Range-shift of a European moth into urban habitats is detected by social media data but not traditional monitoring

Nile Stephenson[1], Nathalie Pettorelli[2], and Regan Early[3]

[1]University of Cambridge
[2]Institute of Zoology
[3]University of Exeter

July 16, 2024

Abstract

As the world's climate changes, species are undergoing range shifts. Range shifts are generally documented using databases such as the Global Biodiversity Information Facility (GBIF), which largely contain data from monitoring schemes and wildlife surveys. Such databases have two major limitations: i) data may be spatially biased because traditionally surveyed areas are in rural habitats, ii) there is a time lag between data collection and assimilation into GBIF, which means rapid range shifts cannot be tracked. Alternative data sources, such as social media, could provide information on species distributions and range shifts that compensate for spatial biases in GBIF records because social media data may be collected outside traditional surveyed areas. Such data are also usually shared online immediately after a wildlife sighting. The complementarity of GBIF and social media data, however, has rarely been assessed, particularly when tracking range shifts. Despite their potential utility, social media data may be particularly prone to temporary trends or geographic variation in behaviour that are not understood. We lack tools with which to counter these biases. To address these knowledge gaps, we compare the habitat usage revealed by biological records of the Jersey tiger moth from GBIF and from multiple social media data sources (Instagram, iNaturalist, and Flickr). We develop a novel method to account for recorder bias in social media data. We find that biological records from iNaturalist and Instagram reveal greater than expected occurrence in urban environments, greatly affecting the accuracy of habitat suitability models. We also develop a method for comparing recorder effort between multiple data sources. Recorder effort differs notably between data sources, and Instagram complements GBIF by recording species in areas unaccounted for by GBIF. By incorporating recorder effort metrics, data from social media sources could be used to improve monitoring of range shifting species in urban spaces.

Introduction

Species around the globe are redistributing in response to anthropogenic climate change[1–3]. Range shifting species illicit positive[4] and negative[5,6] ecological and societal impacts[7], thus there is a need to track range shifts. Tracking range shifts requires large, high-quality occurrence datasets, such as those provided by online databases like the Global Biodiversity Information Facility (GBIF)[8,52]. While GBIF collates occurrence data from a range of sources, the majority of data originate from scientific surveys[9]. The vast majority of scientific surveys occur in a species' "natural" habitat – where a species is historically likely to be found - which may bias occurrence records from databases such as GBIF towards rural locations. However, recent studies report that urban environments are important to range shifting species; many range shifters have been found to be human-associated, often occurring in gardens or unintentionally transported into cities as passengers on trade vessels[10,11]. Therefore, the possibility of relatively urban environments being under-represented in databases such as GBIF may cause a gap within occurrence data records for range shifters. Detecting rapidly and monitoring arrivals in human-dominated landscapes such as urbanised areas may therefore reduce spatial bias in predictive models and inform the association between range-shifting species and urban habitats.

Another challenge for sourcing data on range shifts is that many resources such as GBIF have a time lag (up to 3 years) associated with the process of recording, verification, and agglomeration of occurrence data[12,13]. However, the speed and magnitude of range shifts necessitates more rapid data availability[14,15].

One potential solution could be the implementation of community science projects, which have been shown to produce high quality occurrence data quickly[16–18]. However, community science projects often require vast resource expenditure and many willing participants[18]. Another potential avenue to gather occurrence data quickly within a variety of environments is via social media[19]. Social media users may upload georeferenced photographs of a species of interest incidentally[19]. Photos of a focal species are often uploaded to social media immediately, expediting the process of gathering data. Furthermore, because the majority of humans reside in urban environments, and urban environments benefit from a good internet connection, it is likely that social media will survey these environments. Social media sources may reveal use of urban habitat overlooked within traditional surveying methods that target rural areas[20].

Despite the advantages above, social media data could also be patchy and prone to a higher degree of spatial recorder bias than traditional ecological data. Heterogenous recorder effort can cause over- and under-estimation of suitability of particular environmental conditions in Habitat Suitability Models (HSMs). Patchiness could be due to hotspots of social media use within highly urbanised areas and users may be heavily influenced by trends, leading to a period of intense interest in a small number of species[21]. It is therefore particularly important to understand the role of spatial and temporal recorder effort bias in social media data. There may also be variations in spatial bias and the influence of trends between different social media platforms, so we need to understand how recorder effort differs between platforms.

In this study, we compare the information content provided by different sources of occurrence data of a range shifting species, the Jersey tiger moth (JTM), *Euplagia quadripuncteria* (formerly *Callimorpha quadripuncteria*). JTM is a day-flying, recognisable, abundant lepidopteran currently undergoing rapid range shifts due to climate change[22]. JTM is a generalist species, likely to be able to make use of urban environments[23], and is also visually striking, therefore potentially generating interest on social media platforms. We: i) model annual habitat suitability for JTM in a portion of Europe during a period of changing climate using data from GBIF; ii) assess whether occurrences of JTM from four social media data sources (Twitter, Flickr, Instagram, and iNaturalist) are found in areas that GBIF models predict to have poor habitat suitability; and iii) investigate how recorder effort affects JTM occurrence across all data sources. We predict that: i) occurrence data from social media platforms are found in areas that models based on GBIF data would predict to be of low habitat suitability; and ii) accounting for recorder effort will be particularly important for the modelling of species distribution using social media data.

Materials and methods

Data Collection

Occurrence data were collected from five sources: GBIF, iNaturalist, Twitter, Instagram, and Flickr. iNaturalist, Twitter, Instagram, and Flickr were selected because biological records could be extracted with relative ease. Records from each source were collected from between 2000 and 2018 as these were the years where comparable environmental data could be gathered and where JTM had been sufficiently sampled ($>$ 50 occurrences per year) from GBIF across the selected study region.

The study region included the UK, Republic of Ireland, France, Belgium, the Netherlands, Luxembourg, Switzerland, Czech Republic, Austria, Germany, Denmark, and Italy (Fig. 1). This region represents a large proportion of the known distribution of JTM and includes nations from which biological records were reported to GBIF throughout 2000 – 2018. Although the region does not encompass the hottest component of the species' climate niche (as records in this region were too sparse), this should not affect predictions of habitat suitability for the range shift of JTM into the UK, where conditions are cooler.

Data from iNaturalist were removed from the GBIF dataset to avoid any duplication within the two datasets. Search terms (Table 1) were applied for Twitter, Instagram, and Flickr to both original posts and their subse-

quent comments to account for individuals who were unable to identify JTM and were seeking identification. We only used occurrences derived from posts and tweets that included an image of adult JTMs. Duplicates from social media data arising from people sharing the same information on different platforms were removed. All occurrence data derived from social media platforms were manually checked to ensure that identification of adult JTM was correct. Only occurrences that fell within months of the year when JTM adults fly were retained in the instances of Flickr, Twitter, and Instagram. All larval records were removed from our GBIF dataset. We only obtained 16 records from Twitter, so we did not include this data source in in any further analyses. For data from Flickr, georeferences were automatically extracted using a custom script, but for Instagram and Twitter, georeferenced data was manually collected from individual posts where such information was provided. Where georeferences were absent, no data were collected.

Table 1 | Summary of the search terms and processes used to collect biological records of JTM across the study region. Hits refers to the quantity of successful occurrences that contained all of the required information for the study within the time span of the study (2000 – 2018) and within the study region (the UK, Republic of Ireland, France, Belgium, the Netherlands, Luxembourg, Switzerland, Czech Republic, Austria, Germany, Denmark, and Italy). Note that searches on Instagram are limited to hashtags rather than caption text. Data are available from https://figshare.com/s/94529defd9aa93d18426, except GBIF data which are available from the link in the table

Data source	Search terms(s)	Process
GBIF	*Euplagia quadripunctaria Callimorpha quadripunctaria*	Downloaded from GBIF http
iNaturalist	*Euplagia quadripuncteria Callimorpha quadripunctaria*	Downloaded from GBIF – iN
Twitter	*Euplagia quadripuncteria* Jersey Tiger Russischer Bär Spanische Flagge	Manual search
Flickr	*Euplagia quadripuncteria* Jersey Tiger Russischer Bär Spanische Flagge	API query using python code
Instagram	#Euplagiaquadripuncteria #Jerseytiger #RussischerBär #SpanischeFlagge	Manual search

To represent JTM's climatic niche, we used four climatic variables: average maximum temperature, coefficient of variation in average maximum temperature, total precipitation, and coefficient of variation in total precipitation. These four variables have been found to be dominant factors in the range shift and migration of other lepidoptera[24,25]. Climatic data were all calculated per year for the flying time of JTM (July – September)[22]. Climatic data were gathered from WorldClim at a 2.5 minute spatial resolution (~21 km^2)[26,27]. We note that the CHELSA dataset is an improvement on WorldClim, but the differences are very small within Europe, and within the non-mountainous habitat that JTM largely occupies. Our goal was to understand the impact of social media data and recorder effort on biomonitoring of range-shifts, rather than make the most accurate range-shift prediction possible, and it is highly unlikely the small, unsystematic, differences between CHELSA and WorldClim in our study region would affect these results. Other climatic layers were not included to avoid overfitting HSMs, under-predicting potential distributions and tolerances under climatic conditions where species may be underreported[28]. A fifth environmental layer, night light, was used to capture the degree of urbanisation[29]. Night light data were collected from December of every year (data from summer months may not be an accurate representation due to the lighter summers in the northernmost parts of the study region). Data were collected from the National Centers for Environmental Information[30] and converted to a 2.5 minute spatial resolution (~21 km^2) by averaging. Stray light, lightning, lunar illumination, and cloud cover are all removed from the average measure of illumination prior to calculation of averages for each layer. Only data from 2012 onwards were comparable between years, so for all years prior to 2012, the night light dataset from 2012 was used.

Calculating recorder effort for data sources

Accurate estimations of recorder effort have been a significant quandary for many previous studies[2,31–33]. Here, we defined recorder effort as a ratio between the number of records of a species in a location, and the species' estimated abundance in that location. High ratios indicate grid-cells where a species is detected frequently relative to its abundance, and thus recorder effort is high. Recorder effort could not be calculated

for JTM itself since there are no independent estimates of its abundance across the study region. Therefore, we used a surrogate species: the Eurasian blackbird, *Turdus merula* , which has a consistent range and abundance between 2000 and 2018, is easily identifiable, is charismatic (thus of interest to social media users), and has been recorded across all data sources considered between 2000 and 2018 across the study region. Furthermore, the blackbird occupied both urban and rural environments, so using blackbirds to estimate recorder effort should minimise the difference in abundance recording between urban and rural areas. We therefore judged records for this species' occurrence to reflect the interest in recording wildlife in a given time or location[34]. Estimations of blackbird abundance throughout Europe were acquired from the European Breeding Bird Atlas 2[35].

In order to calculate the recorder effort ratio for each data source, we collected blackbird occurrences using the search terms and processes in table S1. A ratio between the number of blackbird records for each data source and the estimated abundance was calculated for each UTM grid-cell (Figure S1 – 3) from the European Breeding Bird Atlas (~50 km^2 resolution, although some cells varied in size). Recorder effort for GBIF was calculated for all years, whereas the recorder effort for other sources was produced for 2016, 2017, and 2018 (the years for which social media data sources were studied; figures S3 – S5). Other approaches to recorder effort have used the number of species recorded in an area[31], however our approach has the advantage that it is not affected by species richness. Moreover, if social media users are indeed more likely to record eye catching or charismatic species, their recorder effort may not be reflected by the overall number of species recorded in a given time or location.

There is a potential confound within this measure of recorder effort given that traditional data are used to estimate blackbird relative abundance: blackbird abundance may underestimated in urban environments as per our own hypotheses. Using blackbird abundance as the denominator in recorder effort calculations could mean we over-estimate recorder effort in urban areas, relative to rural areas. However, this shouldn't affect the relative difference in recorder effort between data sources within urban areas.

Comparing JTM's habitat usage obtained from different data sources

In order to ask whether social media data sources included more urban records than did GBIF, we compared the logged intensity of night light between records from each source.

To ask whether GBIF data underestimated the urban component of JTM's range-shift, we compared social media records to habitat suitability calculated using GBIF records. GBIF HSMs were produced using bioclim in the *dismo* package. Bioclim is a distance-based, boxcar method for assessing habitat suitability based on the similarity of bioclimatic variables between points in space[36]. Thus, bioclim is simple and robust, which is ideal for comparing habitat suitability at points in different regions and time periods, when the placement of pseudo-absences might strongly affect habitat suitability estimates. First, a single historic ('GBIF-calculated') HSM was calculated for 2000-2009 using the average climatic variables and occurrences of JTM in GBIF from these years, and night light data from 2012. This model was a suitable baseline as it would average out any unusual bioclimatic conditions that could occur within a single year and boasted a relatively large sample size (N = 775). We used a randomly selected 80% of the data points as training data to construct the historic model. Model performance (measured as area under the receiver operating curve; AUC, and calculation of the Boyce Index[51]) was calculated using the remaining 20% of the data as a testing dataset. In order to calculate the AUC and Boyce Index, pseudo-absences were generated by selecting random points from the same study region as the presence data with a 50% prevalence. When predicting suitable and unsuitable habitat, we used a sensitivity threshold of 0.9. This maximised the potential suitable habitat for JTM and partially accounted for underreporting. A threshold of 0.95 was also attempted but discarded since it classified areas that are almost certainly unsuitable for JTM (such as the Scottish Highlands[22]) as suitable.

The historic model was then used to predict the relative habitat suitability for JTM for each year between 2010 and 2018 across the study region using the climatic and night light variables for each year (with the exceptions of 2010 and 2011, which used the night light data from 2012). We extracted habitat suitability

from HSMs at the coordinates of each occurrence of JTM from each data source across the study region for the years 2016–2018. The years 2016 – 2018 were selected as these had relatively large sample sizes for all data sources. In order to test if different data sources recorded JTM in areas of differing habitat suitability across the study region, a linear model was constructed with predicted habitat suitability at each occurrence of JTM as the response variable and the source of the occurrence data as a predictor variable. Any differences between sources were then investigated via Tukey's post-hoc test. Predicted habitat suitability data extracted from JTM occurrence locations were log transformed to homogenise the variance and meet assumptions of linearity. Following this, to investigate if any differences were due to urbanisation, a linear model was produced with night light extracted from JTM occurrence locations as the response variable and the source of the occurrence data as the predictor variable. Night light data were square root transformed to meet the assumptions of linearity. Any differences between sources were then investigated via Tukey's post-hoc test.

Any geographical area with extremes of climate could generate a bias when testing between predicted habitat suitability if one data source happened to be overrepresented in this extreme. For example, if iNaturalist was overrepresented in Italy and Italy was predicted to have a low habitat suitability for JTM (based on data from GBIF) due to extreme temperature, then this could confound a result which suggested that data from iNaturalist were located in areas of significantly lower habitat suitability. Since Italy represented the hottest parts of JTM's range in the study area, we repeated all the above analyses without Italy included in the models and then compared the output of both Italy-included and Italy-omitted analyses. We did not do this for the coldest part of the range of JTM since the range shift into these colder climates (e.g. the UK) is foundational to our questions.

Assessing the contribution of recorder effort to the occurrence of JTM

In order to assess whether recorder effort affected the distribution of known occurrences of JTM throughout the study region, four generalised linear mixed models (GLMMs; one for each data source) were constructed. GLMMs were constructed for the years 2016–2018, with the presence/pseudo-absence of JTM throughout the study region as a binary response variable. Recorder effort, habitat suitability from the historic model, and the interaction between the two were predictor variables in each model. Year was included as a random effect to account for a lack of independence between years. As above, AIC selection was then implemented to select the best model. Habitat suitability and recorder effort were both standardised by subtracting their mean and dividing by their standard deviation. GLMMs were constructed using the*glmmTMB* package and had a binary error structure and a logit link function.

All analyses were conducted in R version 4.0.0[37]. Code is available from https://github.com/nis38/JTM.

Results

The intensity of night light varied significantly between post-2009 records from each data source (ANOVA; $F_{3,\ 1952} = 22.76$, $p < 0.001$, figure 1a), as did GBIF-calculated habitat suitability (ANOVA; $F_{3,\ 1947} = 73.14$, $p < 0.001$, figure 1b). Broadly, occurrences obtained from GBIF and Flickr were found in areas of similar habitat suitability and urbanisation. At GBIF and Flickr occurrences, habitat suitability and urbanisation differed from those at Instagram and iNaturalist occurrences. Occurrences from Instagram were from areas with the highest measure of urbanisation (Table 2). When Italy was removed from the study area, occurrences from GBIF and Flickr were found in areas of significantly different measures of urbanisation but no other results were affected (Figure S4).

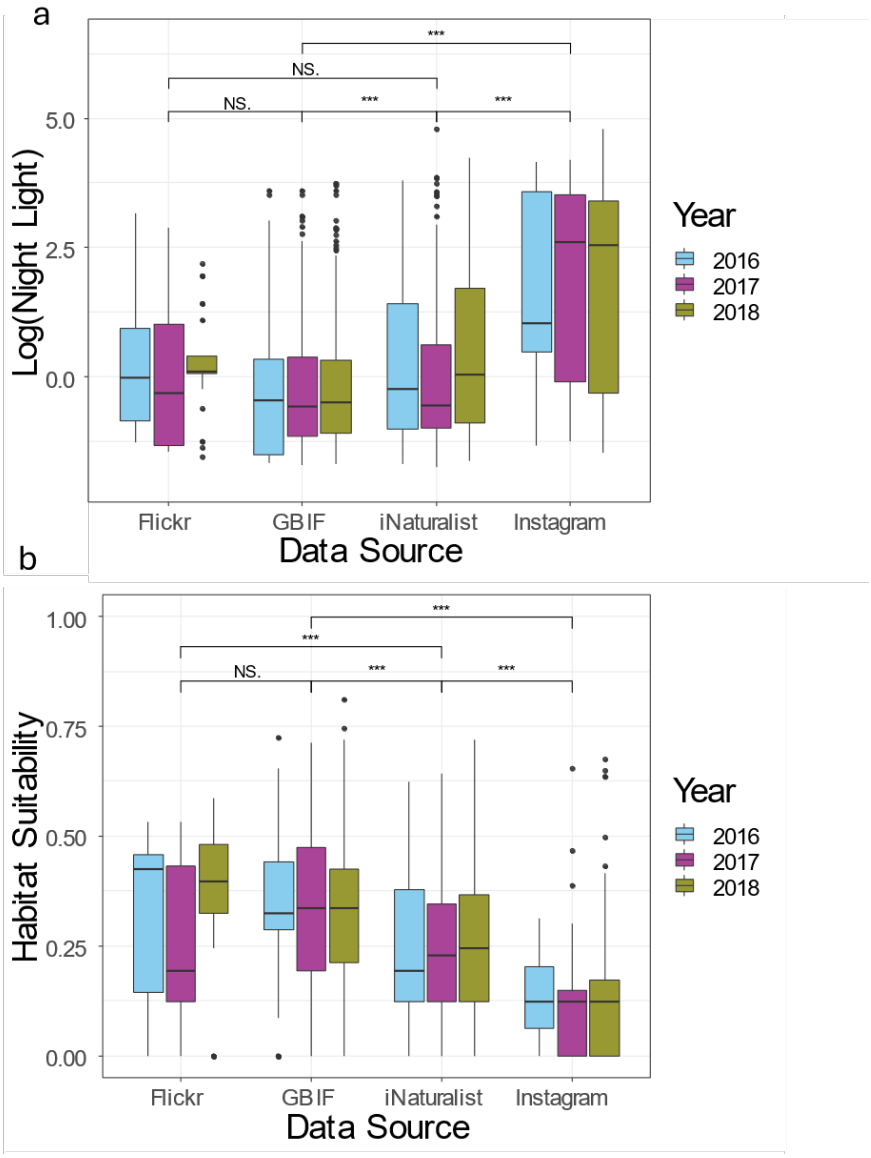

Figure 1 | Differences in (a) night light (urbanisation) and (b) habitat suitability between different sources of occurrence data . Data were taken from the selected study region across 2016 - 2018. (a) and (b): horizontal black bars denote median; vertical bars denote quantiles; NS denotes no significant difference, asterisks denote statistically different variables, and quantity of asterisks denote size of p-value (*** = $p < 0.001$).

Table 2 | Summary of Tukey's post-hoc tests between (log) habitat suitability and urbanisation (square root night light) for four sources of JTM occurrence data.

Comparison	Habitat suitability difference	Habitat suitability p-value	Urbanisation difference
GBIF - Flickr	- 0.000	1.000	- 0.425
iNaturalist - Flickr	- 0.088	**0.002**	0.166
Instagram - Flickr	- 0.193	$<$ **0.001**	1.626

Comparison	Habitat suitability difference	Habitat suitability p-value	Urbanisation difference
iNaturalist - GBIF	- 0.087	$<$ **0.001**	0.591
Instagram - GBIF	- 0.193	$<$ **0.001**	2.051
Instagram - iNaturalist	- 0.105	$<$ **0.001**	1.460

Recorder effort and GBIF-calculated habitat suitability affected the 2016-2018 occurrences of JTM in all data sources (Figure 2, Table S4). Unsurprisingly, occurrence records from GBIF were present in areas of high predicted habitat suitability, but GBIF also mostly recorded JTM where recorder effort was high (Figure 2a and 2e). Likewise, occurrence records from iNaturalist and Flickr were present in areas of relatively high recorder effort and GBIF-calculated habitat suitability (Figures 3b and 3f, 3c and 3g). However, occurrence records from Instagram were more likely to be in areas with low GBIF-calculated habitat suitability. While Instagram records were more likely to be in well-recorded areas, the effect of recorder effort was less than for the other data sources (figures 3d and 3h).

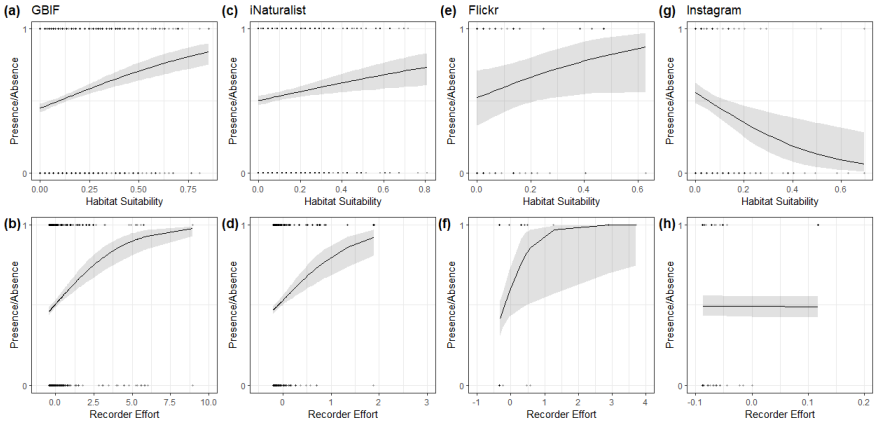

Figure 2 | Effect of standardised values of habitat suitability (panels a, c, e, g, and i) and recorder effort (b, d, f, h, and j) on the presence of JTM for GBIF data (a and b), iNaturalist (c and d), Flickr (e and f), and Instagram (g and h). 0 on the y axis refers to a pseudoabsence, 1 refers to presence of JTM. Lines predicted from GLMMs with year as random effect, however, due to a low variance explained by year (mean standard deviation = 0.020), only the average effect across years was plotted per panel. Grey area denotes 95% confidence interval.

Even though the geographic background for the HSMs did not include the species' entire range, this did not affect our results, since all post-2009 JTM records are found within climate conditions that is analogous to the historical range (Figure S5). GBIF-calculated suitable habitat across the study region was relatively consistent between years, with notable exceptions in the UK, Republic of Ireland, Italy, and Denmark (Figure 2 and S6). The mean ($\pm$ standard deviation) AUC of HSMs was 0.743 ($\pm$ 0.057) and mean Boyce Index was 0.367 ($\pm$ 0.246).

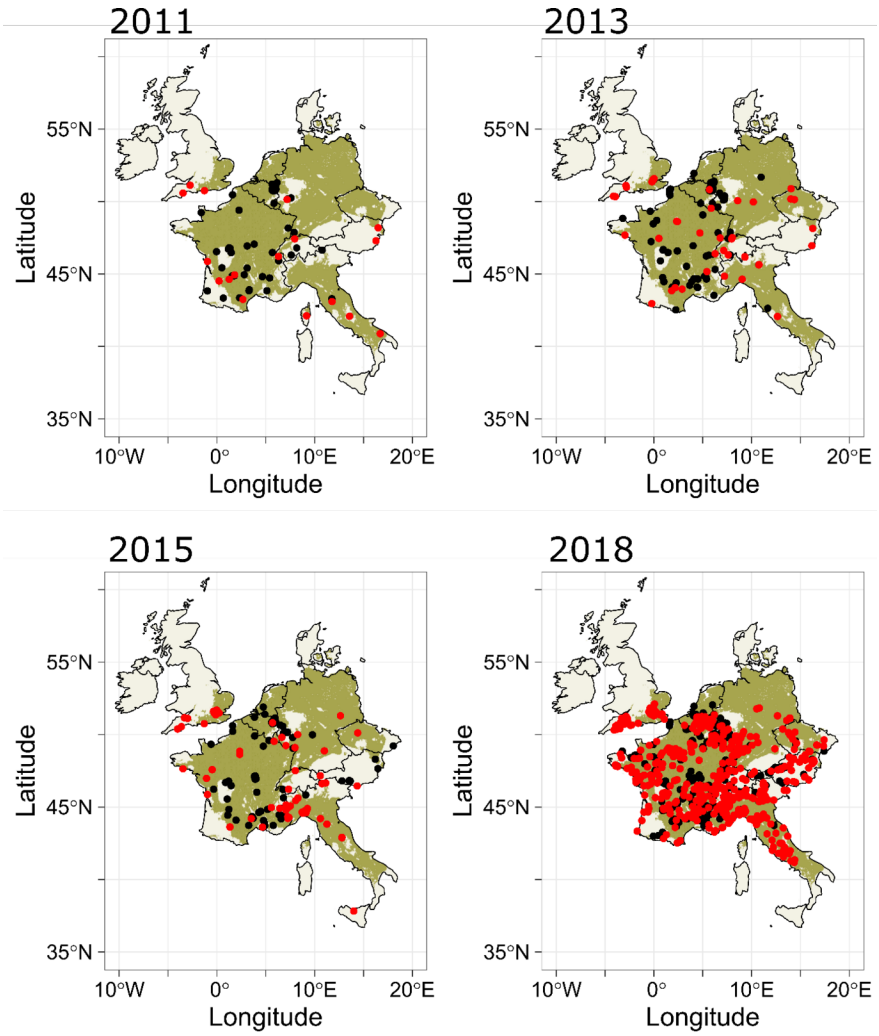

Figure 2 | HSMs for JTM across the study region. Green areas represent suitable habitat for JTM (sensitivity = 0.9); black points are from GBIF; red points originate from Flickr, Instagram, and iNaturalist. Maps presented here selected to display general pattern of changes in habitat suitability (all maps see Figure S6). Habitat suitability was derived from information on maximum temperature, coefficient of variation in maximum temperature, total precipitation, coefficient of variation in total precipitation, and urbanisation (night light) using bioClim models produced with the *dismo* package in R.

Discussion

In this study, we demonstrate that occurrence records from Instagram and iNaturalist are found in different and more urbanised locations compared to occurrences from traditional datasets, such as GBIF. We therefore highlight the utility of these social media platforms as additional and complementary sources of data to traditional databases, such as GBIF. However, contrary to our predictions, occurrence data from Flickr offer a somewhat similar outlook to that provided by GBIF records. There was a notable difference in the environments surveyed by different social media platforms, with Flickr data occurring in more rural locations than data from iNaturalist, and Instagram occurring in still less rural areas.

As predicted, the majority of post-2009 occurrence records from GBIF for JTM within our study region fell within more rural areas, likely due to the majority of GBIF data originating from scientific surveys

and formal recorder efforts, which largely operate in rural zones. Contrary to this, as predicted, Instagram largely contains data from highly urban zones, likely because urban areas are densely populated by humans and have good internet connections leading to geographic trends in human behaviour affecting whether they upload data to social media. However, data from Flickr are more rural than data from iNaturalist and Instagram. Flickr is tailored towards individuals with an interest in the quality of photography, which may attract wildlife photographers intent on capturing wildlife in relatively rural environments, rather than in urban zones. Overall, this demonstrates the utility of social media sites such as iNaturalist and Instagram to fill a void in the occurrence records provided by traditionally used data sources such as GBIF. Although we only studied one species, we think it is likely that the urban-rural differences between databases would remain for similar (colourful, eye-catching) species that would likely be uploaded to social media.

Our results also highlight the importance of accounting for recorder effort. The strong positive effect of recorder effort on GBIF occurrences indicates that JTM is detected where recorders are searching for it. Thus GBIF's predictions of low suitability in urban areas is not necessarily trustworthy. Likewise, iNaturalist and Flickr data also occur in areas where recorder effort on their platforms is high, indicating that these data sources alone may not contain occurrences in all areas the species is found. In contrast to iNaturalist and Flickr, there was a much shallower relationship between the location of JTM records from Instagram and recorder effort. This suggests that Instagram is better at detecting a species in areas where it is not looked for by the majority of users. Instagram's ability to both detect JTM in areas of lower GBIF-calculated habitat suitability and areas of higher urbanisation, as well as a relatively shallow effect of recorder effort makes it an ideal complement to traditional occurrence data for range shifters. The utility of Flickr and iNaturalist should not be discounted though, since both may make species records publically available more rapidly than GBIF.

It should be noted that recorder effort was particularly geographically uneven for social media sources and our results could be affected by this patchiness. It's possible that blackbird recorder effort does not reflect JTM recorder effort, particularly within GBIF, since surveys for different taxa (birds and insects, in this instance) are likely to employ differing sampling techniques and audiences[38]. However, abundance data for insects with which to calculate recorder effort are rarely available. Moreover, using this species, which is well represented in all data sources, would allow for comparison of recorder effort between localities and time periods that could be applied to a wide range of taxa. There may be a novelty bias towards range shifting species, causing geographical and temporal variation in recorder effort. Given the varying, but broadly important, effect of recorder effort, developing improved recorder effort metrics could be particularly important to the use of social media data in biogeography and range-shift ecology. Even if not a precise, quantitative metric of recorder effort, the approach we developed is a useful tool for comparison between data sources, locations, and time periods. This is particularly important when dealing with social media data, which are prone to temporal and spatial trends and uneven geographical use.

Range-shifting and invasive species have previously been found to be human-associated, persisting in urban parks and gardens[10]. Although the extent of this association remains unknown, our results highlight the potential for social media data to track and understand range-shifting species in urban zones. Since Instagram's focus is on photography, it could be used to track the arrival of eye-catching or charismatic taxa in urban area. However, a less recognisable or visually appealing species than JTM could generate fewer occurrences, and thus the repeatability of the use of Instagram data across different taxa requires further investigation. In addition, collection of ad hoc social media data may present opportunities for researchers to assess wildlife management practises in urban and suburban areas. Surveys of bug hotels, bird feeders, and mutualists from social media could be recorded to assess hotspots of positive management in cities, as well as areas that are deficient in their capacity to support biodiversity. Furthermore, social media data could be used to assess the persistence of endangered species in urban and suburban areas, adding to the work already compiled regarding the importance of gardens in supporting threatened or keystone taxa[39]. Our study also suggests that there may even be scope for assessing the potential for urban spaces to propagate range-shifts and invasions further in a similar way to forest corridors[40]. It is clear that, if robust and repeatable methodologies can be applied, social media data sources have a high potential to provide high quality

data at speed. Furthermore, it is likely that these methods will only increase in importance as urbanisation rises globally[41]. It is also noteworthy that social media platforms such as Twitter have been used to promote uptake of the UK ladybird survey, yielding insights into the spread of the Harlequin ladybird[42].

Scientific and policy-maker interest in community-science in urban areas is growing, given that urban environments are increasing, most people live in urban environments, and most nature experiences are close to home[43]. Noticing urban wildlife can improve mental and physical wellbeing[44,45], and increasing engagement with urban nature offers opportunity for improved ecological literacy and nature connectedness, particularly amongst social groups that have historically had inequitable access to nature[46,47]. Our results further reinforce recent findings that social media platforms could be harnessed to assist in urban nature engagement and conservation[48,49]. While our results highlight a promising avenue for future studies and offer novel sources of data with new information, a fundamental area of improvement is the establishment of a rigorous and consistent methodology[19]. A source of uncertainty for this study is that the search terms and the access and use of APIs could not be made consistent across all social media data sources. The process by which data are attained would be benefited by greater consistency; the main barrier here is the expense of using the API services supplied by Instagram and Twitter. Both services have recency constraints and query limits associated with the free-to-use APIs, and the cost of more expansive API usage was outside of the budget of this study, costing up to £2000 per month depending on the service used at the time at which this study was conducted. This could be overcome with additional studies highlighting the importance of access to these data for scientists, thus prompting social media companies to produce an API service that is accessible to scientists. Alternatively, machine learning programmes such as UI Path could provide a more affordable and consistent method to gather data from online sources[50]. Implementation of alternative methodologies and different focal species are likely to increase the utility of Twitter, which was omitted from analyses due to a low sample size, and could permit use of other social media sources not considered here due to data accessibility, such as TikTok or Facebook.

A further potential issue with social media use is that there is not necessarily equal utilisation of these sources throughout all nations, particularly in those outside of Europe and North America. We have attempted to account for unequal usage in our study by using three different sources of social media data, but ideally more could be implemented. Search terms should also be considered with caution. We have included the search terms that yielded the most occurrences of JTM. However, there may be an English bias here, since social media users from non-English speaking individuals will likely submit potential occurrences using English and colloquial terminology. Although various common names are not always simple to incorporate (as was the case here, with German names such as "Spanish flag" and "Russian bear", which yielded countless non-moth results when searched), this is certainly worthy of consideration. Social media data sources are also driven by trends, which may contribute to varying usefulness of different sources over time as the popularity and novelty of range shifting species wax and wane. Such an effect seemed to be apparent with JTM, where the inclusion of the moth on postage stamps in the Channel Islands was associated with an increase in GBIF and iNaturalist occurrences in 2012 and 2013 (which also illustrates that even GBIF is not resistant to trends). Nonetheless, such trends could also be a potential advantage to social media data sources. In theory, governments and scientists could highlight species of interest to the public, thus generating a trend around focal organisms that could be used to generate social media occurrence records. Such strategies could increase the use of social media to record biological phenomena, potentially producing large quantities of community science data.

The results presented here support the idea that the combined use of traditional (GBIF) and social media (particularly Instagram and iNaturalist) data sources to generate a more complete understanding of the habitat-use of range shifting species. Our study suggests that traditional and social media biodiversity data can contain different, but complementary, information regarding habitat usage of a range shifting species. While GBIF captures the rural range of JTM across the study region, Instagram demonstrated that JTM also occupies highly urbanised environments. Social media data may be particularly prone to variation in recorder effort, and we propose a method that can account for this. We suggest that data from social media should be added to occurrence datasets when tracking range shifting species. Implementation of occurrence

records from social media could be particularly important given the human-associated nature of some range shifters, which often occupy parks and gardens in urban zones as well as rural spaces.

Acknowledgements

We thank J. Cranston, L. Hahn, FABio Lab Group for useful discussion. Additional thanks to Sergi Herrando and the European Breeding Bird Atlas, who provided Blackbird data prior to publication of the atlas. For the purpose of open access, the author has applied a 'Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

Author's contribution

NS co-conceived the project, led the data curation, analysis, investigation, visualization, and wrote the original draft. NP co-conceived the project, supported in investigation, methodology, supervision, and reviewed and edited the manuscript. RE co-conceived the project, led the supervision, supported the investigation and methodology, and reviewed and edited the manuscript.

Funding

NP was funded by Research England.

Conflict of interest

The authors declare no known conflict of interest.

Data availability

The code used to run analyses is available at https://github.com/nis38/JTM. The GBIF dataset on Euplagia quadripuncteria we used can be downloaded from https://doi.org/10.15468/dl.dtfjkv. The GBIF dataset on Turdus merula we used can be downloaded from https://doi.org/10.15468/dl.dn3vez. Other data we used are available at https://figshare.com/s/94529defd9aa93d18426.

References

1. Hamann, A. & Wang, T. Potential effects of climate change on ecosystem and tree species distribution in British Columbia.*Ecology* **87** , 2773–2786 (2006).2. Dennis, R. L. H., Sparks, T. H. & Hardy, P. B. Bias in Butterfly Distribution Maps: The Effects of Sampling Effort. *J. Insect Conserv.* **3** , 33–42 (1999).3. Van Der Putten, W. H., Macel, M. & Visser, M. E. Predicting species distribution and abundance responses to climate change: why it is essential to include biotic interactions across trophic levels.*Philos. Trans. R. Soc. B Biol. Sci.* **365** , 2025–2034 (2010).4. Dawson, T. P., Jackson, S. T., House, J. I., Prentice, I. C. & Mace, G. M. Beyond Predictions: Biodiversity Conservation in a Changing Climate. *Science* **332** , 53–58 (2011).5. Pettorelli, N., Smith, J., Pecl, G. T., Hill, J. K. & Norris, K. Anticipating arrival: Tackling the national challenges associated with the redistribution of biodiversity driven by climate change. *J. Appl. Ecol.* **56** , 2298–2304 (2019).6. Wallingford, P. D.*et al.* Adjusting the lens of invasion biology to focus on the impacts of climate-driven range shifts. *Nat. Clim. Change* **10** , 398–405 (2020).7. Cranston, J., Crowley, S. L. & Early, R. UK wildlife recorders cautiously welcome range-shifting species but incline against intervention to promote or control their establishment.*People Nat.* **4** , 879–892 (2022).8. Hirzel, A. H., Helfer, V. & Metral, F. Assessing habitat-suitability models with a virtual species. *Ecol. Model.* **145** , 111–121 (2001).9. Anderson, R. P., Araújo, M., Guisan, A., Lobo, J. M. & Martínez-Meyer, E.*Are Species Occurrence Data in Global Online Repositories Fit for Modeling Species Distributions? The Case of the Global Biodiversity Information Facility (GBIF)* . 1–27 (2016).10. Van Der Veken, S., Hermy, M., Vellend, M., Knapen, A. & Verheyen, K. Garden plants get a head start on climate change. *Front. Ecol. Environ.* **6** , 212–216 (2008).11. Estrada, A., Morales-Castilla, I., Meireles, C., Caplat, P. & Early, R. Equipped to cope with climate change: traits associated with range filling across European taxa. *Ecography* **41** , 770–781 (2018).12. Samy, G. *et al.* Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. *Biodivers. Inform.* **8** , (2013).13. Kusber, W.-H. *et al.* From cleain the valves to cleaning the data: Case studies using diatom biodiversity data on the internet. *Studi Trent Sci Nat* **84** , 111–122.14.

Straub, S. C., Thomsen, M. S. & Wernberg, T. The Dynamic Biogeography of the Anthropocene: The Speed of Recent Range Shifts in Seaweeds. in*Seaweed Phylogeography* (eds. Hu, Z.-M. & Fraser, C.) 63–93 (Springer Netherlands, Dordrecht, 2016). doi:10.1007/978-94-017-7534-2_3.15. Sinka, M. E. *et al.* A new malaria vector in Africa: Predicting the expansion range of*Anopheles stephensi* and identifying the urban populations at risk. *Proc. Natl. Acad. Sci.* **117** , 24900–24908 (2020).16. Delaney, D. G., Sperling, C. D., Adams, C. S. & Leung, B. Marine invasive species: validation of citizen science and implications for national monitoring networks. *Biol. Invasions* **10** , 117–128 (2008).17. Maistrello, L., Dioli, P., Bariselli, M., Mazzoli, G. L. & Giacalone-Forini, I. Citizen science and early detection of invasive species: phenology of first occurrences of Halyomorpha halys in Southern Europe. *Biol. Invasions* **18** , 3109–3116 (2016).18. Sumner, S., Bevan, P., Hart, A. G. & Isaac, N. J. B. Mapping species distributions in 2 weeks using citizen science. *Insect Conserv. Divers.* **12** , 382–388 (2019).19. Jarić, I. *et al.* iEcology: Harnessing Large Online Resources to Generate Ecological Insights. *Trends Ecol. Evol.* **35** , 630–639 (2020).20. Hall, D. M. *et al.* The city as a refuge for insect pollinators.*Conserv. Biol.* **31** , 24–29 (2017).21. Mancini, F., Coghill, G. M. & Lusseau, D. Using social media to quantify spatial and temporal dynamics of nature-based recreational activities. *PLOS ONE* **13** , e0200565 (2018).22. Waring, P. & Townsend, M.*Field Guide to the Moths of Great Britain and Ireland* . (Bloomsbury Publishing, 2017).23. Sorace, A. & Gustin, M. Distribution of generalist and specialist predators along urban gradients.*Landsc. Urban Plan.* **90** , 111–118 (2009).24. Sparks, T. H., Dennis, R. L. H., Croxton, P. J. & Cade, M. Increased migration of Lepidoptera linked to climate change. *Eur. J. Entomol.* **104** , 139–143 (2007).25. Sparks, T. H., Roy, D. B. & Dennis, R. L. H. The influence of temperature on migration of Lepidoptera into Britain. *Glob. Change Biol.* **11** , 507–514 (2005).26. WorldClim. Historical monthly weather data. (2020).27. Harris, I., Jones, P. & Osborn, T. J. Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset. *Int. J. Climatol.* **642** , 623–642 (2014).28. Early, R. & Sax, D. F. Climatic niche shifts between species' native and naturalized ranges raise concern for ecological forecasts during invasions and climate change. *Glob. Ecol. Biogeogr.* **23** , 1356–1365 (2014).29. Gaston, K. J., Visser, M. E. & Hölker, F. The biological impacts of artificial light at night: the research challenge. *Philos. Trans. R. Soc. B Biol. Sci.* **370** , 20140133 (2015).30. National Centers for Environmental Information. Version 1 VIIRS Day/Night Band Nighttime Lights. (2019).31. Isaac, N. J. B. & Pocock, M. J. O. Bias and information in biological records: Bias and information in biological records. *Biol. J. Linn. Soc.* **115** , 522–531 (2015).32. Hassall, C. & Thompson, D. J. Accounting for recorder effort in the detection of range shifts from historical data: *Detecting range shifts from historical data* . *Methods Ecol. Evol.* **1** , 343–350 (2010).33. Casey, L. M. Using citizen science to monitor bumblebee populations. (University of Sussex, 2016).34. BirdLife International. *Turdus Merula* . (2016).35. Keller, V. *et al. European Breeding Bird Atlas 2: Distribution, Abundance and Change* . (European Bird Census Council & Lynx Edicions, Barcelona, 2020).36. Beaumont, L. J. *et al.* Which species distribution models are more (or less) likely to project broad-scale, climate-induced shifts in species ranges? *Ecol. Model.* **342** , 135–146 (2016).37. R Core Team. R: A language and environment for statistical computing. (2020).38. Mair, L. & Ruete, A. Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa.*PLOS ONE* **11** , e0147796 (2016).39. Lowenstein, D. M. & Minor, E. S. Diversity in flowering plants and their characteristics: integrating humans as a driver of urban floral resources. *Urban Ecosyst.* **19** , 1735–1748 (2016).40. Melles, S. J., Fortin, M.-J., Lindsay, K. & Badzinski, D. Expanding northward: influence of climate change, forest connectivity, and population processes on a threatened species' range shift. *Glob. Change Biol.* **17** , 17–31 (2011).41. Goddard, M. A., Dougill, A. J. & Benton, T. G. Scaling up from gardens: biodiversity conservation in urban environments. *Trends Ecol. Evol.* **25** , 90–98 (2010).42. Roy, H. E. *et al.* The harlequin ladybird, Harmonia axyridis: global perspectives on invasion history and ecology. *Biol. Invasions* **18** , 997–1044 (2016).43. Veerkamp, C. J. *et al.* A review of studies assessing ecosystem services provided by urban green and blue infrastructure. *Ecosyst. Serv.* **52** , 101367 (2021).44. Aerts, R., Honnay, O. & Van Nieuwenhuyse, A. Biodiversity and human health: mechanisms and evidence of the positive health effects of diversity in nature and green spaces. *Br. Med. Bull.* **127** , 5–22 (2018).45. Houlden, V., Jani, A. & Hong, A. Is biodiversity of greenspace important for human health and wellbeing? A bibliometric analysis and systematic literature review. *Urban For. Urban Green.* **66** , 127385 (2021).46. Cooper, D. S. *et al.* Large Cities Fall Behind in "Neighborhood Biodiversity". *Front. Conserv. Sci.* **2** , 734931 (2021).47. Amorim

Maia, A. T., Calcagni, F., Connolly, J. J. T., Anguelovski, I. & Langemeyer, J. Hidden drivers of social injustice: uncovering unequal cultural ecosystem services behind green gentrification. *Environ. Sci. Policy* **112** , 254–263 (2020).48. Persson, A. S., Hederström, V., Ljungkvist, I., Nilsson, L. & Kendall, L. Citizen science initiatives increase pollinator activity in private gardens and green spaces. *Front. Sustain. Cities* **4** , 1099100 (2023).49. Langemeyer, J., Calcagni, F. & Baró, F. Mapping the intangible: Using geolocated social media data to examine landscape aesthetics. *Land Use Policy* **77** , 542–552 (2018).50. Sirisuriya, S. de S. A Comparative Study on Web Scraping. (2015).51. Hirzel, A.H., Le Lay, G., Helfer, V., Randin, C., & Guisan A. Evaluating the ability of habitat suitability models to predict species presences. Ecological Modelling 199, 142-152 (2006).

52. GBIF: The Global Biodiversity Information Facility (year) *What is GBIF?* Available from https://www.gbif.org/what-is-gbif

53. GBIF.org (30 January 2020) GBIF Occurrence Download https://doi.org/10.15468/dl.dtfjkv

Supplementary materials

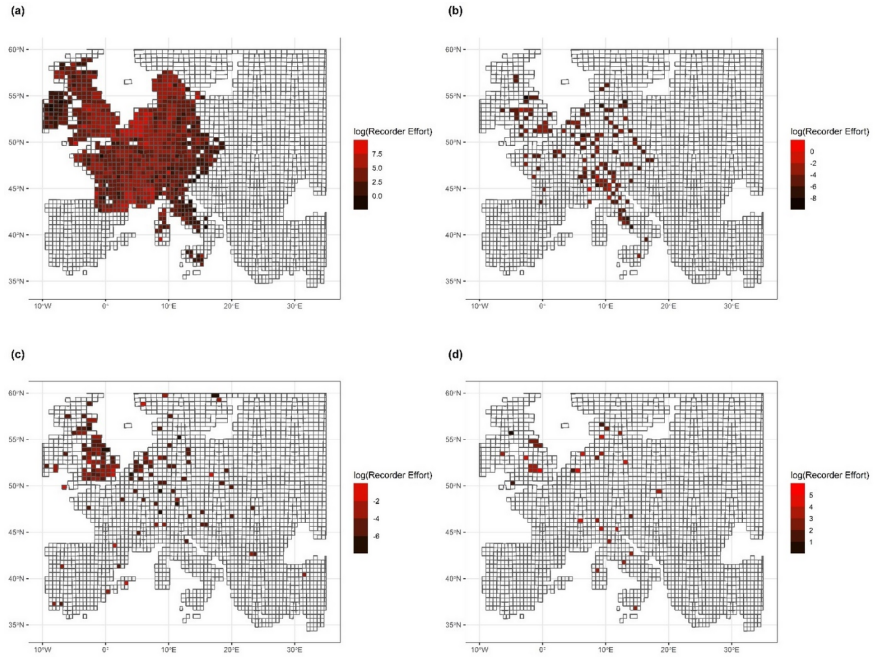

Figure S1 | Recorder effort in 2016 from different data sources. (a) GBIF recorder effort; (b) iNaturalist recorder effort; (c) Flickr recorder effort; (d) Instagram recorder effort. Recorder effort was

calculated as the abundance of blackbirds in a cell as reported by that data source divided by the actual estimated abundance according to data from the European Breeding Bird Atlas[36]. Grid size is approximately 50 km^2, although some gird cells varied in size. Grid was supplied by European Breeding Bird Atlas.

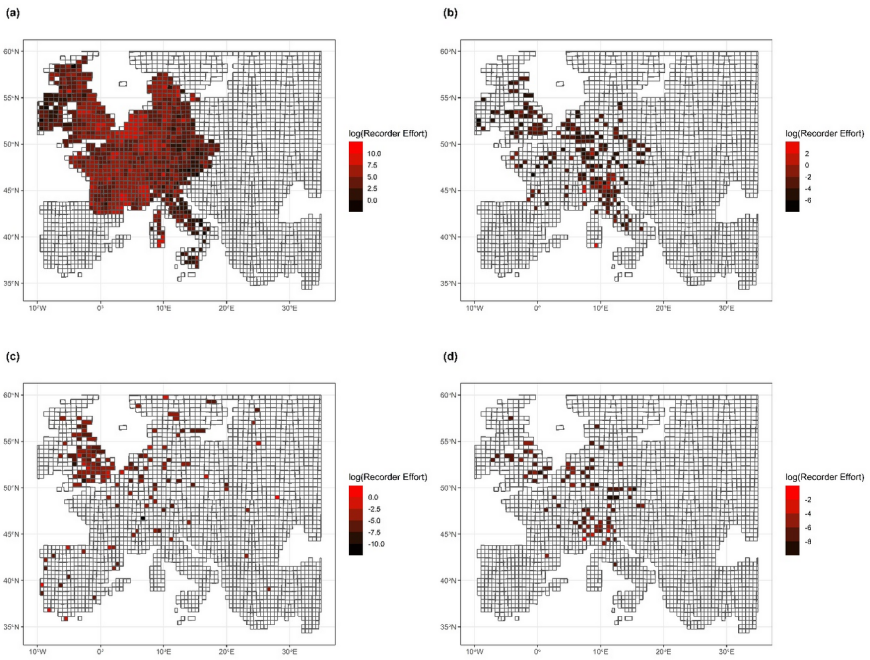

Figure S2 | Recorder effort in 2017 from different data sources. (a) GBIF recorder effort; (b) iNaturalist recorder effort; (c) Flickr recorder effort; (d) Instagram recorder effort. Recorder effort was calculated as the abundance of blackbirds in a cell as reported by that data source divided by the actual estimated abundance according to data from the European Breeding Bird Atlas[36]. Grid size is approximately 50 km^2, although some gird cells varied in size. Grid was supplied by European Breeding Bird Atlas.

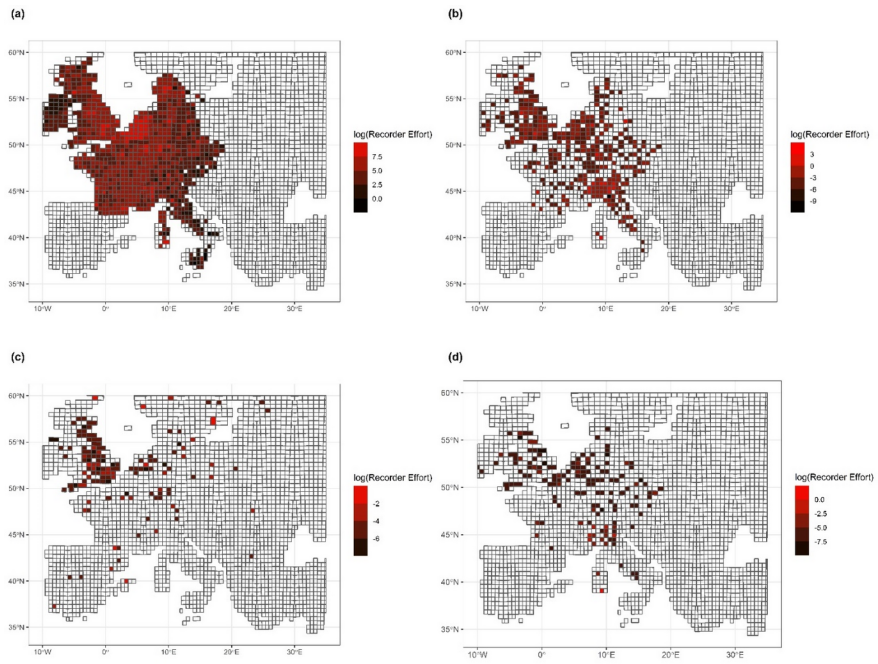

Figure S3 | Recorder effort in 2018 from different data sources. (a) GBIF recorder effort; (b) iNaturalist recorder effort; (c) Flickr recorder effort; (d) Instagram recorder effort. Recorder effort was calculated as the abundance of blackbirds in a cell as reported by that data source divided by the actual estimated abundance according to data from the European Breeding Bird Atlas[36]. Grid size is approximately 50 km^2, although some gird cells varied in size. Grid was supplied by European Breeding Bird Atlas.

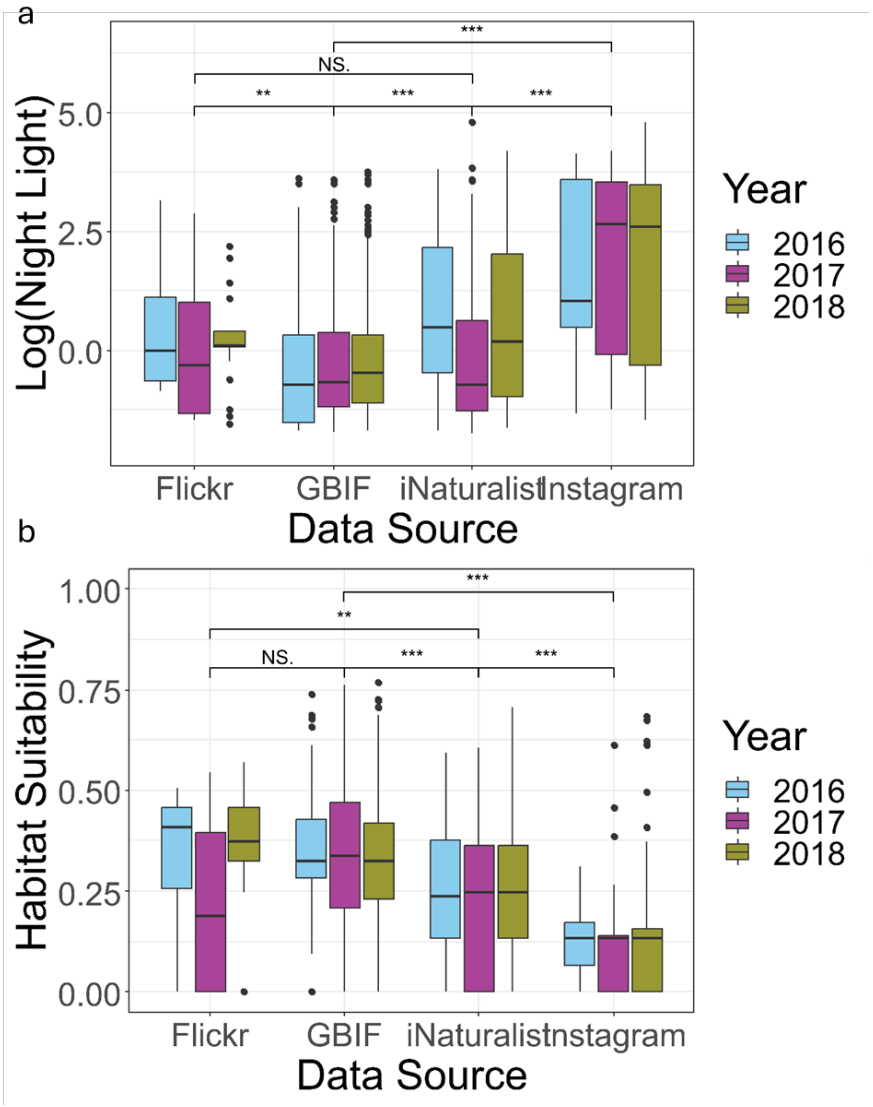

Figure S4 | Differences in (a) night light (urbanisation) and (b) habitat suitability between different sources of data with Italy removed from the study region . (a) and (b): horizontal black bars denote median; vertical bars denote quantiles; NS denotes no significant difference, asterisks denote statistically different variables, and quantity of asterisks denote size of p-value (*** = $p < 0.001$).

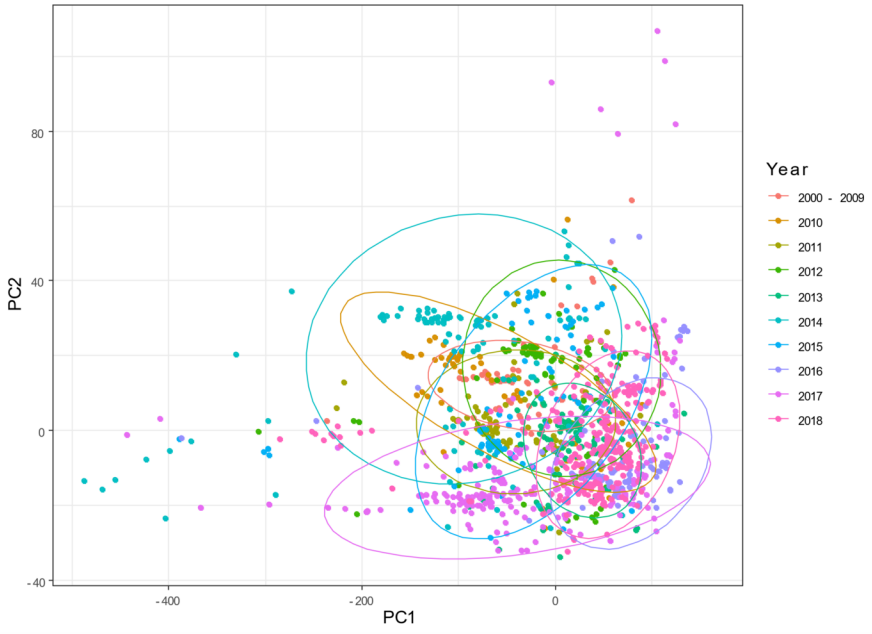

Figure S5 | PCA plot containing environmental variables from occurrences of JTM across years included in this study. Overlap of ellipses suggests that climatic variables where JTM is found have not differed over time. Ellipses plotted using 95% confidence intervals.

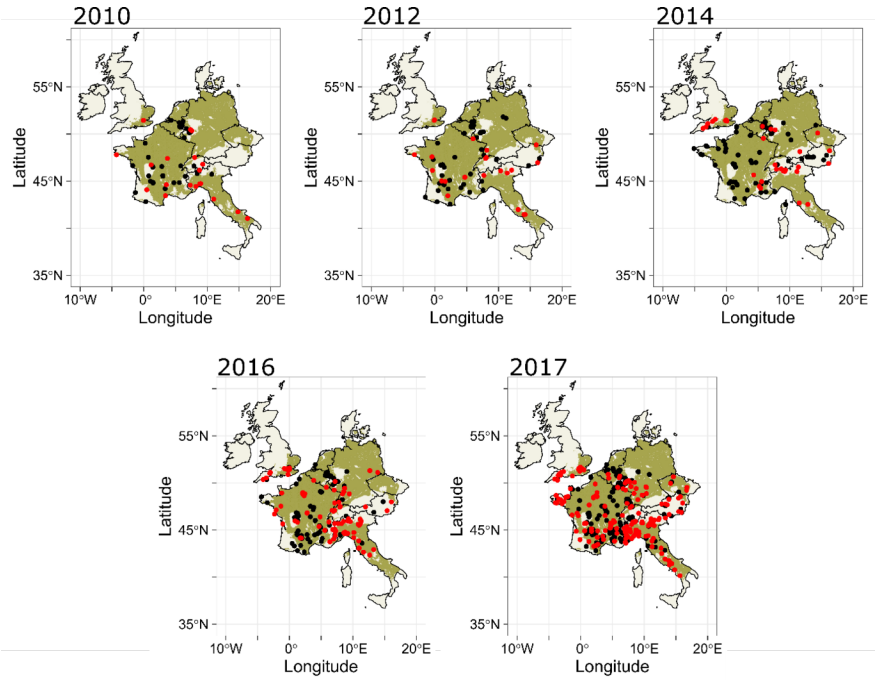

Figure S6 | Additional HSMs for JTM across the study region not included in main text. Green areas represent suitable habitat for JTM (sensitivity = 0.9); black points are from GBIF; red points originate from Flickr, Instagram, and iNaturalist. Habitat suitability was calculated from maximum temperature,

covariance in maximum temperature, total precipitation, covariance in total precipitation, and urbanisation (night light) using BioClim models produced with the *dismo* package in R.

Table S1 | Summary of the search terms and processes used to collect biological records of the Eurasian blackbird across the study region. Note that searches on Instagram are limited to hashtags rather than caption text. Data are available at https://figshare.com/s/94529defd9aa93d18426, except data from GBIF (link in table)

Data source	Search terms(s)	Process
GBIF	*Turdus merula*	Downloaded from GBIF GBIF.org (02 July 2020) GBIF Occurrence Download
iNaturalist	*Turdus merula*	Downloaded from iNaturalist.
Twitter	*Turdus merula*	Manual search
Flickr	*Turdus merula* Blackbird	Automatic API search using python code and then query geographical data usi
Instagram	#Turdusmerula	Manual search

Table S2 | AIC scores for model selection process for recorder effort models. (a) Models produced with GBIF data; (b) models produced with iNaturalist data; (c) models produced with Flickr data; (d) models produced with Instagram data. Interaction term refers to the interaction between habitat suitability and recorder effort. Models with convergence errors were disregarded as no AICc score could be concluded.

(a)

Habitat suitability	Recorder effort	Degrees of freedom	Log Likelihood	AICc	ΔAIc
+	+	3	- 1232.702	2471.417	0.000
-	+	2	- 1252.152	2508.311	36.893
+	-	2	- 1257.461	2518.928	47.511
-	-	1	- 1287.866	2577.735	106.317

(b)

Habitat suitability	Recorder effort	Degrees of freedom	Log Likelihood	AICc	ΔAIc
+	+	3	-1143.693	2293.401	0.000
-	+	2	-1148.777	2301.561	8.159
+	-	2	-1176.011	2356.029	62.628
-	-	1	-1181.121	2364.245	70.844

(c)

Habitat suitability	Recorder effort	Degrees of freedom	Log Likelihood	AICc	ΔAIc
+	+	3	-58.686	123.622	0.000
-	+	4	-60.504	125.133	1.511
+	-	3	-66.541	137.207	13.585
-	-	2	-69.314	140.670	17.048

(d)

Habitat suitability	Recorder effort	Degrees of freedom	Log Likelihood	AICc	ΔAIC
+	-	3	-170.771	345.590	0.000
+	+	3	-170.424	346.944	1.354
-	-	2	-176.751	355.517	9.925
-	+	5	-176.516	357.079	11.489