

Can data mining from various internet platforms systematically accelerate detection of alien species invasions across the EU?

Simon Reynaert^{1*}, Niels Billiet¹, Pavel Pipek^{2,3}, Ana Novoa^{2,4}, Philip Hulme⁵, Sofie Meeus¹, & Quentin Groom¹

¹Meise Botanic Garden, Nieuwelaan 38, B-1860 Meise, Belgium

²Department of Invasion Ecology, Institute of Botany, Czech Academy of Sciences, Průhonice, Czech Republic

³ Department of Ecology, Faculty of Science, Charles University, Prague, Czech Republic

⁴ Estación Experimental de Zonas Áridas, Consejo Superior de Investigaciones Científicas (EEZA-CSIC), Almería, Spain

⁵ Bioprotection Aotearoa, Department of Pest-Management and Conservation, Lincoln University, PO Box 85084, Christchurch 7648, Canterbury, New Zealand

* Corresponding author contact details: +32485 68 28 93 | simon.reynaert@plantentuinmeise.be

Simon Reynaert (Orcid ID: 0000-0003-4690-0955)

Pavel Pipek (Orcid ID: 0000-0003-1116-1013)

Ana Novoa (Orcid ID: 0000-0001-7092-3917)

Keywords: invasive alien species, List of invasive alien species of Union Concern, iEcology, European Union, Bio-informatics, social media, biological invasions, invasion fronts, citizen science

Abstract

Invasive alien species (IAS) expansions are increasingly impacting the biodiversity and economy of Europe. To more effectively allocate the limited resources available for their management, it is pertinent to accelerate detection of IAS spread and distribution. One largely untapped secondary data source showing much potential lies in the automated tracking of internet activity such as IAS search intensity or mentions across different internet platforms. In this study, we tested if internet activity increases systematically when IAS expand into new EU countries utilizing the combined data of 88 invasive species from various internet platforms. In total, 14 internet platforms were screened and evaluated based on their database accessibility, mined data quality and utility for systematic IAS expansion tracking. We found that the procedure to obtain researcher access to minimal data required for IAS tracking (i.e., information about location, time and place) varies widely across platforms, and is particularly difficult without incurring significant costs for many of the larger ones (X, Google and Tiktok). From the explored species, more charismatic species (i.e., mammals) overall gained more online traction than more cryptic ones (i.e., plants), though online activity of the first proved a worse representation of real-world occurrence patterns. Moreover, while the final five selected internet platforms showed increased activity surrounding the year of invasion in many of the explored invasion scenarios (particularly Wikipedia and Facebook), inconsistencies between species groups, trends per platform and the large variability in data quality currently still hampers systematic integration of such data into existing databases. We conclude that combining IAS activity data from various internet platforms shows potential to accelerate IAS expansion detection across the EU (especially for fish, crustaceans, reptiles, birds and plants). However, incorporation in automated early warning systems is currently hampered by variable data quality, limited researcher access to online data and the few open, accurate and generalizable species classification algorithms with API access.

1. Introduction

The cost of invasive alien species (IAS) to the global economy – ranging from damage to infrastructure and crops to depletion of natural resources (e.g., pollinator populations) and deterioration of human health (e.g., disease transmission) – is mounting every year, with Europe taking the heaviest losses (Diagne et al., 2021; Seebens et al., 2017; Soto et al., 2025). In addition, biological invasions are recognized as one of the five principal direct drivers of biodiversity loss (Roy et al., 2024) with impacts encompassing profound chemical, physical, and structural alterations of ecosystems, including changes to nutrient cycling, or fire regimes (Kumschick et al., 2024). To more effectively allocate the limited resources available for IAS management, impact mitigation and knowledge building, accelerating detection of new occurrences is becoming increasingly important (Jarić et al., 2021; Novoa et al., 2025).

Previous studies indicate that the emerging field of iEcology – exploring ecological information hidden in mentions and activity patterns of various internet platforms (e.g., social media, search engines or encyclopedias) –, shows much potential for accelerating detection of IAS outbreaks and better understanding their spread and population dynamics (Jarić et al., 2021; Novoa et al., 2025; Tomojiri & Takaya, 2025). For instance, the presence of invasive alien fish and flies was previously detected solely based on images from angler websites and Facebook groups across different EU countries (Kalous et al., 2018; Schifani & Paolinelli, 2018). Many internet platforms have also proven their utility for more systematic monitoring, such as Facebook for rare whales (Morais et al., 2021) or Flickr, Google images and Instagram for invasive plants (Allain, 2019; Cardoso et al., 2024).

Despite the abundance of potential information available online, iEcology data is fragmented (both spatially and temporarily), of highly variable quality, and usually requires thorough manual cleaning before it is suitable for ecological analysis (see e.g., issues with bots mentioned in Tomojiri & Takaya, 2025). Moreover, secondary data generated from social media mentions and internet activity will inherently be dominated by human bias (Olteanu et al., 2019). For iEcology analyses, data tends to be skewed towards larger-bodied, charismatic and easily recognizable taxa from densely populated areas with sufficient internet access (Jarić et al., 2021; Novoa et al., 2025). Thus, it is virtually impossible to use iEcology data by itself for answering many typical ecological questions relating to IAS dynamics such as, e.g., reliably estimating population size or the full extent of a species' distribution.

Until a few years ago, monitoring efforts were mostly limited to a relatively small number of trained experts. Recently developed and widely used citizen science platforms, such as iNaturalist provide accurate data which are often useful for early detection (González-Moreno et al., 2025; López-Guillén

et al., 2024). However, citizen-science platform related curation and data publishing policies often introduce lags between the time of observation and the occurrence being publicized. Any delays in data publishing may hamper rapid IAS detection (Kourantidou et al., 2022). An alternative source of real-time data, showing potential to become a useful extra component in early warning systems tracking IAS spread, lies in the monitoring and detection of changes in online activity regarding IAS (Cardoso et al., 2024; Cerri et al., 2022; Jarić et al., 2021). To our knowledge, it has not been evaluated yet to what extent iEcology datamining approaches can be applied for the systematic monitoring of IAS dynamics at a large scale (i.e., across multiple species, countries and internet platforms), and by doing so directly complement evidence-based informing of policy makers and other stakeholders regarding IAS occurrences and decisions.

Here, we investigated whether data on IAS mentions and activity from five internet platforms (Facebook, Flickr, iNaturalist, Wikipedia and YouTube) could aid in accelerating country-level detection of IAS expansions across Europe. Our overarching goals were to 1) develop automated and repeatable geolocated IAS data mining workflows and 2) evaluate whether IAS related online activity could serve as a systematic indicator for IAS expansions across Europe. Based on previous case studies (Allain, 2019; Cardoso et al., 2024; Cranswick et al., 2022; Daume, 2016; Fukano & Soga, 2019; Kalous et al., 2018; Kotowska et al., 2021; Schifani & Paolinelli, 2018), we hypothesized that IAS activity patterns across internet platforms would (i) reflect patterns in IAS occurrence data collected through active monitoring available from the Global Biodiversity Information Facility (GBIF) and the European Alien Species Information Network (EASIN) and therefore (ii) show systematic increases in IAS activity surrounding the year of invasion into new EU countries.

2. Methods and Materials

Before data collection, 14 online platforms were screened for their suitability as potential IAS data sources based on qualitative characteristics. These characteristics included i) the type of data (text, images or video), ii) the cost of accessing these data, iii) the geolocation precision associated with online activity, iv) the query limit for the respective application programming interfaces (APIs), v) the temporal resolution of the data, and vi) whether low-cost data mining for research purposes was authorized by the platform (Table S1).

Due to various reasons including (i) declined applications for obtaining researcher API access to perform IAS research (X, Google Ads, Google Search Researcher result API, TikTok), (ii) the lack of accurate geolocation (Reddit, Bluesky, Mastodon, eBay), and (iii) unreliable automated data mining (Google Trends; Alam & Hulme (2025)), we only incorporated data from a total of five online

platforms in our final analyses. These platforms included Facebook, Flickr, iNaturalist ('casual' and 'needs ID' observations only), Wikipedia (Wikimedia) and YouTube.

The complete methodological framework included three steps (Fig. 1). In the first step, information on the 88 species included in the list of invasive alien species of the Union concern (pre-june 2025; Regulation (EU) 2016/1141) was automatically collected from Wikipedia and manually enriched using EASIN data. In a second step, we mined data through APIs about daily videos, photos, posts and activity on these species from the short-listed internet platforms mentioned above. In a third step, we processed and summarized data, conducted time series analysis and compared patterns and trends with reference datasets from the European Alien Species Information Network (EASIN) and GBIF. More details regarding all individual steps are outlined below. Data mining and processing scripts can be found in this Github repository: <https://github.com/Simon-Reynaert/iEcology-IAS-miner>. To protect the privacy of individuals in the public domain, only the raw count per species × country × platform data files are made available for re-analysis, with coordinates anonymized to country level and any identifiers which could potentially lead to the identification of individuals removed. Note also that the scripts for collecting and cleaning the Facebook data mining are not publicly released since Facebook data mining was performed within the Social Media Archive's (SOMAR) secure virtual development environment at the University of Michigan (500 S State St, Ann Arbor, MI 48109, Michigan, USA)(Meta, 2025).

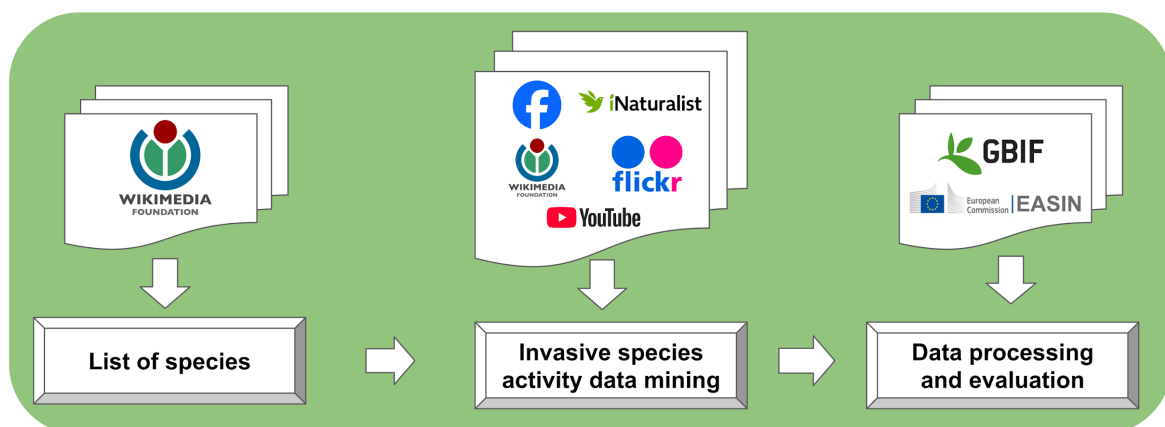


Figure 1. Three step methodological framework for the data collection and analysis performed in this study.

2.1. Data collection

All initial data collection, cleaning, anonymizing and summarizing to daily counts / activities was conducted in Python v 3.13(van Rossum, 2025). All anonymized, GDPR-compliant internet platform

activity datasets, reference data from GBIF and EASIN, as well as the generated output figures for all explored EASIN invasion scenarios are available in a Zenodo repository (Link TBA).

First, the scientific names (“*Genus species*”) of 88 species included in the list of invasive alien species of Union Concern (i.e., the list pre-dating June 2025) were mined from Wikipedia and their associated Wikidata Q identifiers were stored in a data frame (see `species_wikipedia_sitelinks.csv`). Note that sometimes former scientific names were mined due to outdated names circulating on Wikipedia still (e.g., *Eichhornia crassipes* instead of *Pontederia crassipes*). We then collected all posts/observations/videos containing exact matches with these scientific names for countries on the European continent (Table S2) through the respective APIs of Facebook, Flickr (normal tag search), iNaturalist (only non-research grade observations) and YouTube (approximate country bounding boxes) between 01/01/2016 and 15/07/2025. Where possible, we used exact string matching within the API call to only fetch records containing the full scientific name, thus limiting collection of false positives. However, in cases where the API lacked such functionality (e.g., Facebook and YouTube), we performed post-hoc filtering to include only those posts that mentioned the full species name. For Facebook, posts were collected querying all posts containing the scientific names using the `surface_countries` parameter in the API to retrieve posts per country. For Flickr, we collected photos that had the species name in one of the tags such as title, description, etc. (normal tags in API call). Given every posted image on Flickr gets a new ID, we had to filter out ‘repeat’ observations made at approximately the same time and location to only count them once, in case someone had posted multiple pictures of the same observation event (i.e., one post containing multiple pictures). To collect geolocated videos from YouTube we specified a circle radius surrounding a single point coordinate for every European country in our queries, often splitting countries into multiple circles to ensure proper coverage (Table S3).

Next, we collected IAS activity data from Wikipedia. Given it is not social media (i.e., not built around posting videos/images/text), this process was slightly different. For Wikipedia, we collected daily pageview estimates over time as an indicator of platform activity, where possible by specifying “user” in the API endpoint URL to minimize false matches related to bot searches. Wikipedia daily pageviews were collected in two ways, including both geolocated pageviews (for pages with > 90 pageviews / day) and language-based pageviews (all daily pageviews except if only the editor accessed the page). Geo-located pageview data was collected by downloading historical pageview data (from 09/02/2017 until 15/07/2025) from the wikidata pageview repository (Wikimedia, 2025a) utilizing the previously mined Q identifiers to filter out pageviews related to individual species. Given the greater detection capacity (also pageviews < 90 / day) and the generally good correlations with geolocated pageview data (Fig. S1), we decided to also include pageviews that were collected

through EU language filtering. To this end, we collected data from all pages linked to the wiki Q-identifier that existed in the Wikidata database and performed language to country mapping based on plausible matches (see 2.2). The accuracy of this method is known to be language dependent (e.g., we assumed pageviews to the Finnish page of invasive species primarily reflected pageviews originating from Finland, whereas this correlation is very poor for pageviews to the English page; Fig. S1)(Vardi et al., 2021). While some noise is introduced on top of geolocated pageview volumes for privacy protection reasons (Wikimedia, 2025b), we believe the effect of this on overall activity trends over time to be minimal, which was further confirmed by strong correlations with the vast majority of language based pageviews (Fig. S1).

Finally, for collection of reference and validation datasets, we directly received a dataset from the European Alien Species Information Network (EASIN) containing the year in which each species was first recorded per country for the 85 species on the Union list of Concern that were already present in Europe between 2016 and 2025. Furthermore, we collected all observations across all EU countries for the selected species from GBIF through the API and summarized them to daily counts per species and country between 01/01/2016 and 15/07/2025. After collection and initial cleaning of activity data, the daily number of activities/observations/posts per species $\times$ country combination were counted and saved into .csv files ready for analysis for all platforms except Facebook. Due to relatively small post volumes and concerns regarding user's privacy, Facebook posts were counted and grouped per month and normalized by dividing the monthly number of posts by the maximum number of posts retrieved across the entire time series.

2.2. Calculations and statistical analysis

After initial data collection and cleaning which resulted in activities / observations / post counts for all queried species $\times$ country combinations across the various internet platforms, all further analyses were performed in R v 4.4.2(R Core Team, 2024). All plots were constructed using *ggplot2* (Wickham, 2016). Models were constructed with the packages *nlme* (Pinheiro et al., 2025), *lme4* (Bates et al., 2025) and *glmmTMB* (Brooks et al., 2025) and 2-by-2 differences further explored using post-hoc Tukey tests from the *emmeans* (Lenth et al., 2025) package to correct for multiple testing.

First, Wikipedia language-based pageviews were mapped with plausible candidate countries from which we assumed the majority of pageviews had originated (see also 2.1) (Vardi et al., 2021). All activities were summed per month, binning observations to increase signal detection strength while maintaining enough temporal resolution to allow detection of phenological patterns in the data. Datasets were then joined together, species synonyms matched and monthly counts used to calculate 'popularity', defined as the number of unique months with non-zero activity over the entire

period of data collection (01/01/2016 - 15/7/2025). Differences in popularity were then compared by performing ANOVA on a GLMM (family = “negative binomial”, link = “log”) including country, taxon, habitat and internet platform as fixed effects and species nested within taxon as random intercept.

Next, we used this binned monthly data to compare individual internet platform trends with monthly GBIF occurrence data by performing Spearman rank correlations for all queried species × country combinations on the raw data and standardizing data from all platforms between 0 and 1 by dividing every monthly data point by the maximum monthly activity / post count / observation value across the entire time period. The differences in Spearman’s rank correlations were further explored by performing ANOVA on a GLMM (family = “gaussian”, link = “identity”) including country, taxon and internet platform as fixed effects and species nested within taxon as random intercept.

Finally, time series analysis was performed to detect anomalous increases in platform activity surrounding the year of reported invasion in the EASIN data for a total of 112 species × country combinations (i.e., recorded EU IAS invasion scenarios) where new invasions had occurred for the given set of species between 2016 and 2025. This was done for all raw monthly activity data for platforms individually (i.e., all 112 invasion scenarios were explored per platform) as well as for the lumped monthly sums of normalized activities across all platforms.

To assess if activity increased over time for each species on any platform for every European country, every species × country combination where a platform had sufficient data (at least one month with non-zero activity and available data for three separate months) had its counts summed per month to perform anomaly detection with the package *anomalize* (Dancho, 2023) (method = “twitter”, frequency = “auto” and trend = “auto” in the time series decomposition step; method = “gesd” with alpha = 0.25 in the anomaly detection step). The “twitter” method performs seasonal time series decomposition by piecewise medians and was specifically designed for the analysis of social media data (Dancho, 2023). Then, for all 112 EU species × country combinations with known invasions between 2016 - 2025, we calculated if a positive anomaly was present within a two year window surrounding the 1st of January of the EASIN reported year of invasion (i.e., the earliest possible invasion date +/- 1 year) and if so, how many days there was between the closest increase in anomalous platform activity and that earliest possible invasion date. In other words, we detected the presence and nature of the shortest ‘lead’ or ‘lag’ expressed in number of days between the earliest possible invasion date (1st of January of the EASIN reported year of invasion) and anomalous increase in platform activity within a two year window surrounding the earliest possible invasion date. The proportion of detected anomalous activity increases during this 2-year invasion window, as well as the shortest lead and lag lengths with the earliest possible invasion date were then compared

between taxa, platforms and countries using ANOVA on a GLMM (family = “gaussian”, link = “identity”) including country, taxon, habitat and internet platform as fixed effects and species as random intercept.

For the lumped data activity increase detection, normalized activity trends were summed on a monthly basis across all platforms and their behaviour characterized during the 2-year invasion window using Generalized Additive Models (GAM)(bs = “tpsr”, k = 9)(Wood, 2015) to verify if combining activity data from all internet platforms improved predictability of known IAS invasions. More specifically, for every of the 112 individual known EU invasion scenarios between 2016 and 2025, we detected regions in the summed time series where either the first or second derivative showed a positive increase greater than a relative threshold (based on the total variation in observations) as well as where variance increased significantly. Per time series, only the region exhibiting the first region with a strong positive first derivative was retained. This approach prioritizes maximal instantaneous activity increases over sustained or accelerating trends, as we consider the steepest rate of increase to be the clearest and most important indicator of an immediate, anomalous activity spike. For every invasion scenario out of the 112 possible, we then verified if this region with the strongest slope and variance increases (if any) corresponded with the two-year invasion window surrounding the first of January of the EASIN reported year of invasion. Based on this ruleset, detections were ultimately classified as true positive (strong statistical signal of increased internet activity detected during invasion window), false positive (strong statistical signal of increased internet activity detected outside of the invasion window) or false negative (no significant change in variance together with first or second derivative detected across the entire timeseries).

3. Results

3.1. Mammals are the most popular species online in the EU

Our data mining exercise returned a total of 183 247 months with some online activity for the 88 species of union concern across all internet platform and country combinations (excluding the reference GBIF data). Popularity varied between platforms and was clearly skewed towards a few platforms with Wikipedia being the most popular platform on average across all unique species × country combinations (Fig. 2a; $P < 0.001$, Table S4). Moreover, in line with our expectations, Mammals were more popular on average across all platforms and countries compared to plants (Fig. 2b; $P = 0.01$; Table S5). These trends were also reflected in the rankings of individual species, with more mammals and fewer plants coming out on average as the most popular across all internet

platform × country combinations on average (Fig. S2 & S3). On average, the investigated species gained much more traction online in West-European countries (Fig. S4; $P < 0.001$; Table S5).

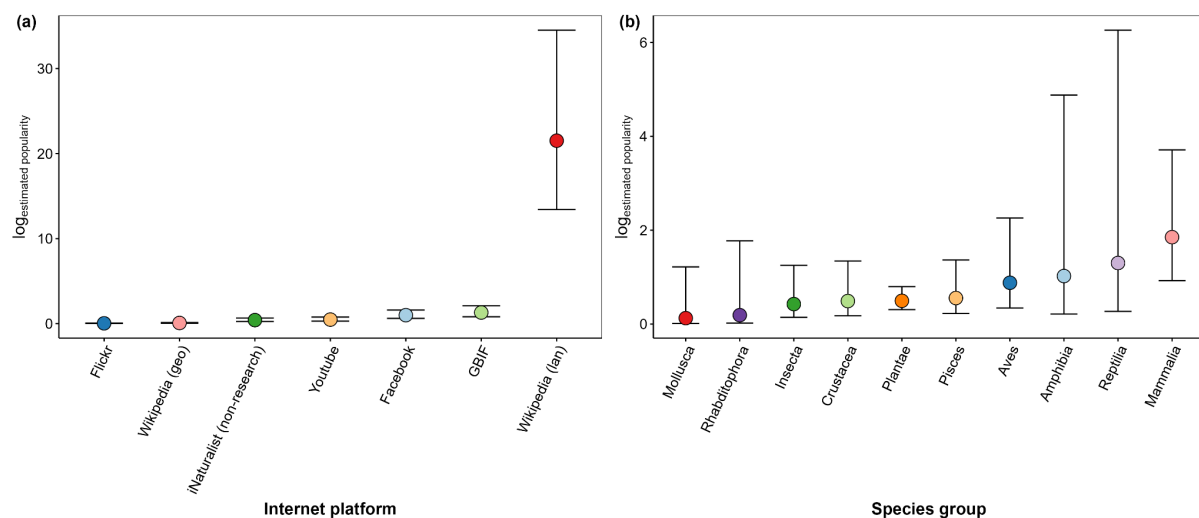


Figure 2. Estimated mean popularity per a) internet platform and b) species group (excl. GBIF). Error bars indicate $\pm 95\%$ CI on the modelled mean per species. Wikipedia (geo) and Wikipedia (lan) refer to geolocated and language-filtered Wikipedia pageviews.

3.2. Internet activity of less popular species better reflects real-world GBIF occurrences across the EU

When verifying if internet activity trends reflected country-level occurrence patterns on GBIF across all queried species and country combinations, we observed that also here Wikipedia language filtered pageviews performed best (Fig. 3a; $P < 0.001$; Table S6), despite on average relatively small Spearman's rho (ρ) values and large variation across all species × country observations for all platforms (Fig. 3a). Interestingly, patterns in iNaturalist non-research grade observations correlated the second best with GBIF occurrence data on average (Fig. 3a; $P < 0.001$; Table S6). Moreover, despite Mammalia being amongst some of the most popular species groups across internet platforms (Fig. 2b), their activity patterns aligned rather poorly with GBIF observations on average (Fig. 3b). In fact, insect activity patterns resembled most closely the available GBIF occurrence data on average – though driven by one species (*Vespa velutina*) –, and the per species correlation data also showed that 8 out of the 10 species with strongest correlations on average were Plantae, indicating that their internet activity patterns most closely resembled real-world country-level occurrences across all available data (Fig. S5). These per species rankings also indicated a tendency for internet activity to reflect trends in GBIF occurrences poorly for mammals and comparatively strongly for plants (Fig. S5), contrasting with the popularity trends (Fig. 2b). However, the per country trends were generally comparable to their popularity (Fig. S6; $P < 0.001$; Table S6).

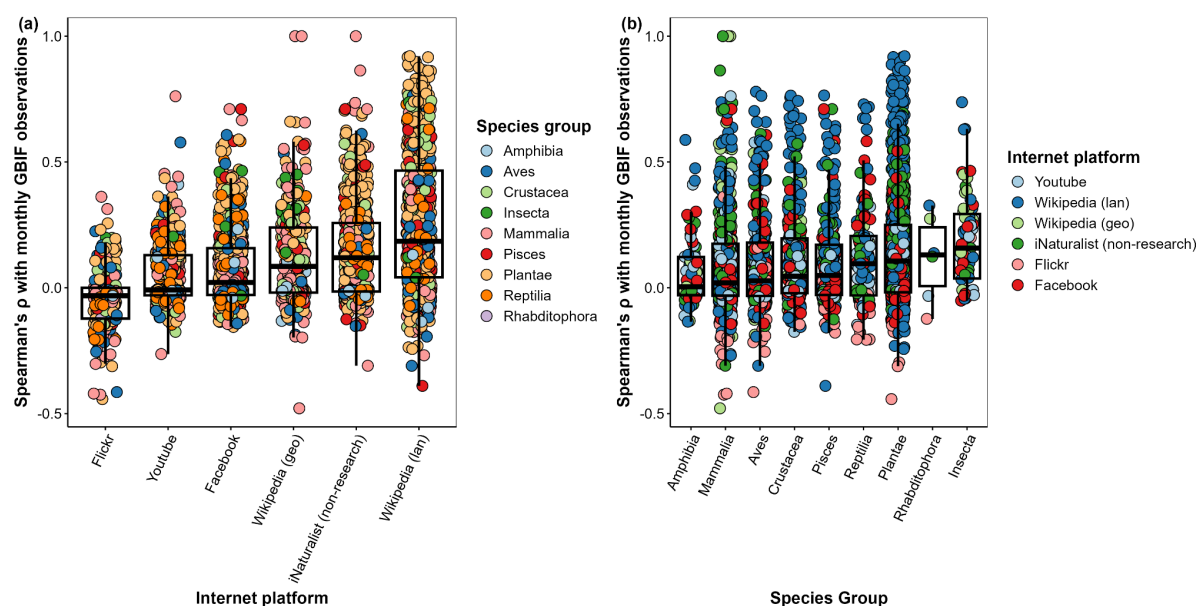


Figure 3. Spearman's rho (ρ) values and boxplots of correlations between monthly summed GBIF observations and unique species $\times$ country combinations a) per platform and b) per species group. Observation points are jittered for improved clarity. Wikipedia (geo) and Wikipedia (lan) refer to geolocated and language-filtered Wikipedia pageviews.

3.3. Varying capacity for systematic detection of invasions from EU internet activity across species groups

Next, we explored to what extent internet activity increased during the 2-year invasion window surrounding the 112 reported EASIN invasions of 85 IAS from the list of invasive alien species of Union concern since 2016 (see e.g., Fig. S7 for the EU invasion timeline of *Pontederia crassipes*). Out of all potential species $\times$ platform combinations (excl. GBIF) where the queries returned sufficient data to be analyzed statistically, 45% ($N = 275$) showed an anomalous increase in online activity in the 2-year window surrounding the EASIN reported year of invasion, with proportionally most anomalies detected within (geolocated) Wikipedia and Facebook activity (60%, $N = 25$ and 56%, $N = 81$; Fig. 4a) and the least for iNaturalist non-research grade observations (26%, $N = 43$; Fig. 4a). Despite large variation between observations across species and countries, the closest recorded anomalous increases in internet platform activity fell on average within the year of the first record for all platforms except iNaturalist ($\text{DOY} = 38 \pm 178$ (SD); Fig. 4b), with the reference GBIF dataset unsurprisingly being the most sensitive regarding detection of anomalous activity increases surrounding the year of invasion (69%, $N = 78$; Fig. 4a,b). However, anomalous data points were often also detected outside of the invasion window, leading to high false positive rates in many invasion scenarios (Fig. S8).

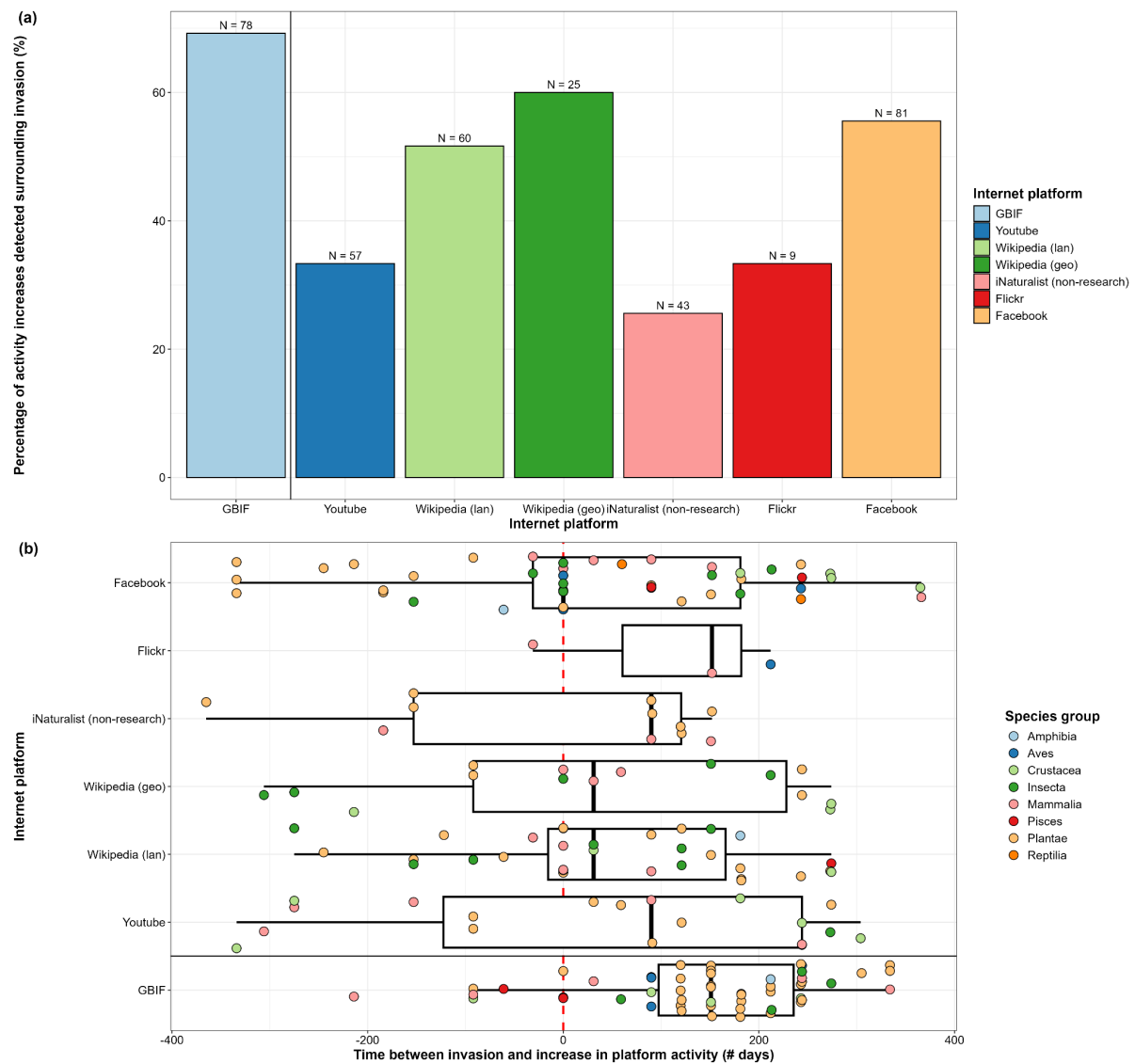


Figure 4. Anomaly detection data a) expressed as percentage of detections per platform and b) expressed as leads or lags from the earliest possible invasion moment in number of days per platform. In a) the total number of potential species × country combinations with sufficient data for anomaly detection per platform are indicated as N above each bar. The vertical red dashed line in b) indicates the first of January of the EASIN reported year of invasion. Observations are jittered for improved clarity. Wikipedia (geo) and Wikipedia (lan) refer to geolocated and language-filtered Wikipedia pageviews.

To tackle the issue of anomalous internet activity spikes outside of invasion windows inflating false positive rates on a per platform time series basis (see e.g., Fig. S9), we pooled internet platform data together and performed activity increase detection on the cumulative sum of normalized monthly platform activity data for each species × country combination. When pooling data and looking at combined slope and variance increases rather than isolated anomalies, overall true positive

detection rate was 45% (N = 112; Fig. 5a, b). However, while the first strong signal of increasing internet activity overlapped with the invasion window for many species × country combinations, our detection algorithm still returned 31% false positives (N = 112; Fig. 5b), due to comparatively stronger activity increase signals in non-target regions of the respective time series. In addition, for 24% of the explored scenarios, no strong activity increases were detected throughout the entire time series (N = 112; Fig. 5b). Among taxonomic groups, true positive detection rates (TPR) were greater than false positive rates for Aves (TPR: 80%, N = 5; Fig. S7), Pisces (TPR: 75%, N = 4; Fig. S7), Reptilia (TPR: 67%, N = 3; Fig. S7), Crustacea (TPR: 57%, N = 14; Fig. S7) and Plantae (TPR: 25%, N = 64; Fig. S7), equal to false positive rates for mammals (TPR: 36%, N = 11; Fig. S7) but smaller compared to false positive rates for Insecta (TPR: 10%, N = 10; Fig. S7).

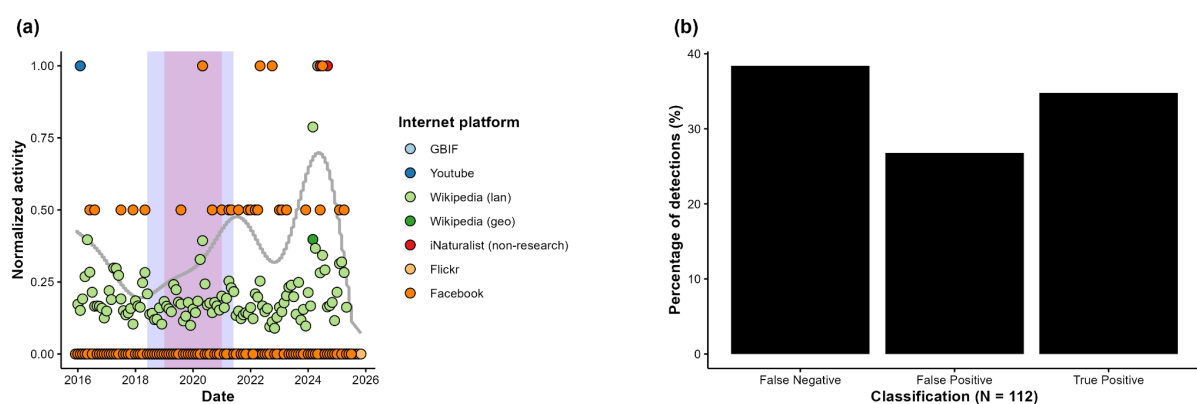


Figure 5. Example of a) the mean fitted GAM predictions (grey line) on the sum of normalized monthly activities for the combined internet platform activity data (excl. GBIF) of *Pontederia crassipes* in Poland (see also Fig. S4). The red box outlines the 2-year invasion window (reported EASIN invasion year in 2020) and the purple box indicates the first region of strong activity increase in the summed normalized data timeseries based on our pre-defined threshold. GBIF data was not used in the models but is shown for illustrative purposes. In b) the sensitivity of the detection algorithm is summarized across the total of 112 unique EASIN reported invasion cases of species of the Union list of concern between 2016 and 2025.

4. Discussion

Tracking online activity (mentions / searches) regarding IAS across internet platforms shows much potential to accelerate detection of novel species invasions (Daume, 2016; Jarić et al., 2021). We explored whether combining geolocated activity data (posts, mentions, videos, pageviews, searches) on IAS from different internet platforms (Wikipedia, Youtube, iNaturalist, Facebook, Flickr) could be used as a means to systematically identify new invasions across the EU. Out of all platforms, Wikipedia and Facebook were the most popular (i.e., had the most unique months with non-zero

activity) for the explored species × country combinations. However, despite an abundance of IAS activity data across various internet platforms, data availability and quality varied strongly across species and platforms, with a tendency for bias towards sustained interest for more active and often considered charismatic species groups (i.e., mammals) but more accurate reflections of real-world occurrence patterns for less popular species such as plants. When exploring recent early invasion scenarios across the EU, we found that Wikipedia followed by Facebook were the most useful platforms for detecting new species invasions, detecting anomalous activities in about 60% of cases with sufficient data. However, even when pooling data together for all platforms, false positives were still flagged in approximately one third of the invasion scenarios, with the highest likelihood of detecting true IAS invasions (> 50% of cases with clear activity signal) from internet activity data for birds, fish, reptiles and crustaceans versus the lowest for plants, mammals, insects and amphibians (< 50% of cases with clear activity signal).

Contrary to previous case studies (see e.g., Jarić et al. (2021)) and our first hypothesis, correlations between internet platform activity data and real-world GBIF occurrences per country were overall relatively weak for all platforms. Despite that some platforms such as Wikipedia, iNaturalist and Facebook showed a tendency to correlate better with real-world GBIF occurrences – likely related to persistent identifiers aggregating activity data per species topic rather than individual keywords or tags such as in other platforms –, their mined data was still of highly variable quality. Some species did show strong correlations with real-world occurrences across platforms and countries (e.g., *Microstegium vimineum* or *Vespa velutina*), but these were exceptions rather than the rule. Nonetheless, internet activity data relating to plants showed a tendency to correlate better with occurrences mined from GBIF compared to mammals, despite the latter being considerably more popular across all platforms. One potential explanation for this observation could be that internet activity data of more popular species are inflated with false positives, leading to less accurate representations of real-world occurrence patterns. Given these inconsistencies and the inability to manually verify observations with such big data volumes, these results highlight the urgent need for generalizable, open-source text and image based machine-learning tools that allow verification of rare invasive species identity with reasonable certainty from internet activity data, similar as to what has been already developed in more specific contexts (e.g., Cardoso et al., (2024)).

Interestingly, only three internet platforms (Flickr, Youtube and iNaturalist) allowed mining of IAS images, videos and occurrences with exact geolocated coordinates. Whilst the use of geolocated internet activity data without anonymization is problematic within the public space due to conflicts with privacy legislation (e.g., GDPR) and the use of internet platform data outside of the user-consented context (Di Minin et al., 2021; Sandbrook et al., 2021; Thompson et al., 2021; Xue,

2024), the lack of real coordinates associated with most mineable iEcology observations hampers direct integration into secure databases that could be used in specific national or international contexts where research regarding environmental issues and public health (i.e., matters of ‘public interest’) outweigh privacy protection rules. Thus, the utility of using automated iEcology data harvesting workflows to collect primary species observations remains overall likely limited to only a handful of platforms.

In that regard, although our research questions fall under the assessments of risks from digital data outlined in the EU’s digital service act (EU, 2022), and that we ensured secure and anonymized data processing and publication practices, multiple platforms still rejected our researcher access API applications (X, Google Ads, Google Search) or only allowed us to extract already anonymized (i.e., to country level), normalized or summarized activity data points (Facebook; Vardi et al. (2021)). As such, not only has internet platform data quality declined significantly over the past few years due to exceedingly greater numbers of bot accounts, users deliberately pushing misinformation and manipulative algorithms maximizing profit and political gain over content moderation (Koo et al., 2025; Ye et al., 2025), but the access to such data has also been significantly restricted (Hase et al., 2024). Two recent examples of such restrictions are directly impacting prominent iEcology and culturomics data sources. First, is the loss of researcher access to X’s data, where iEcology researchers used to have access to through the X’s API in the past (Novoa et al., 2022; Tomojiri & Takaya, 2025). Second, is the phasing out of popular Python (*Pytrends*, 2025) and R (Massicotte & Eddelbuettel, 2022) libraries for the automated mining of Google Trends data, since this feature is being integrated into the Google cloud system as a dedicated API with (currently) unknown researcher accessibility options (Google, 2025).

Our second hypothesis was confirmed to some extent. In approximately half of the explored scenarios that yielded sufficient data, there were spikes of platform activity surrounding the first report of the invasion of an IAS into a new EU country. However, a large slice of the tested scenarios did not return any useful data across platforms. In line with the popularity and the correlations with GBIF records, anomaly detections were most informative for Wikipedia, Facebook and Youtube, with Facebook returning data for most species × country combinations out of any platform including GBIF (81 out of 112), despite much of these combinations showing no spikes of activity during the actual invasion window. In addition, our results suggest that the occurrence datasets mined from GBIF should not be used as the sole data source for early warning systems: only 78 out of 112 species × country combinations yielded data and we could not capture anomalous occurrence signals after the IAS were first detected in each country in about 30 % of the cases with sufficient data. Moreover, while many platforms detected useful anomalous activity increases during the invasion windows,

most internet platform activity time series also showed anomalies activities in non-target regions, inflating the number of false positives massively and limiting the use of such an approach within early warning systems. Despite these limitations, citizen science platform data mined through GBIF has shown its use for capturing first records of invasive species, albeit inconsistently across countries and taxa (González-Moreno et al., 2025).

To tackle the issue of numerous false positives, we pooled activity data together and verified if this improved invasion detection capabilities. Overall, this method indeed improved overall true positive to false positive rates, despite still returning many false negatives. As such, the likelihood of flagged anomalous activity windows reflecting real-world invasions became somewhat higher. However, the ability of our algorithm to detect real invasion windows still differed between species groups. Interestingly, despite mammals being among the most popular species across platforms, our algorithm only detected a few real invasion windows, while the processing of fish, crustaceans and reptile invasion scenarios virtually only returned true positives. This observation is in line with how well the activity data seemed to reflect real world occurrences, further confirming that internet activity data on the most popular species often offers a worse representation of real world patterns (though with exceptions such as *Vespa velutina*).

Some shortcomings of the methodology used in this study should be highlighted. First, due to reasons relating to API rate and query limits (YouTube), disproportionate amounts of false positives requiring manual verification (Flickr) as well as reasons of practical implementation within VDE's (Facebook), most of our queries were limited to a keyword list only including the scientific names of the target species. Although this in all likelihood resulted in some missed 'real' IAS observations, we believe that the detected activity changes would still follow the same patterns as when including all synonyms and common names, given a single observation was enough to be classified as anomalous activity. Additionally, querying on scientific names assumes that users investigating the subject online have at least some intermediary level of knowledge about the species, increasing the likelihood of such observations actually reflecting the species of interest. Nonetheless, one of the reasons why Wikipedia returned disproportionate volumes of data compared to other platforms may be related to the fact that they work with unique identifiers linking up common names, synonyms and scientific species names within one topic, summing all activity relating to these topics together under one persistent umbrella. Given that species taxonomy changes on a regular basis and that keyword lists are per definition always finite and incomplete, the ability to extract activity data relating to a topic is a large benefit of this platform over others. Secondly, data was processed in ways that could influence the outcomes of this study. It is possible that binning observations to a monthly sum diluted important anomalous signals that occurred on a finer temporal resolution. However, binning

was necessary to allow comparison between all platforms, and we are quite confident that large anomalies were still visible, given that phenological signals were also still present in the monthly time series data. Thirdly, we considered the ‘invasion window’ to be the two-year period surrounding the earliest possible invasion date as reported by EASIN, which is somewhat an arbitrary middle point. As such, it is possible that some platform activity spikes were missed if e.g., the species only invaded towards the end of the second year and activity spiked with a lag. However, expanding these detection windows would go beyond investigating whether internet platform activity increases systematically during the early stages of real-world invasions. Finally, after having explored multiple options and benchmarked our thresholds, we chose to characterize our increases by looking at the first variance increase that coincided with positive increases in the first and second derivative within our time series (similar to the method proposed in Vanderhoeven et al., 2017). While the concept of ‘invasion fronts’ is commonly used in invasion biology (Arim et al., 2006), no studies thus far have explored whether such fronts are also systematically found in iEcology data. Given the more noisy nature of iEcology data per definition (Cardoso et al., 2024; Jarić et al., 2021), these criteria may thus not be the ultimate way of detecting alien species invasions from internet platform activity data.

5. Conclusion

We explored whether geolocated IAS internet platform activity patterns can be used as a systematic early warning system for detecting real-world species invasion across the EU. While our results indicate that online activity patterns of some taxonomic groups (fish, crustaceans, reptiles, birds and plants) reflect real-world early invasion patterns quite well, inconsistencies across species and countries currently hamper development of systematic detection algorithms and implementation in decision tools. Moreover, given the noisy nature of iEcology data, the recent trend across internet platforms to limit researcher access to well-documented databases through automated programming interfaces (Hase et al., 2024), likely worsens sensitivity of detection of real-world invasions due to the increasingly more limited access to geolocated online observations. To improve the utility of iEcology data for informing researchers and policy makers regarding invasive species expansions across the EU, future studies should (i) aim to develop more open, accessible and generalized invasive species machine vision models, (ii) work on systematic improvement of data processing algorithms that minimize retention of false positives across platforms, and (iii) push for improved researcher access to detailed internet platform data within the context of identifying systemic risks from online data as outlined by the digital service act.

6. Data availability statement

The data and code associated with this study is openly available in the Zenodo repository: TBA.

7. Acknowledgments

This study was supported financially by the EU horizon project OneSTOP (Grant agreement ID: 101180559).

8. Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

9. References

- Alam, A., & Hulme, P. E. (2025). Can Google Trends be used to estimate the geographic distribution of alien plants in the United States? *Lincoln University*.
- Allain, S. (2019). Mining Flickr: A method for expanding the known distribution of invasive species. *Herpetological Bulletin*, (148, Summer 2019), 11–14. <https://doi.org/10.33256/hb148.1114>
- Arim, M., Abades, S. R., Neill, P. E., Lima, M., & Marquet, P. A. (2006). Spread dynamics of invasive species. *Proceedings of the National Academy of Sciences*, 103(2), 374–378. <https://doi.org/10.1073/pnas.0504272102>
- Bates, D., Maechler, M., Bolker [aut, B., cre, Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., simulate.formula), P. N. K. (shared copyright on, Tanaka, E., Jagan, M., Boylan, R. D., & Ly, A. (2025). *lme4: Linear Mixed-Effects Models using “Eigen” and S4* (Version 1.1-38) [Computer software]. <https://cran.r-project.org/web/packages/lme4/index.html>
- Brooks, M., Bolker, B., Kristensen, K., Maechler, M., Magnusson, A., McGillicuddy, M., Skaug, H., Nielsen, A., Berg, C., Benth, K. van, Sadat, N., Lüdtke, D., Lenth, R., O’Brien, J., Geyer, C. J., Jagan, M., Wiernik, B., Stouffer, D. B., Agronah, M., ... Krieger, N. (2025). *glmmTMB*:

- Generalized Linear Mixed Models using Template Model Builder* (Version 1.1.13) [Computer software]. <https://cran.r-project.org/web/packages/glmmTMB/index.html>
- Cardoso, A. S., Malta-Pinto, E., Tabik, S., August, T., Roy, H. E., Correia, R., Vicente, J. R., & Vaz, A. S. (2024). Can citizen science and social media images support the detection of new invasion sites? A deep learning test case with *Cortaderia selloana*. *Ecological Informatics*, 81, 102602. <https://doi.org/10.1016/j.ecoinf.2024.102602>
- Cerri, J., Carnevali, L., Monaco, A., Genovesi, P., & Bertolino, S. (2022). Blacklists do not necessarily make people curious about invasive alien species. A case study with Bayesian structural time series and Wikipedia searches about invasive mammals in Italy. *NeoBiota*, 71, 113–128. <https://doi.org/10.3897/neobiota.71.69422>
- Cranswick, A. S., Constantine, R., Hendriks, H., & Carroll, E. L. (2022). Social media and citizen science records are important for the management of rarely sighted whales. *Ocean & Coastal Management*, 226, 106271. <https://doi.org/10.1016/j.ocecoaman.2022.106271>
- Dancho, M. (2023). *anomalize: Tidy Anomaly Detection* (Version 0.3.0) [R]. 10.32614/CRAN.package.anomalize
- 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>
- Di Minin, E., Fink, C., Hausmann, A., Kremer, J., & Kulkarni, R. (2021). How to address data privacy concerns when using social media data in conservation science. *Conservation Biology*, 35(2), 437–446. <https://doi.org/10.1111/cobi.13708>
- Diagne, C., Leroy, B., Vaissière, A.-C., Gozlan, R. E., Roiz, D., Jarić, I., Salles, J.-M., Bradshaw, C. J. A., & Courchamp, F. (2021). High and rising economic costs of biological invasions worldwide. *Nature*, 592(7855), 571–576. <https://doi.org/10.1038/s41586-021-03405-6>
- EU. (2022). *Article 34, the Digital Services Act (DSA)*. https://www.eu-digital-services-act.com/Digital_Services_Act_Article_34.html

Fukano, Y., & Soga, M. (2019). Spatio-temporal dynamics and drivers of public interest in invasive alien species. *Biological Invasions*, 21(12), 3521–3532.

<https://doi.org/10.1007/s10530-019-02065-y>

González-Moreno, P., Anđelković, A. A., Adriaens, T., Botella, C., Demetriou, J., Bastos, R., Bertolino, S., López-Cañizares, C., Essl, F., Fišer, Ž., Glavendekić, M., Herremans, M., Hulme, P. E., Jani, V., Katsada, D., Kleitou, P., La Porta, N., Lapin, K., López-Darias, M., ... Pocock, M. J. O. (2025).

Citizen science platforms can effectively support early detection of invasive alien species according to species traits. *People and Nature*, 7(1), 278–294.

<https://doi.org/10.1002/pan3.10767>

Google. (2025). *Introducing the Google Trends API (alpha): A new way to access Search Trends data* | Google Search Central Blog. Google for Developers.

<https://developers.google.com/search/blog/2025/07/trends-api>

Hase, V., Ausloos, J., Boeschoten, L., Pfiffner, N., Janssen, H., Araujo, T., Carrière, T., de Vreese, C.,

Haßler, J., Loecherbach, F., Kmetty, Z., Möller, J., Ohme, J., Struminskaya, B., Trilling, D.,

Welbers, K., & Haim, M. (2024, September 16). *Fulfilling data access obligations: How could (and should) platforms facilitate data donation studies?* [Info:eu-repo/semantics/article].

Alexander von Humboldt Institute for Internet and Society gGmbH.

<https://doi.org/10.14763/2024.3.1793>

Jarić, I., Bellard, C., Correia, R. A., Courchamp, F., Douda, K., Essl, F., Jeschke, J. M., Kalinkat, G.,

Kalous, L., Lennox, R. J., Novoa, A., Proulx, R., Pyšek, P., Soriano–Redondo, A., Souza, A. T.,

Vardi, R., Veríssimo, D., & Roll, U. (2021). Invasion Culturomics and iEcology. *Conservation*

Biology, 35(2), 447–451. <https://doi.org/10.1111/cobi.13707>

Kalous, L., Nechanská, D., & Petrýl, M. (2018). Survey of angler’s internet posts confirmed the

occurrence of freshwater fishes of the genus *Ictiobus* (Rafinesque, 1819) in natural waters of

Czechia. *Knowledge & Management of Aquatic Ecosystems*, (419), Article 419.

<https://doi.org/10.1051/kmae/2018019>

- Koo, G. H., Lukito, J., Masullo, G. M., Overgaard, C. S. B., & Orr, B. (2025). The consequences of “The Bird is Free”: A computational analysis of aversive LGBTQIA+ tweets and engagement trends before and after Elon Musk dismantled the platform’s moderation system. *New Media & Society*, 14614448251356240. <https://doi.org/10.1177/14614448251356240>
- Kotowska, D., Pärt, T., & Žmihorski, M. (2021). Evaluating Google Street View for tracking invasive alien plants along roads. *Ecological Indicators*, 121, 107020. <https://doi.org/10.1016/j.ecolind.2020.107020>
- Kourantidou, M., Haubrock, P. J., Cuthbert, R. N., Bodey, T. W., Lenzner, B., Gozlan, R. E., Nuñez, M. A., Salles, J.-M., Diagne, C., & Courchamp, F. (2022). Invasive alien species as simultaneous benefits and burdens: Trends, stakeholder perceptions and management. *Biological Invasions*, 24(7), 1905–1926. <https://doi.org/10.1007/s10530-021-02727-w>
- Kumschick, S., Bertolino, S., Blackburn, T. M., Brundu, G., Costello, K. E., de Groot, M., Evans, T., Gallardo, B., Genovesi, P., Govender, T., Jeschke, J. M., Lapin, K., Measey, J., Novoa, A., Nunes, A. L., Probert, A. F., Pyšek, P., Preda, C., Rabitsch, W., ... Bacher, S. (2024). Using the IUCN Environmental Impact Classification for Alien Taxa to inform decision-making. *Conservation Biology*, 38(2), e14214. <https://doi.org/10.1111/cobi.14214>
- Lenth, R. V., Piaskowski, J., Banfai, B., Bolker, B., Buerkner, P., Giné-Vázquez, I., Hervé, M., Jung, M., Love, J., Miguez, F., Riebl, H., & Singmann, H. (2025). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (Version 2.0.1) [Computer software]. <https://cran.r-project.org/web/packages/emmeans/index.html>
- López-Guillén, E., Herrera, I., Bensid, B., Gómez-Bellver, C., Ibáñez, N., Jiménez-Mejías, P., Mairal, M., Mena-García, L., Nualart, N., Utjés-Mascó, M., & López-Pujol, J. (2024). Strengths and Challenges of Using iNaturalist in Plant Research with Focus on Data Quality. *Diversity*, 16(1), Article 1. <https://doi.org/10.3390/d16010042>

Massicotte, P., & Eddelbuettel, D. (2022). *gtrendsR: Perform and Display Google Trends Queries*

(Version 1.5.1) [Computer software].

<https://cran.r-project.org/web/packages/gtrendsR/index.html>

Meta. (2025). *Get access—Meta Content Library and API - Documentation*. Meta for Developers.

<https://developers.facebook.com/docs/content-library-and-api/get-access/>

Morais, P., Afonso, L., & Dias, E. (2021). Harnessing the Power of Social Media to Obtain Biodiversity Data About Cetaceans in a Poorly Monitored Area. *Frontiers in Marine Science*, 8, 765228.

<https://doi.org/10.3389/fmars.2021.765228>

Novoa, A., Canavan, S., Jarić, I., Pipek, P., & Pyšek, P. (2022). Musk’s Twitter takeover jeopardizes culturomics. *Nature*, 612(7939), 211–211. <https://doi.org/10.1038/d41586-022-04361-5>

Novoa, A., Jarić, I., Pipek, P., & Pyšek, P. (2025). Culturomics and iEcology provide novel opportunities to study human and social dimensions of alien species introductions. *Trends in Ecology & Evolution*, 40(1), 18–26. <https://doi.org/10.1016/j.tree.2024.08.012>

Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2. <https://doi.org/10.3389/fdata.2019.00013>

Pinheiro, J., Douglas Bates, Saikat DebRoy, Deepayan Sarkar, Siem Heisterkamp, Johannes Ranke, & R Core Team. (2025). *nlme: Linear and Nonlinear Mixed Effects Models* (Version 3.1-168) [Computer software]. <https://cran.r-project.org/web/packages/nlme/index.html>

*Pytrend*s. (2025). <https://pypi.org/project/pytrend/>

R Core Team. (2024). *R: A language and Environment for Statistical Computing* (Version 4.4.2.) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>

Roy, H. E., Pauchard, A., Stoett, P. J., Renard Truong, T., Meyerson, L. A., Bacher, S., Galil, B. S., Hulme, P. E., Ikeda, T., Kavileveetil, S., McGeoch, M. A., Nuñez, M. A., Ordonez, A., Rahlao, S. J., Schwindt, E., Seebens, H., Sheppard, A. W., Vandvik, V., Aleksanyan, A., ... Ziller, S. R. (2024). Curbing the major and growing threats from invasive alien species is urgent and achievable. *Nature Ecology & Evolution*, 8(7), 1216–1223. <https://doi.org/10.1038/s41559-024-02412-w>

- Sandbrook, C., Clark, D., Toivonen, T., Simlai, T., O'Donnell, S., Cobbe, J., & Adams, W. (2021). Principles for the socially responsible use of conservation monitoring technology and data. *Conservation Science and Practice*, 3(5), e374. <https://doi.org/10.1111/csp2.374>
- Schifani, E., & Paolinelli, R. (2018). Los foros y las redes sociales ayudan a descubrir especies exóticas en Europa y monitorear su propagación: El caso de *Exaireta spinigera* (Wiedemann, 1830) (Diptera, Stratiomyidae) en la península italiana y Sicilia. *Graellsia*, 74(2), e079. <https://doi.org/10.3989/graellsia.2018.v74.213>
- Seebens, H., Blackburn, T. M., Dyer, E. E., Genovesi, P., Hulme, P. E., Jeschke, J. M., Pagad, S., Pyšek, P., Winter, M., Arianoutsou, M., Bacher, S., Blasius, B., Brundu, G., Capinha, C., Celesti-Grapow, L., Dawson, W., Dullinger, S., Fuentes, N., Jäger, H., ... Essl, F. (2017). No saturation in the accumulation of alien species worldwide. *Nature Communications*, 8(1), 14435. <https://doi.org/10.1038/ncomms14435>
- Soto, I., Courtois, P., Pili, A., Tordoni, E., Manfrini, E., Angulo, E., Bellard, C., Briski, E., Buřič, M., Cuthbert, R. N., Kouba, A., Kourantidou, M., Macêdo, R. L., Leroy, B., Haubrock, P. J., Courchamp, F., & Leung, B. (2025). Using species ranges and macroeconomic data to fill the gap in costs of biological invasions. *Nature Ecology & Evolution*, 9(6), 1021–1030. <https://doi.org/10.1038/s41559-025-02697-5>
- Thompson, R. M., Hall, J., Morrison, C., Palmer, N. R., & Roberts, D. L. (2021). Ethics and governance for internet-based conservation science research. *Conservation Biology*, 35(6), 1747–1754. <https://doi.org/10.1111/cobi.13778>
- Tomojiri, D., & Takaya, K. (2025). Aspects of public attention on the most mentioned nonindigenous species, as determined by a comprehensive assessment of Japanese social media. *Biological Invasions*, 27(1), 62. <https://doi.org/10.1007/s10530-024-03520-1>
- van Rossum, G. (2025). *Python Language Reference, version 3.13* (Version 3.13) [Computer software]. Python Software Foundation. <http://www.python.org>

- Vanderhoeven, S., Adriaens, T., Desmet, P., Strubbe, D., Backeljau, T., Barbier, Y., Brosens, D., Cigar, J., Coupremanne, M., Troch, R. D., Eggermont, H., Heughebaert, A., Hostens, K., Huybrechts, P., Jacquemart, A.-L., Lens, L., Monty, A., Paquet, J.-Y., Prévot, C., ... Groom, Q. J. (2017). Tracking Invasive Alien Species (TrIAS): Building a data-driven framework to inform policy. *Research Ideas and Outcomes*, 3, e13414. <https://doi.org/10.3897/rio.3.e13414>
- Vardi, R., Mittermeier, J. C., & Roll, U. (2021). Combining culturomic sources to uncover trends in popularity and seasonal interest in plants. *Conservation Biology*, 35(2), 460–471. <https://doi.org/10.1111/cobi.13705>
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>
- Wikimedia. (2025a). *Pageviews Differential Privacy—Current*. Pageviews Differential Privacy. https://analytics.wikimedia.org/published/datasets/country_project_page/00_README.html
- Wikimedia. (2025b). *Pageviews Differential Privacy—Current*. https://analytics.wikimedia.org/published/datasets/country_project_page/00_README.html
- Wood, S. (2015). *Package “mgcv”* (Version 1) [Computer software]. <http://cran.uib.no/web/packages/mgcv/mgcv.pdf>
- Xue, M. (2024). Assessing the Recreational Fishers and their Catches based on Social Media Platforms: Privacy and Ethical Data Analysis Considerations. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, 15(3), 521–542. <https://doi.org/10.58346/JOWUA.2024.I3.033>
- Ye, J., Luceri, L., & Ferrara, E. (2025). Auditing Political Exposure Bias: Algorithmic Amplification on Twitter/X During the 2024 U.S. Presidential Election. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25*, 2349–2362. <https://doi.org/10.1145/3715275.3732159>

Supplementary material

Table S1. Characteristics utilized to evaluate the suitability of different internet platforms for mining iEcology data through APIs. Text or numerical values, images and videos are abbreviated by txt, img and vid, respectively.

Platform	Data types	API Cost	Geolocation	Query limit	Temporal resolution	Mining authorized?
X (Twitter)	txt/img/vid	free*	full coordinates	NA**	any	No
Meta	txt/img/vid	free*	country	60/min	any	Yes
Google ads	txt	free*	country/region/city	NA**	monthly	No
Google health	txt	NA**	country/region/city	NA**	any	NA**
Google trends	txt	free*	country/region/city	NA**	hourly	No
Youtube	txt/img/vid	free	full coordinates	10k/day	any	Yes
Tiktok	txt/img/vid	free*	country/region/city	20k/day	any	No
Reddit	txt/img/vid	free	language based	100/min	any	Yes
Wikidata	txt	free	page dependent	unlimited	daily	Yes
Flickr	txt/img	free	full coordinates	60/min	any	Yes
Bluesky	txt/img/vid	free	language based	600/min	any	Yes
Mastodon	txt/img/vid	free	language based	60/min	any	Yes
iNaturalist	txt/img	free	full coordinates	60/min	any	Yes
eBay	txt/img	free	country	5k/day	live listings	Yes

*Platform specific limitations affecting utility apply.

**NA values indicate lack of relevant information about query limits due to poor documentation, de-activated APIs or rejected researcher API applications.

Table S2. The Alpha-2 ISO codes and country names of European countries included in the API data collection queries.

Alpa-2 ISO Code	Country name
AT	Austria
BE	Belgium
BG	Bulgaria
HR	Croatia
CY	Cyprus
CZ	Czechia
DK	Denmark
EE	Estonia
FI	Finland
FR	France
DE	Germany
GR	Greece
HU	Hungary

IE	Ireland
IT	Italy
LV	Latvia
LT	Lithuania
LU	Luxembourg
MT	Malta
NL	Netherlands
PL	Poland
PT	Portugal
RO	Romania
SK	Slovakia
SI	Slovenia
ES	Spain
SE	Sweden
IS	Iceland
LI	Liechtenstein
NO	Norway
CH	Switzerland
GB	United Kingdom

Table S3. Different European countries and their approximate coverage by different circles used for the Youtube API queries.

Country	Locations (Latitude, Longitude, Radius)
Austria	47.5162, 14.5501 (180km)
Belgium	50.5039, 4.4699 (120km)
Bulgaria	42.7339, 25.4858 (180km)
Croatia	45.1000, 15.2000 (200km); 43.5000, 16.5000 (100km)
Cyprus	35.1264, 33.4299 (70km)
Czechia	49.8175, 15.4730 (150km)
Denmark	56.2639, 9.5018 (180km); 55.0000, 10.0000 (80km)
Estonia	58.5953, 25.0136 (150km)
Finland	61.9241, 25.7482 (350km); 67.0000, 26.0000 (200km)
France	46.6034, 1.8883 (350km); 48.5000, 6.0000 (150km); 43.5000, -1.0000 (150km); 42.0000, 9.0000 (70km)
Germany	51.1657, 10.4515 (300km); 53.5000, 9.5000 (150km); 48.5000, 11.5000 (150km)
Greece	39.0742, 21.8243 (250km); 35.5000, 24.0000 (150km); 37.5000, 26.0000 (150km)

Hungary	47.1625, 19.5033 (150km)
Ireland	53.1424, -7.6921 (150km)
	41.8719, 12.5674 (300km); 45.5000, 10.5000 (150km); 38.5000, 16.0000 (150km);
Italy	40.0000, 9.0000 (120km)
Latvia	56.8796, 24.6032 (150km)
Lithuania	55.1694, 23.8813 (150km)
Luxembourg	49.8153, 6.1296 (50km)
Malta	35.9375, 14.3754 (20km)
Netherlands	52.1326, 5.2913 (100km)
Poland	51.9194, 19.1451 (250km); 53.5000, 22.0000 (150km)
Portugal	39.3999, -8.2245 (200km); 38.7000, -28.0000 (150km); 32.7000, -17.0000 (100km)
Romania	45.9432, 24.9668 (250km); 46.0000, 28.0000 (100km)
Slovakia	48.6690, 19.6990 (150km)
Slovenia	46.1512, 14.9955 (100km)
	40.4637, -3.7492 (400km); 42.5000, -7.5000 (150km); 37.0000, -4.0000 (150km);
Spain	28.5000, -16.0000 (150km); 39.7000, 2.9000 (100km)
Sweden	60.1282, 18.6435 (400km); 65.0000, 18.0000 (300km); 56.0000, 14.0000 (150km)
United Kingdom	54.0000, -2.5000 (300km); 51.0000, 0.0000 (150km); 57.0000, -4.5000 (200km);
	54.5000, -7.0000 (80km)
Switzerland	46.8182, 8.2275 (100km)
	60.4720, 8.4689 (400km); 67.0000, 14.0000 (300km); 70.0000, 24.0000 (200km);
Norway	59.0000, 10.0000 (100km)

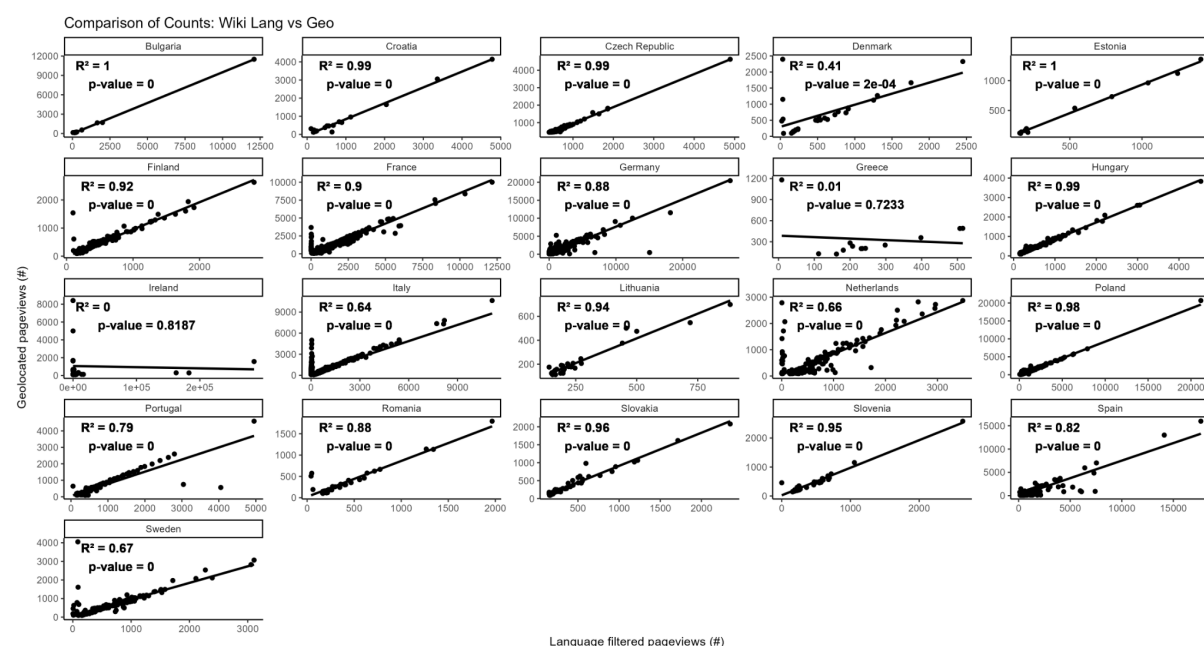


Figure S1. Correlations of non-NA observations between wikipedia pageviews per country based on language filtering vs direct geolocation (> 90 pageviews per day) for all species on the Union List of Concern.

Table S4. Type II ANOVA outcome table of the GLMM (family = negative binomial, link = ‘log’) testing for the effects of internet platform (incl. GBIF), taxon, country and habitat on popularity.

Response variable	Fixed effect	df	Chisq	P - value
Popularity	Internet platform	6	7870.825	< 0.001
	Taxon	9	16.981	0.049
	Country	50	3647.887	< 0.001
	Habitat	2	3.075	0.215

Table S5. Type II ANOVA outcome table of the GLMM (family = negative binomial, link = ‘log’) testing for the effects of internet platform (excl. GBIF), taxon, country and habitat on popularity.

Response variable	Fixed effect	df	Chisq	P - value
Popularity	Internet platform	5	8229.114	< 0.001
	Taxon	9	21.680	0.01
	Country	50	3081.641	< 0.001
	Habitat	2	2.641	0.267

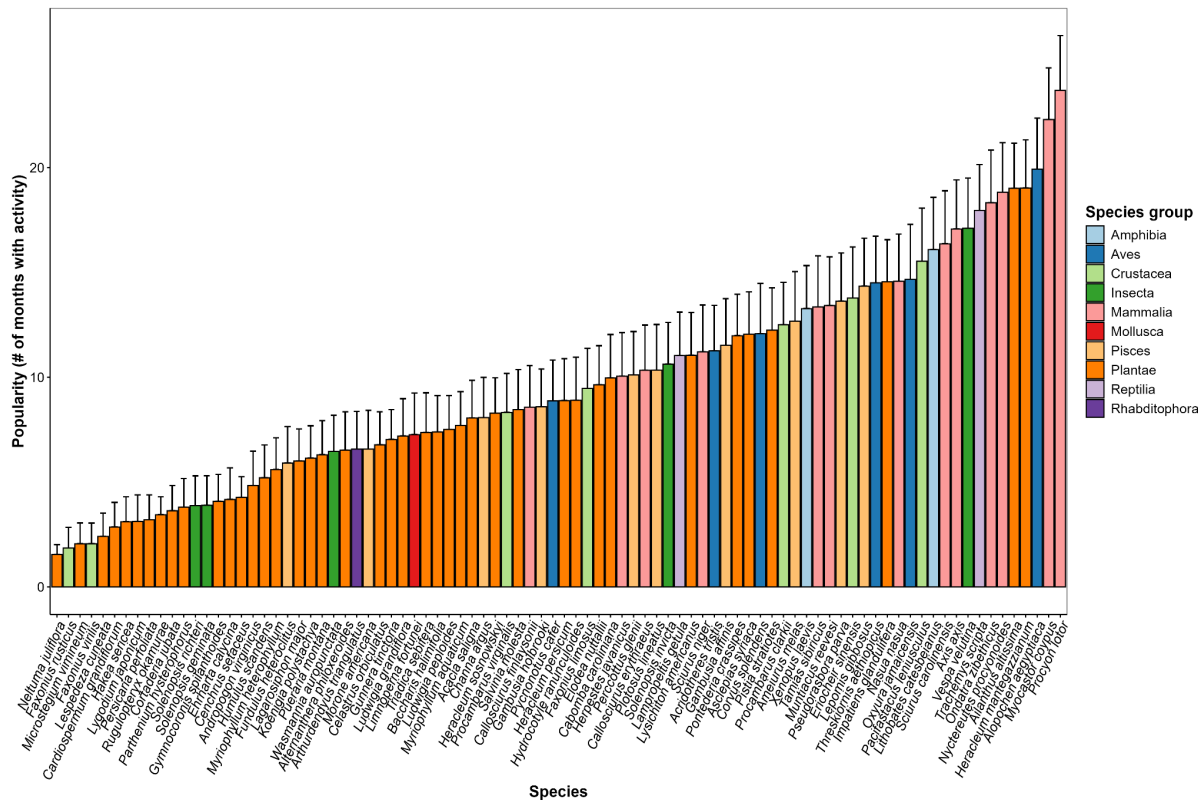


Figure S2. Average popularity of species across all internet platform x country combinations ranked by increasing popularity. Error bars represent ± 1 SE on the mean.

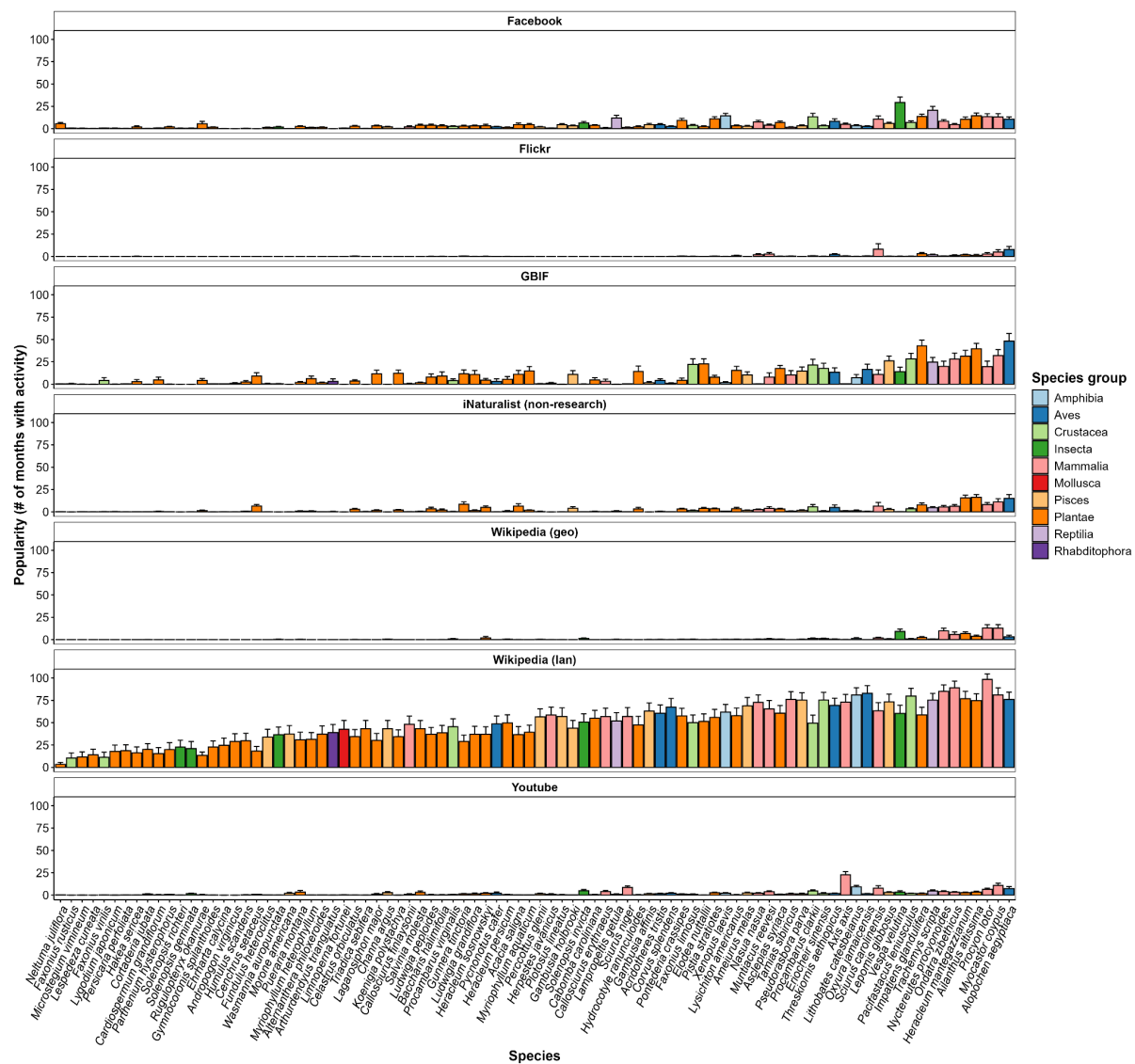


Figure S3. Average popularity of species per platform across all countries ranked by increasing mean popularity across all observations. Error bars represent ± 1 SE on the mean.

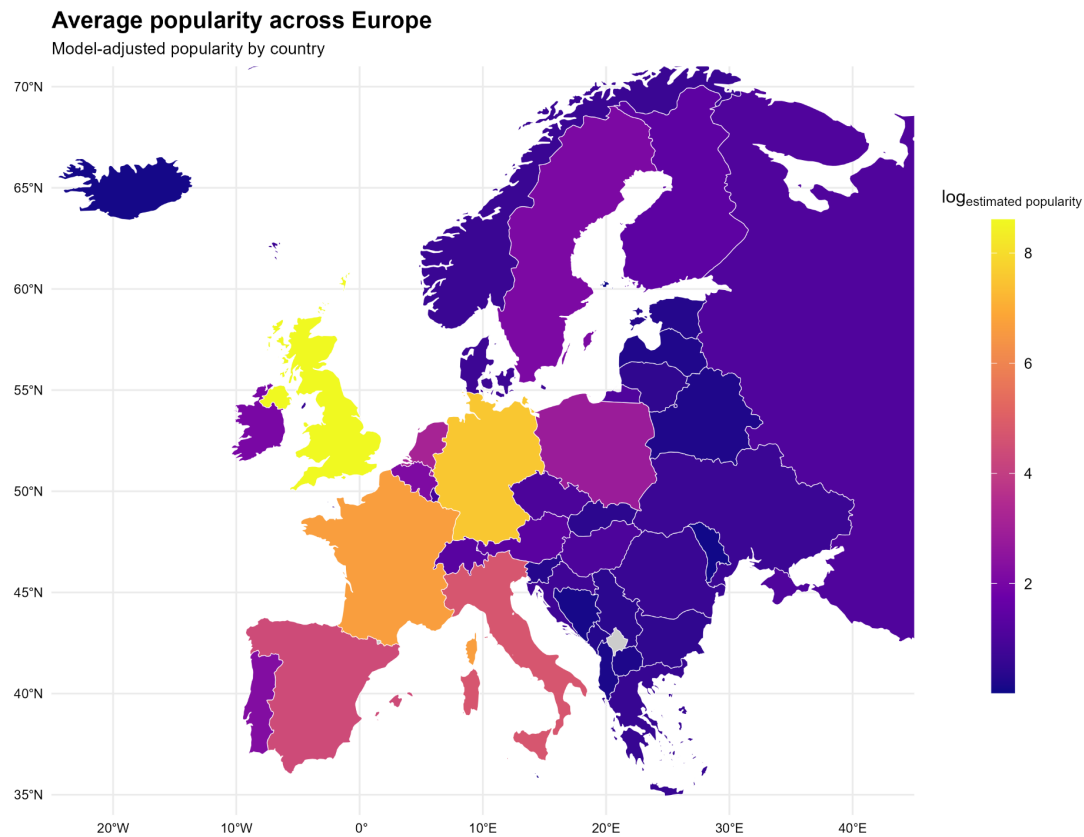


Figure S4. Map of modelled mean natural logarithm of popularity per country across all species.

Table S6. Type III ANOVA outcome table of the GLMM (family = gaussian, link = 'identity') testing for the effects of internet platform, taxon, country and habitat on rho value outcomes of the correlations between monthly internet platform activity data and GBIF occurrences.

Response variable	Fixed effect	df	F-value	P - value
Rho values	Internet platform	5, 3031	171.861	< 0.001
	Taxon	8, 61	0.7846	0.620
	Country	37, 3046	5.922	< 0.001
	Habitat	2, 54	2.108	0.132

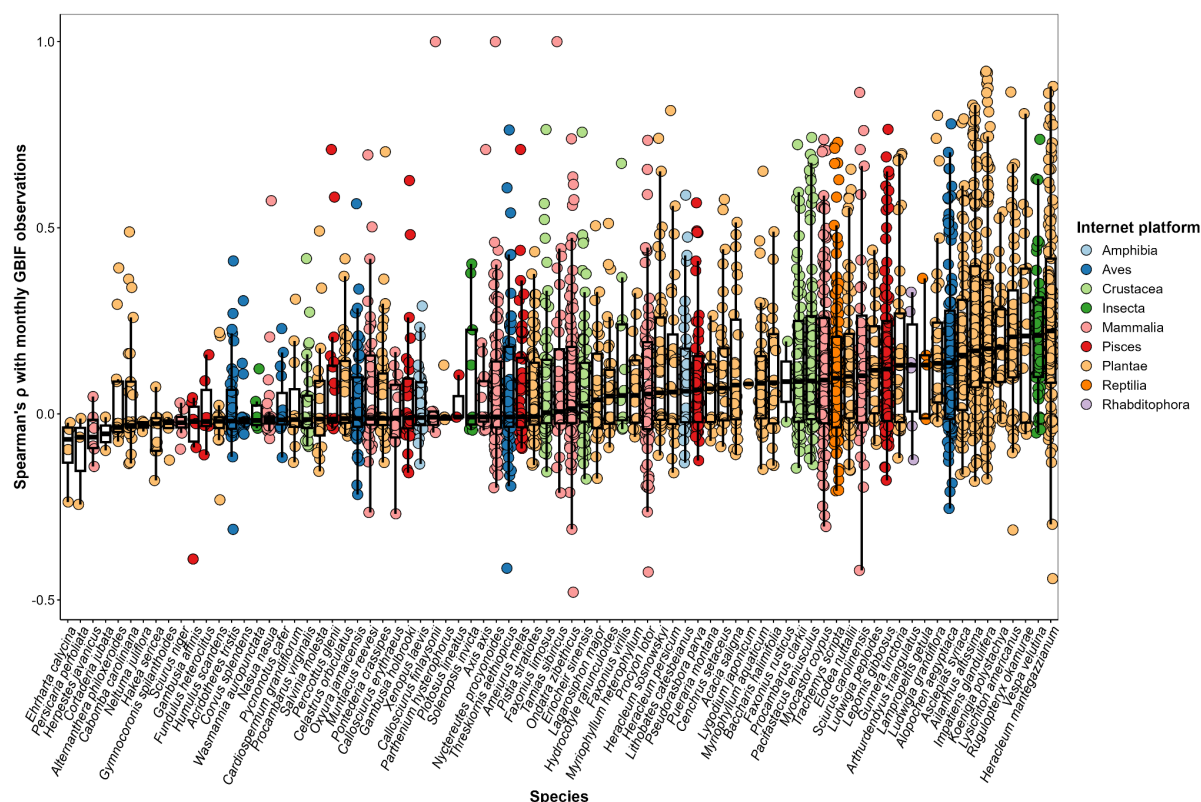


Figure S5. Spearman's rho (ρ) values and boxplots of correlations between monthly summed GBIF observations and unique platform x country combinations per species.

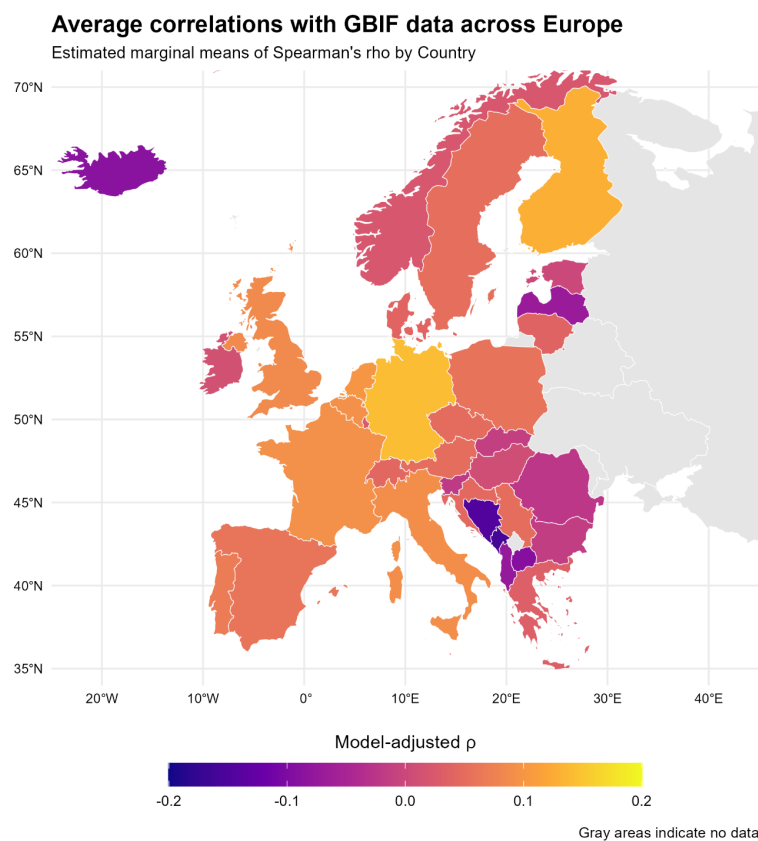


Figure S6. Map of Europe showing mean modelled Rho values of average monthly internet activity across all platforms and species with GBIF occurrences.

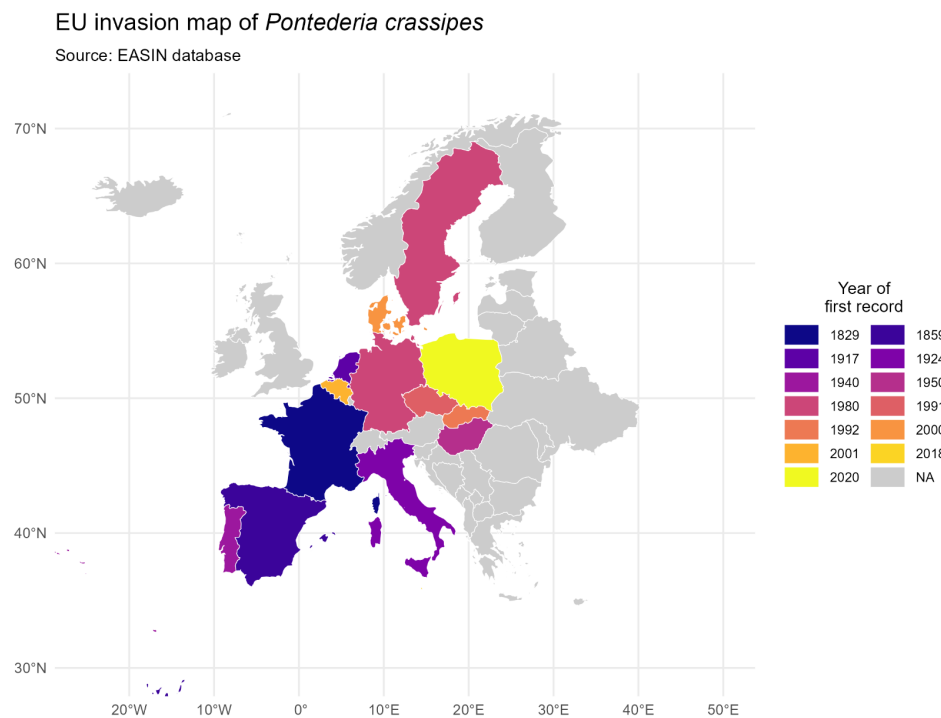


Figure S7. Invasion map of *Pontederia crassipes* across the EU with the EASIN reported invasion years highlighted. This species invasion in France from 2019 onwards was used as an example for the sensitivity analysis (see also section 3.3).

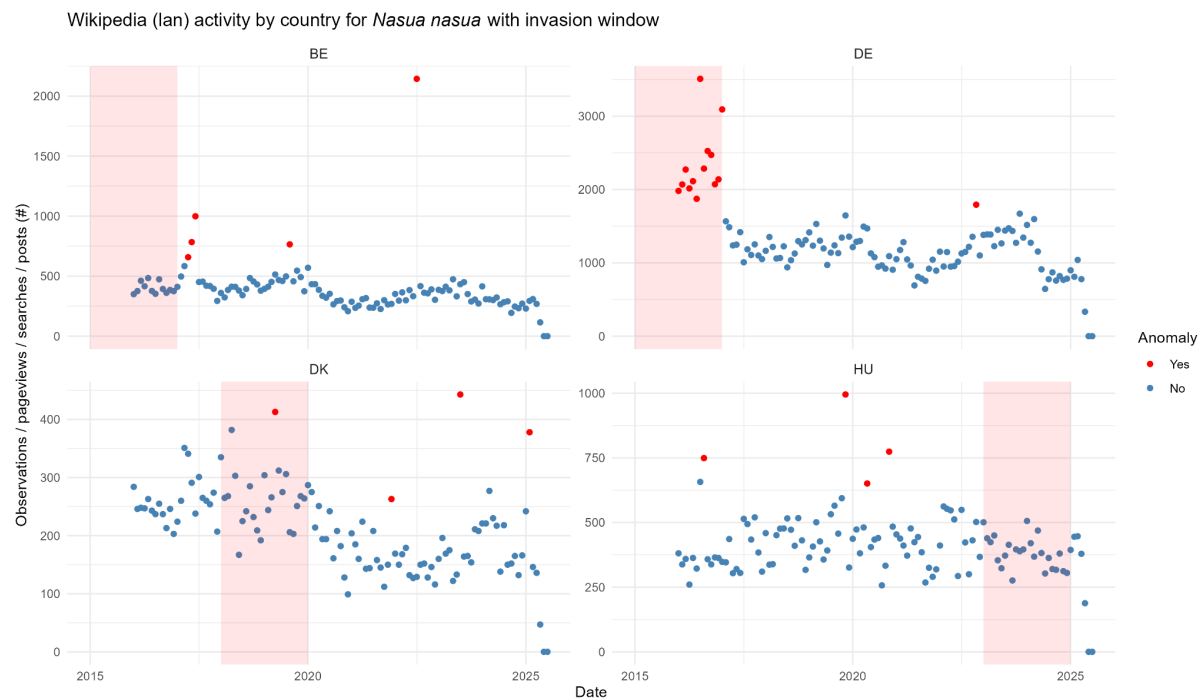


Figure S8. Wikipedia language filtered pageview timeseries and their anomalies highlighted in red for the 7 invasion scenario's of *Nasua nasua* in the EU since 2016. The ISO-3166-1 alpha-2 country codes

are given above the individual facets. The red box outlines the two-year invasion window including both the EASIN reported year of invasion and the year prior.



Figure S9. Sensitivity analysis of activity increase detection rates by taxonomic group.