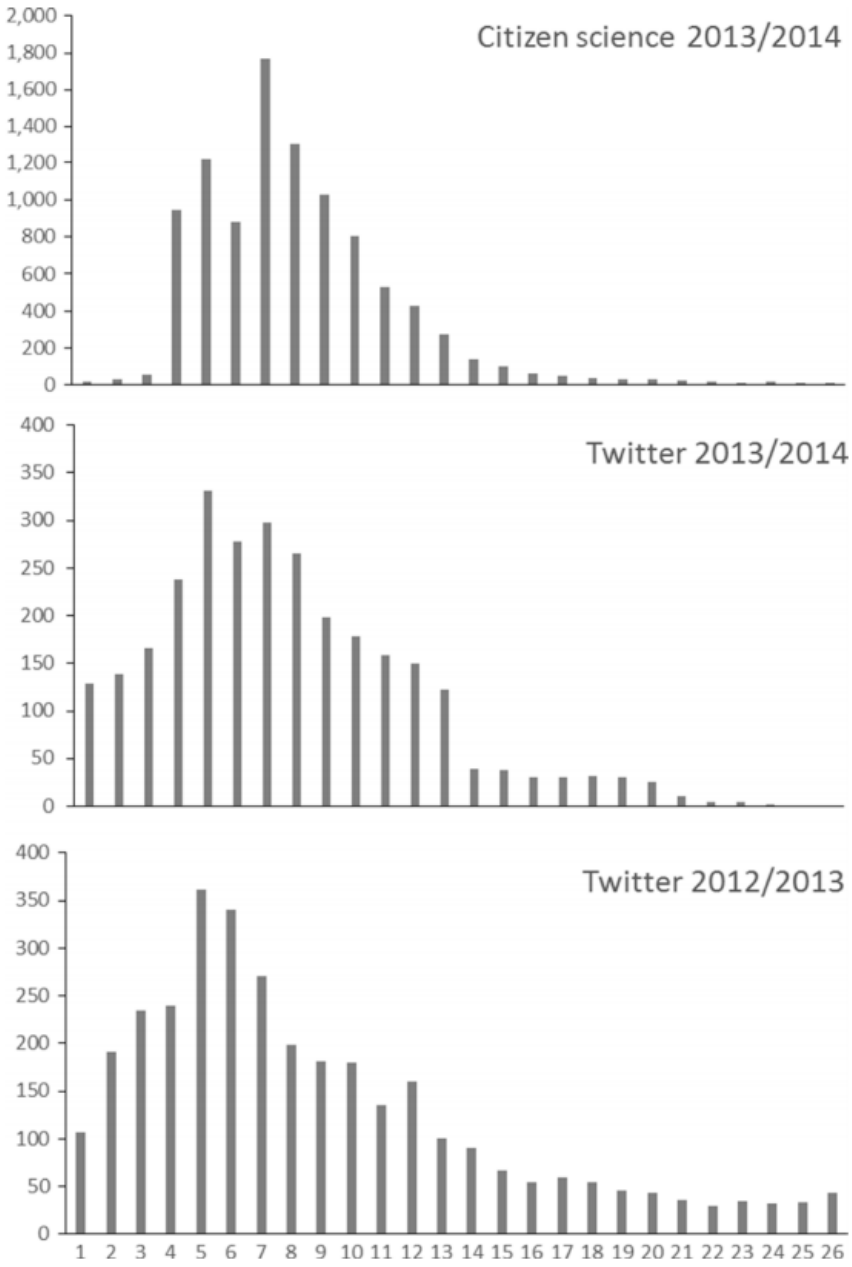

Figure 3 Sightings of house spiders (likely *Tegenaria* and *Eratigena* were reported via a citizen science study (Hart et al., 2018) (top) and derived from mentions within in tweets for the same year (middle) and the previous year (bottom). Week 1 begins on August 1st. The citizen science data and Twitter data are not significantly different (see Results)

3.2.5 Location within the house

Hart et al. (2018) asked respondents about where in the house their spider was seen made and 8,241/9,905 respondents (83.2%) provided this information. Useable room information derivable from tweets was far lower but was available for 417/2,898 tweets (14.4%). The top five rooms in Hart et al. (2018) were, in descending order, living room, bathroom, bedroom, hallway/stairs and kitchen. Using Twitter-derived data, the top five rooms, in descending order, were bedroom,

bathroom, living room, hallway/stairs, and kitchen. The distribution of sightings was significantly different (chi- square test for association: $\chi^2 = 130$, $df = 8$, $p < 0.001$), a pattern driven by the higher frequency of sightings in bedrooms and lower frequency of sightings in living rooms for Twitter-derived data.

Hart et al. (2018) also asked respondents about the location of reported spiders within a room (options were wall, floor, ceiling, furniture, door/window and sink/bath). This gave 7,789 records from 9,905 (78.6%) that could be analysed. Useable information derivable from tweets gave 265 records from 2,898 (9.1%). The order of locations, in declining order was floor, wall, ceiling, sink, door/window, furniture (Hart et al., 2018) and furniture, floor, wall, door/window, ceiling, sink (this study). This difference was significant (chi-square test for association: $\chi^2 = 1720$, $df = 5$, $p < 0.001$) and was driven by the higher proportion of furniture-related sightings associated with Twitter-derived data (this was the most reported location in this study and only the fifth highest reported location in Hart et al. (2018)).

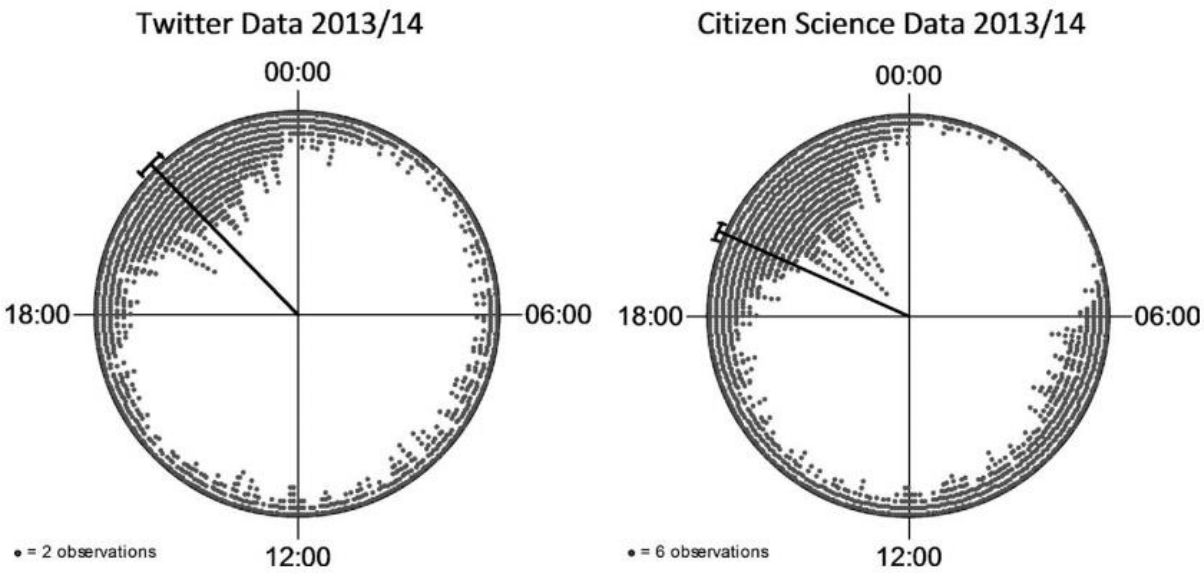

Figure 4 The time of day of sightings of house spiders (likely *Tegenaria* and *Eratigena*) given by participants in a citizen science study (Hart et al., 2017) (left) and derived from tweets. There was a statistically significant difference in the circular (time) distributions between the datasets (see Results)

3.3 Starlings

The number of tweets referencing starling murmurations was considerably lower ($N = 135$) than the number of records obtained by Goodenough et al. (2017) ($N = 1,066$). In contrast to tweets on the other taxa, geographical location was almost always given in murmuration tweets (2014/15 = 90.3%; 2015/16 = 80.8%). Accordingly, it was possible to map murmuration sightings from both original

records and Twitter-derived data (Figure 5). The main spatial patterns in the large dataset reported in Goodenough et al. (2017) were also present in the Twitter-derived data. Specifically, key murmuration hotspots (including Blackpool, Aberystwyth, Brighton, the Somerset levels and East Anglia) were identified in both datasets as were the limited sightings in Scotland (probably reflecting a lack of people recording murmurations rather than an absence of the phenomenon).

Murmuration size and duration, were important components of Goodenough et al.'s (2017) analysis and compulsory questions in that study. Conversely, murmuration size, was mentioned in just nine tweets across 2 years (6.7%), while duration was only mentioned in one tweet. Given this lack of data, it was not possible to test the relationship between size and duration, nor the spatiotemporal patterns in these parameters, using Twitter-derived data to replicate Goodenough et al.'s (2017) analyses.

3.3.1 Predator presence and temperature

To test whether predator presence or temperature influenced murmuration size or duration, Goodenough et al. (2017) asked respondents to record these variables during murmuration events. Just five tweets mentioned birds of prey (specifically sparrowhawk *Accipiter nisus* (*N* = 3), peregrine *Falco peregrinus* (*N* = 1) and kestrel *Falco tinnunculus* (*N* = 1)), a total of 3.7% of records versus 29.6% in Goodenough et al. (2017). Only one tweet specifically mentioned the absence of birds of prey. Temperature information was never given. The lack of information on predators and temperature (and murmuration size/duration) meant that no analyses could be undertaken to examine relationships between these variables.

4. Discussion

It has been suggested previously that Twitter-derived data has potential for ecological research (e.g. Daume, 2016), but this has not been extensively tested. Here, by comparing Twitter-derived data with published datasets on three ecological phenomena we have, for the first time, been able to provide a robust analysis of the value of Twitter mining in ecology. The comparator datasets were gathered using citizen science studies so we are in effect comparing primary, direct citizen science data (collected during nationwide citizen science campaigns) with secondary, indirect citizen science data (collected from Twitter). Citizen science is proving to be a reliable technique for gathering data across large spatial scales (e.g. Cooper et al., 2014) but is not without shortcomings (data are rarely validated, for example) (Lukyanenko et al., 2016). However, the citizen science studies used here provide the only data with which we can meaningfully compare nationwide Twitter-derived data, and the ecological insights derived from these studies have been published previously. If we are

willing to accept that citizen science provides useful ecological data, then using such studies as the basis for a cautious comparison of methods is an acceptable approach. With this caveat in mind, we found that such data can be used successfully but there some important limitations.

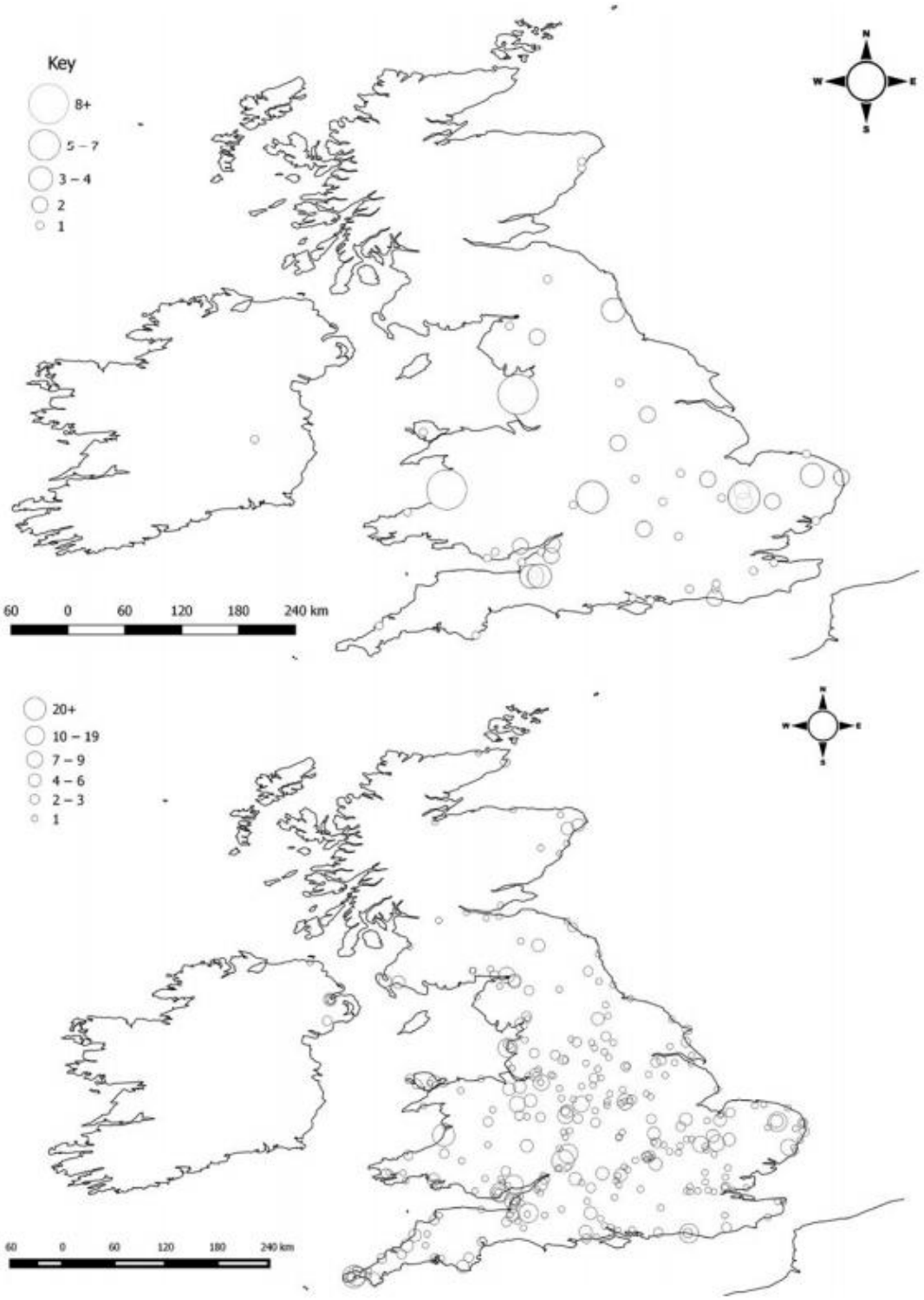

Figure 5 Map of starling murmurations based on Goodenough et al. (2017) (top) and Twitter-derived data (bottom); the size of the symbol denotes the number of records from that location

Central to the winged ants and house spider focal comparator studies were temporal aspects of the phenomena and we were able to replicate the main temporal findings of both studies using Twitter data. Indeed, the complex pattern of peaks and troughs in ant emergence and the autumnal peak in spider sightings within homes as reported on Twitter were so similar to the published data that Twitter mining would have yielded conclusions identical to those in the published dataset. This was also the case for winged ant emergences from the London subset, indicating that Twitter data can be robust at sub-national scales provided that there are sufficient tweets. It is possible that people could both be tweeting about a phenomenon and recording identical data in the citizen science studies, leading to a form of pseudoreplication. We doubt that this is a substantial effect since in cases where we had tweets from the year before the citizen science study, there was no pronounced increase in the subsequent year. In any case, anonymity (in the case of the citizen science studies) and identify ambiguity (Twitter names are not necessarily relatable to an individual) means we were unable to investigate this further.

For both ants and spiders, it is perhaps the immediacy of Twitter, the "urgency" of the phenomena in question and the desire to connect to other users (Chen, 2011) that contributes to this success. The emergence of winged ants is popular in the media (e.g. Vulliamy, 2017) and frequently evokes an emotional response from tweeters (e.g. "#Flyingants have taken over London today!"). Likewise, the annual appearance of house spiders has garnered considerable media attention (e.g. Duell, 2017). Many people have negative attitude to spiders in their homes and often share spider sightings on this basis (a typical tweet being "Just found a MASSIVE house spider in my bedroom, I'm too scared to sleep now"). Consequently, tweets are generally posted on the same day as the event. At a finer scale, the time of spider sightings using Twitter-derived data yielded a significantly later mean time than that reported in Hart et al. (2018). We suspect that while tweets may be posted soon after the actual event, they are not always posted immediately. Thus, asking people to report an exact time retrospectively is a more accurate measure of the timing than using tweet posting time, at least for this phenomenon.

All tweets have an automatic date/time stamp and temporal data do not rely on additional user input. Conversely, spatial data were not, at the time of the study, automatically available. Other social media sites such as Flickr, which geotag uploaded images, have been used in studies of species distributions using either a keyword tagging or mining tag approach (El Qadi et al., 2017; Stafford et al., 2010). It is possible to use the World Wide Web Consortium (W3C) Geolocation API to enable device information from web browsers of mobile phones or laptops (Doty & Wilde, 2010) to be accessed but this was not available in this retrospective study. More recently, Twitter has launched the option of having latitude and longitude automatically added to tweets via "share precise

location" in version 6.26.0. This only became live on 9 December 2016 and does not apply to retrospective mining, such as that used here, but might open new research avenues in the future provided sufficient users opt in.

Despite the above limitations, we were able to replicate most baseline spatial and spatiotemporal findings of the comparator studies by trawling tweet content and tweeter biography to determine likely location. Like Hart et al. (2017) we found latitude was significantly positively related to date of winged ant emergence (i.e. emergences move northwards) for all 3 years using Twitter data and significantly negatively related to longitude for 1 year (2014) but we were unable to replicate their findings for longitude for the other 2 years where original effect sizes were much smaller. This suggests that weak trends might be harder to quantify using Twitter, possibly because of the smaller sample sizes involved or the courser spatial scale inherent in determining location indirectly. We were able to replicate both the latitude and longitude relationships for spider sightings found by Hart et al. (2018), probably because of the much higher sample size. However, although baseline spatial findings could be replicated, more complex analyses were not possible. For example, the paucity of spatial data precluded replication of the spatial clustering analyses or modelling environmental triggers of ant flights, which were major components of the original study.

In stark contrast to ant and spider tweets, location information was provided in ca. 90% of tweets relating to starling murmurations. Given the fact that murmurations often become hotspots for people wanting to view them, and thus location is relevant to both the tweeter and the followers, it is not perhaps surprising that most tweets gave locational information. We were able to replicate the broad spatial occurrence map of Goodenough et al. (2017) and to identify a number of the same key locations. However, our ability to further analyse starling murmurations spatially was hampered by the relatively small number of people tweeting about them. There is no a priori reason to assume that murmurations are less "tweet worthy" than house spiders but it is possible that many of those visiting murmurations are not typical tweeters. Delving in to the motivations behind tweeting (e.g. Toubia & Stephen, 2013) is beyond the scope of this paper and has not been carried out for ecological phenomena but our findings throughout strongly suggest that it would be a sensible approach if Twitter mining is to be used for ecology research. It might be, for example, that there is considerable bias in tweeting about ecological phenomena perceived negatively or that tweeters are more prone to exaggerate or embellish reports compared to those respondents motivated to fill in details in a bespoke citizen science campaign.

The ant and spider studies were primarily spatiotemporal explorations but the starling murmuration study tested two causal hypotheses viz. the "safer together" hypothesis (murmurations protect

against predators) and the “warmer together” hypothesis (murmurations advertise roost location). Goodenough et al. (2017) were able to support the former because they had specifically requested data on murmuration size, duration, predator presence and temperature. Such information was rarely recorded on tweets, probably because the tweeter’s motivations for tweeting and the restricted character count did not encourage sharing this ecologically useful information (also found by Fuka et al., 2013), and it was thus not possible to replicate their findings.

All three comparator studies had additional idiosyncratic ecological findings that we were not always successful in replicating. The winged ant study (Hart et al., 2017) was able to identify the species for a substantial subset of 1 year’s records. Tweets did not provide species identifications although uploaded media allowed identification in a few cases. Hart et al. (2017) also found that ant nests in urban areas or associated with heat-retaining structures emerged earlier but few tweets provided relevant information and so we were unable to replicate these findings. Neither Twitter data nor the comparison citizen science data were able to identify spiders to species, although comparison with other studies led Hart, Nesbit and Goodenough to conclude that their data were likely reflecting the ecology of *Tengenaria* and *Eratigena*, the spiders most commonly referred to as house spiders. The sex ratio of house spiders (Hart et al., 2018) was a result we were able to replicate with Twitter data even though the number of tweets mentioning sex was relatively low (N = 57). The clear sexual dimorphism in many of the larger spider species possibly contributed to tweeters including that information on their tweets. Hart et al. (2018) analysed the rooms in which spiders were sighted position in those rooms and, perhaps because of the immediate relevance of this finer-scale locational information (e.g. “I’ve just seen a giant spider on the floor of my bathroom!”), 14.4% of tweets reported room location and 9.1% reported location within room. Comparative analysis, however, showed a difference in the distribution between and within rooms relative to the original dataset. This might be due to the difference between the objective report of a sighting and a tweet. Twitter is not an objective reporting medium and the emotional response of seeing a spider somewhere close to sleeping or bathing areas might be sufficient to tip tweeters over a “tweeting threshold”. As discussed above, the motivation and behaviour of people on social media are likely to affect when and what people post and, in turn, the availability and reliability of Twitter-derived information.

We conclude that Twitter mining has great potential for providing data on ecological phenomena, especially temporal data and, for some phenomena, spatial data. This is especially true for phenomena that have a specific date of occurrence (e.g. winged ant emergence) or a specific location (e.g. a murmuration site). However, while broad-scale spatial patterns can be identified at both national and sub-national levels, low sample sizes can preclude detailed analysis of spatial data

in relation to, for example, environmental parameters or temporal data, especially when effect sizes are small. Tweets were very much less successful in gathering data needed for testing specific hypotheses or answering specific question (e.g. abiotic cues for ant emergence, biotic cues for murmuration behaviour) simply because so few tweeters provided the necessary (possibly rather obscure) biological detail. Overall, we conclude that Twitter mining, and likely other social media data mining, is a form of indirect citizen science that represents a real opportunity for ecological research, especially for phenological studies of relatively charismatic events and species, provided that the pattern of Twitter usage is well-matched to the questions and phenomena of interest.

Acknowledgements

A.G.H. and A.E.G. conceived the idea and designed methodology; M.R. oversaw the collection the Twitter data; E.H.S. helped with cleaning the Twitter data, A.G.H. and A.E.G. analysed the data; W.S.C. produced the maps, A.G.H. and A.E.G. led wrote the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Data accessibility

Data used for this study are archived in the University of Gloucestershire Repository http://eprints.glos.ac.uk/id/eprint/5772.

ORCID

Adam G. Hart http://orcid.org/0000-0002-4795-9986

References

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen science: A developing tool for expanding science knowledge and scientific literacy. BioScience, 59, 977–984. https://doi.org/10.1525/bio.2009.59.11.9

Catlin-Groves, C. L. (2012). The citizen science landscape: From volunteers to citizen sensors and beyond. International Journal of Zoology, 2012. https://doi.org/10.1155/2012/349630

Chen, G. M. (2011). Tweet this: A uses and gratifications perspective on how active Twitter use gratifies a need to connect with others. Computers in Human Behavior, 27(2), 755–762. https://doi. org/10.1016/j.chb.2010.10.023

Cooper, C. B., Shirk, J., & Zuckerberg, B. (2014). The invisible prevalence of citizen science in global research: Migratory birds and climate change. PLoS ONE, 9, e106508. https://doi.org/10.1371/journal. pone.0106508

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a distributed sensor system. Transactions in GIS, 17(1), 124–147. https://doi. org/10.1111/j.1467-9671.2012.01359.x

Daume, S. (2016). Mining twitter to monitor invasive alien species—an analytical framework and sample information topologies. Ecological Informatics, 31, 70–82. https://doi.org/10.1016/j.ecoinf.2015.11.014

Dennis, E. B., Morgan, B. J., Brereton, T. M., Roy, D. B., & Fox, R. (2017). Using citizen science butterfly counts to predict species population trends. Conservation Biology, 31, 1350–1361. https://doi. org/10.1111/cobi.12956

Doty, N., & Wilde, E. (2010). Geolocation privacy and application platforms. In Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS (pp. 65–69) ACM.

Duell, M. (2017). Wet weather sparks huge spider invasion in homes across UK as creatures rush inside for warmth and shelter. The Daily Mail, 25th August http://www.dailymail.co.uk/news/article-4822640/UK-weather-Bank-Holiday-weekend-SPIDERinvasion.html

El Qadi, M. M., Dorin, A., Dyer, A., Burd, M., Bukovac, Z., & Shrestha, M. (2017). Mapping species distributions with social media geo-tagged images: Case studies of bees and flowering plants in Australia. Ecological informatics, 39, 23–31.

Fellenor, J., Barnett, J., Potter, C., Urquhart, J., Mumford, J. D., & Quine, C. P. (2017). The social amplification of risk on Twitter: The case of ash dieback disease in the United Kingdom. Journal of Risk Research, 1–21. https://doi.org/10.1080/13669877.2017.1281339

Fournier, A., Sullivan, A. R., Bump, J. K., Perkins, M., Shieldcastle, M. C., & King, S. L. (2017). Combining citizen science species distribution models and stable isotopes reveals migratory connectivity in the secretive Virginia rail. Journal of Applied Ecology, 54, 618–627. https:// doi.org/10.1111/1365-2664.12723

Fuka, M. Z., Osborne-Gowey, J. D., & Fuka, D. R. (2013). Shifting species ranges and changing phenology: A new approach to mining social media for ecosystems observations. In AGU Fall Meeting Abstracts.

Goodenough, A. E., Little, N., Carpenter, W. S., & Hart, A. G. (2017). Birds of a feather flock together: Insights into starling murmuration behaviour revealed using citizen science. PLoS ONE, 12, e0179277. https://doi.org/10.1371/journal.pone.0179277

Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2017). The untold story of USA presidential elections in 2016-insights from twitter analytics. In A. K. Kar, P. Vigneswara Ilavarasan, M. P. Gupta, Y. K. Dwivedi , M. Mäntymäki, M. Janssen, A. Simintiras, & S. Al-Sharhan, (Eds.), Conference on e-Business, e-Services and e-Society (pp. 339– 350).

Cham, Switzerland: Springer. Hart, A. G., Hesselberg, T., Nesbit, R., & Goodenough, A. E. (2017). The spatial distribution and environmental triggers of ant mating flights: Using citizen-science data to reveal national patterns. Ecography, 40, https://doi.org/10.1111/ecog.03140

Hart, A. G., Nesbit, R., & Goodenough, A. E. (2018). Spatiotemporal variation in house spider phenology at a national scale using citizen science. Arachnology, 17, 331–334. https://doi.org/10.13156/ arac.2017.17.7.331

Kuchner, M. J., Faherty, J. K., Schneider, A. C., Meisner, A. M., Filippazzo, J. C., Gagné, J., … Mokaev, K. (2017). The first brown dwarf discovered by the backyard worlds: Planet 9 citizen science project. The Astrophysical Journal Letters, 841(2), L19. https://doi. org/10.3847/2041-8213/aa7200

Kumar, S., Morstatter, F., & Liu, H. (2014). Twitter data analytics. New York: Springer. https://doi.org/10.1007/978-1-4614-9372-3

Lukyanenko, R., Parsons, J., & Wiersma, Y. F. (2016). Emerging problems of data quality in citizen science. Conservation Biology, 30, 447–449. https://doi.org/10.1111/cobi.12706

McKinley, D. C., Miller-Rushing, A. J., Ballard, H. L., Bonney, R., Brown, H., Cook-Patton, S. C., … Ryan, S. F. (2017). Citizen science can improve conservation science, natural resource management, and environmental protection. Biological Conservation, 208, 15–28. https:// doi.org/10.1016/j.biocon.2016.05.015

Newson, S. E., Moran, N. J., Musgrove, A. J., Pearce-Higgins, J. W., Gillings, S., Atkinson, P. W., … Baillie, S. R. (2016). Long-term changes in the migration phenology of UK breeding birds detected by largescale citizen science recording schemes. Ibis, 158, 481–495. https:// doi.org/10.1111/ibi.12367

Parr, J., & Sewell, J. (2017). Citizen sentinels: The role of citizen scientists in reporting and monitoring invasive non-native species. In J. A. Cigliano, & H. L. Ballard (Eds.), Citizen Science for Coastal and Marine Conservation (pp. 59–76). London, UK: Routledge.

Pescott, O. L., Walker, K. J., Pocock, M. J., Jitlal, M., Outhwaite, C. L., Cheffings, C. M., … Roy, D. B. (2015). Ecological monitoring with citizen science: The design and implementation of schemes for recording plants in Britain and Ireland. Biological Journal of the Linnean Society, 115, 505–521. https://doi.org/10.1111/bij.12581

Purohit, H., Hampton, A., Shalin, V. L., Sheth, A. P., Flach, J., & Bhatt, S. (2013). What kind of# conversation is Twitter? Mining# psycholinguistic cues for emergency coordination. Computers in Human Behavior, 29, 2438–2447. https://doi.org/10.1016/j.chb.2013.05.007

Rowbotham, S., McKinnon, M., Leach, J., Lamberts, R., & Hawe, P. (2017). Does citizen science have the capacity to transform population health science? Critical Public Health, 1–11. https://doi.org/10.1080 /09581596.2017.1395393

Stafford, R., Hart, A. G., Collins, L., Kirkhope, C. L., Williams, R. L., Rees, S. G., … Goodenough, A. E. (2010). Eu-social science: The role of internet social networks in the collection of bee biodiversity data. PLoS ONE, 5, e14381. https://doi.org/10.1371/journal.pone.0014381

Theobald, E. J., Ettinger, A. K., Burgess, H. K., DeBey, L. B., Schmidt, N. R., Froehlich, H. E., … Parrish, J. K. (2015). Global change and local solutions: Tapping the unrealized potential of citizen science for biodiversity research. Biological Conservation, 181, 236–244. https://doi. org/10.1016/j.biocon.2014.10.021

Toubia, O., & Stephen, A. T. (2013). Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter? Marketing Science, 32, 368–392. https://doi.org/10.1287/mksc.2013.0773

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. ICWSM, 14, 505–514.

Vulliamy, E. (2017). Flying ant day 2017: When is it? What is it? Everything you need to know. The Independent July 5th https://www.independent.co.uk/news/science/flying-ant-day-2017-when-is-it-what-is-ituk-ants-infestation-a7824186.html

Willis, C. G., Law, E., Williams, A. C., Franzone, B. F., Bernardos, R., Bruno, L., … Davis, C. C. (2017). CrowdCurio: An online crowdsourcing platform to facilitate climate change studies using herbarium specimens. New Phytologist, 215, 479–488. https://doi.org/10.1111/nph.14535

Zipper, S. C. (2018). Agricultural research using social media data. Agronomy Journal, 110, 349–358. https://doi.org/10.2134/ agronj2017.08.0495