NeoBiota 86: 151–168 (2023)
doi: 10.3897/neobiota.86.101471
https://neobiota.pensoft.net

R PACKAGE

A peer-reviewed open-access journal
NeoBiota
Advancing research on alien species and biological invasions

lydemapr: an R package to track the spread of the invasive spotted lanternfly (*Lycorma delicatula*, White 1845) (Hemiptera, Fulgoridae) in the United States

Sebastiano De Bona[1], Lawrence Barringer[2], Paul Kurtz[3], Jay Losiewicz[4], Gregory R. Parra[5], Matthew R. Helmus[1]

1 *Integrative Ecology Lab, Center for Biodiversity, Department of Biology, Temple University, 1925 N 12 Street, Philadelphia, PA, USA* **2** *Pennsylvania Department of Agriculture, Entomology Department, Bureau of Plant Industry, 2301 N Cameron Street, Harrisburg PA 17110, USA* **3** *New Jersey Department of Agriculture, Division of Plant Industry, PO Box 330, Trenton, NJ 08625, USA* **4** *Pennsylvania Department of Agriculture, Communications Office, 2301 N Cameron Street, Harrisburg PA 17110, USA* **5** *USDA APHIS PPQ Science and Technology, 920 Main Campus Drive, Raleigh, NC 27606, USA*

Corresponding author: Sebastiano De Bona (sebastiano.debona@gmail.com)

Academic editor: Maud Bernard-Verdier | Received 5 February 2023 | Accepted 12 June 2023 | Published 20 July 2023

Citation: De Bona S, Barringer L, Kurtz P, Losiewicz J, Parra GR, Helmus MR (2023) lydemapr: an R package to track the spread of the invasive spotted lanternfly (*Lycorma delicatula*, White 1845) (Hemiptera, Fulgoridae) in the United States. NeoBiota 86: 151–168. https://doi.org/10.3897/neobiota.86.101471

Abstract

A crucial asset in the management of invasive species is the open-access sharing of data on the range of invaders and the progression of their spread. Such data should be current, comprehensive, consistent and standardised, to support reproducible and comparable forecasting efforts amongst multiple researchers and managers. Here, we present the lydemapr R package containing spatiotemporal data and mapping functions to visualise the current spread of the spotted lanternfly (*Lycorma delicatula*, White 1841) in the Western Hemisphere. The spotted lanternfly is a forest and agricultural pest in the eastern Mid-Atlantic Region of the U.S., where it was first discovered in 2014. As of 2023, it has been found in 14 states according to State and Federal Departments of Agriculture. However, the lack of easily accessible, fine-scale data on its spread hampers research and management efforts. We obtained multiple memoranda-of-understanding from several agencies and citizen-science projects, gaining access to their internal data on spotted lanternfly point observations. We then cleaned, harmonised, anonymised and combined the individual data sources into a single comprehensive dataset. The resulting dataset contains spatial data gridded at the 1 km^2 resolution, with yearly information on the presence/absence of spotted lanternflies, establishment status and population density across 658,390 observations. The lydemapr package will aid researchers, managers and the public in their understanding, modelling and managing of the spread of this invasive pest.

Keywords
Biological invasions, crop pest, data science, forecasting, *Lycorma delicatula*, management, open access data, reproducibility, spread modelling

Introduction

Due to the globalisation of trade and the homogenisation of urban and suburban habitats, the accidental introduction and establishment of invasive species is ever more likely (Hulme 2009). When establishment goes undetected and eradication becomes less viable, the goal should be to mitigate the negative effects generated by invasive species (Diagne et al. 2020; Fantle-Lepczyk et al. 2022; Leroy et al. 2022). In doing so, one of the main challenges is tracking the spread of established invasive alien species so that control measures to slow spread, reduce impact and conserve biodiversity can be effectively enacted (Robertson et al. 2020). High quality data on past and present spread of invasives are key to model invasive spread accurately enough to provide robust forecasts on which to base management decisions.

A multitude of modelling techniques to forecast spread is available to researchers (Fisher 1937; Skellam 1951; Higgins and Richardson 1996; Kot et al. 1996; Neubert et al. 2000; Travis and Dytham 2002; Clark et al. 2003; Jongejans et al. 2011; Rodrigues and Johnstone 2014; Hudgins et al. 2017). Despite different assumptions and approaches to the modelling itself, fitting and validating models rely on longitudinal, spatially-explicit data on the occurrence or density of the spreading invasive species. Different models need to be built upon the same standardised data for comparisons between models to reflect genuine differences in model assumptions (e.g. Norberg et al. 2019). Comparing models with standardised data highlights which biological aspects of spread coded in each model are crucial to manage (Sakai et al. 2001). In addition, building models on the same data provides a more solid ground to combine them into ensemble models, which offer a higher degree of reliability compared to a single model (Araújo and New 2007). However, there are three hurdles that must be overcome before such standardised data for modelling be made available.

The first hurdle that must be overcome when developing a standardised dataset on invasive spread is to develop relationships with the agencies, institutions and citizen-science projects collecting data on the invasive of interest. For pests with negative impact on agricultural activity or forest habitats, local agencies, state departments and research institutions associated with the species first discovery are likely to operate data collection. If the pest is spreading across geopolitical boundaries, multiple organisations with different jurisdictions and areas of operation are likely to collect field data. In addition, easy-to-identify pests are likely to attract public attention and involvement, fostering the collection of citizen-science data (Dickinson et al. 2010; Catlin-Groves 2012; Sullivan et al. 2014; Kobori et al. 2016; Johnson et al. 2020; Norman-Burgdolf and Rieske 2021; Santaoja 2022). Obtaining access to the data often requires directly contacting the maintainer of the dataset in the relevant institution and obtaining memoranda-of-

understanding to use the data once shared. Each agency will follow unique data sharing agreements, which need to be discussed in-depth at this stage.

Once the data are obtained, the heterogeneity of the data collection protocols adopted by different agencies requires several additional steps to harmonise the survey results before they can be combined into a single dataset (Kelling et al. 2009). This second hurdle is often the most time-consuming and requires a high degree of eco-informatic skill in data handling and management (Michener and Jones 2012). Non-standardised data collection demands an in-depth understanding of the collection protocols used in order to match the information collected across different surveys (Hampton et al. 2013). For this reason, harmonisation often demands an active collaboration with the agencies that collected the data, to ensure the data are interpreted correctly, especially when surveys lack metadata (Jones et al. 2019).

The third hurdle is essential, yet not often acknowledged: data anonymisation. Calls to make scientific knowledge more accessible and transparent have pushed ecological data to be published alongside many scientific papers (Reichman et al. 2011). This process is paramount to improve collaboration and repeatability of scientific studies, although some limitations need to occur to ensure sharing open access data is done safely (Lindenmayer and Scheele 2017; Lunghi et al. 2019). One such limitation concerns data at high spatial resolution, the publication of which could infringe upon individual privacy and personal interests (Zipper et al. 2019). Due to this, invasive spread data need to be carefully and fully anonymised to ensure stakeholders are protected and served. This is especially true when knowledge on the infested state of a property could cause its value to decrease or the value of the goods produced to be affected (Zhang and Boyle 2010; Kovacs et al. 2011). Anonymisation practices include the removal of personal information, as well as data handling that reduces the spatial resolution to an optimal compromise between conveying relevant information and safeguarding privacy.

The spotted lanternfly (*Lycorma delicatula*, White 1845; often referred to as SLF in literature) was first discovered in the United States in Berks County, Pennsylvania, in 2014 (Barringer et al. 2015; Dara et al. 2015) and, by 2023, spread to 14 states across the Northeastern, South-Atlantic and Midwestern United States (Urban et al. 2021; NYIPM 2023). This phloem-feeding planthopper is native to China and was likely introduced accidentally via a shipment of landscaping materials. The spotted lanternfly is known to feed on over 100 species of plants (Barringer and Ciafré 2020; Murman et al. 2020; Huron and Helmus 2022) and poses a major economic burden on viticulture as it feeds on grapevines reducing total yield and plant vigour (Urban 2020). There is a high risk of spotted lanternfly impacting the global wine market by spreading to areas like California and Europe (Huron et al. 2022).

State agencies and the United States Department of Agriculture (USDA) have collected large amounts of data on spotted lanternfly spread through field surveys. In addition, given the species is easily recognised and hard to misidentify, an extensive campaign to educate the public has promoted the collection of citizen-science data. Data are collected through individual use of well-established applications such as iNaturalist, which allow for users to record geo-referenced observations of wildlife sightings, as well as

through the use of applications developed *ad hoc* by State Departments of Agriculture to collect data on the spotted lanternfly. Given the variety of sources and the refinement of protocols for data collection, the data on this species are heavily heterogeneous. Currently, any research team analysing the spread of the pest has to invest a significant amount of time processing the data before using them in model construction and validation (Wakie et al. 2020; Cook et al. 2021; Huron et al. 2022; Jones et al. 2022; Ramirez et al. 2023).

Here, we describe the R package **lydemapr** (***Ly****corma* ***de****licatula* **map**ping in **R**), containing an up-to-date, fully anonymised and regularly refined, longitudinal, spatially-explicit dataset of spotted lanternfly records throughout the United States since its first discovery. The dataset includes information derived from field surveys and citizen-science observations and reports observed presence/absence of this invasive species in surveyed areas, as well as the presence of established populations and estimates of population density. In addition, the package contains tools to visualise the data by mapping them and to obtain summary tables of the dataset. The goal of this package is to provide a baseline for future modelling efforts to forecast the spread of the spotted lanternfly and to foster more effective collaboration between agencies and researchers. The **lydemapr** package was fully developed in R (R Core Team 2021) and is available as an online repository at https://github.com/ieco-lab/lydemapr.

Data and metadata

The dataset contained in the package represents an anonymised and condensed comprehensive record of data collected by several federal agencies, state agencies and citizen-science projects on the presence, establishment and population density of the spotted lanternfly in the United States (Fig. 1). Sources include the Departments of Agriculture for the States of Pennsylvania, Delaware, Indiana, Maryland and New Jersey; the New York State Department of Agriculture and Markets; the Virginia Department of Agriculture and Consumer Services; the Virginia Polytechnic Institute and State University; the United States Department of Agriculture; and public reporting from iNaturalist. The field data were collected through a variety of methods, including surveys aiming to estimate establishment status and spotted lanternfly population density, control actions to manage population through egg mass destruction and trapping of nymphs and adults and citizen-science observations collected through self-reporting or direct involvement through research projects. Self-reporting tools include two separate platforms developed by the Pennsylvania Department of Agriculture (PDA) in association with Penn State University (PSU) and the New Jersey Department of Agriculture (NJDA). In addition, we included data collected through an independent citizen-science projects of limited duration run by the Virginia Polytechnic Institute and State University and the Virginia Cooperative Extension.

At the date of this publication, the aggregated and anonymised dataset contained 658,390 individual observations pertaining to 61,715 point-locations throughout the United States collected between 2014 and 2021. These 61,715 point-locations represent centroids of a 1 km^2 grid at which the geospatial data were aggregated for

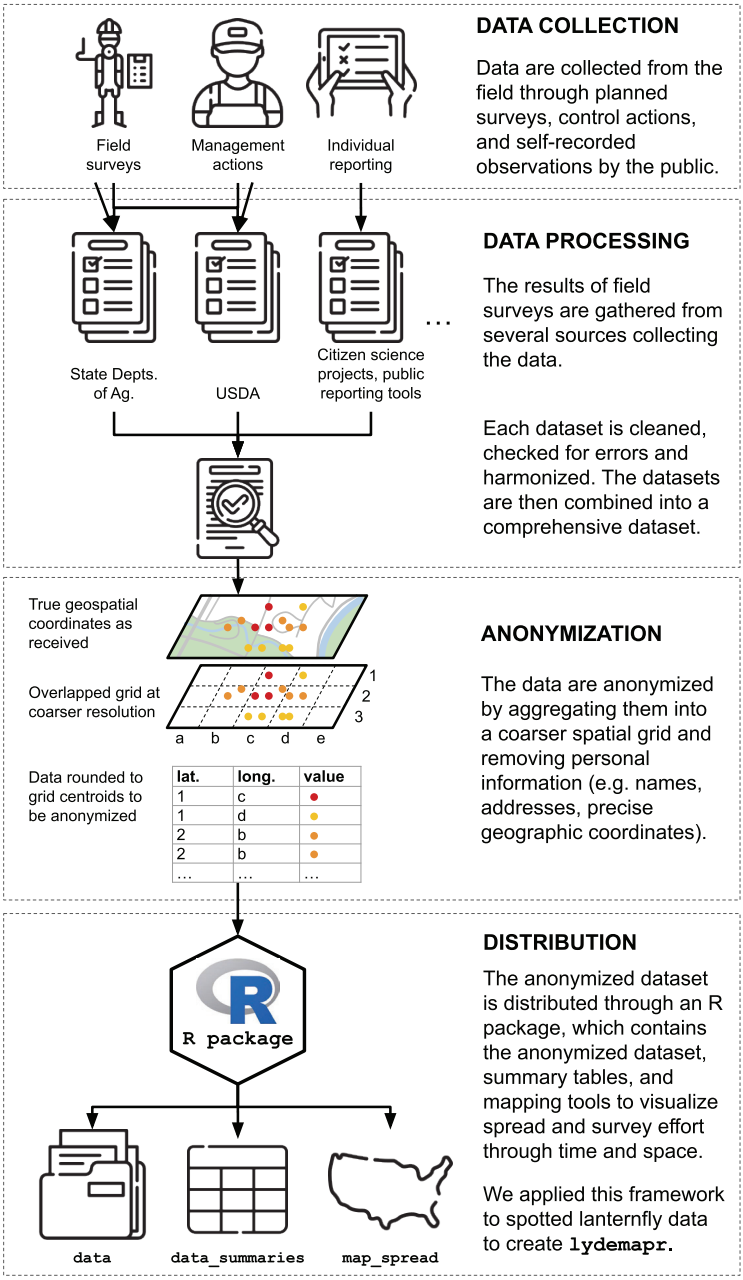

Figure 1. Conceptual graph describing the process leading to the distribution of the R package **lydemapr**. Data are collected by individual sources through multiple surveying processes. The datasets compiled this way are gathered from the sources and individually processed, then combined into a single comprehensive dataset. This is anonymised through both a censoring step and a spatial transformation to reduce spatial resolution. For the spatial transformation, latitude and longitude of individual survey points are rounded to the centroids of a 1-km^2 resolution grid. The aggregated and anonymised dataset is distributed through the package, together with functions to visualise the spread of the invasion through time.

anonymisation. The exact latitude and longitude of each survey contained in the geospatial data collected by the sources were rounded to the coordinates of the centroids. This approach, while removing the ability to derive property-level information from the dataset, allowed us to distribute survey-level information the data users can summarise as it best fits their needs. All variables containing traceable information regarding personal names, business names, contact information and comments were also removed from the dataset. The choice of 1 km^2 was agreed upon by all data contributing agencies to represent a compromise that provides high-resolution spatial data to enable precise spatial forecasting modelling while preserving privacy of the distributed data.

The individual observations recorded in the dataset derive from surveys and individual reporting conducted in 25 states across 8 years. The data points organised by year and state are summarised in Table 1. The distribution of data points by state is greatly skewed towards highly-impacted states. While Pennsylvania and the neighbouring states of Delaware, Maryland, New Jersey, New York and Virginia account for over 95% of data points (630,688 out of 658,390), other states in the western part of the country only account for a handful of surveys, mostly as a result of anecdotal reporting. Across time, it is easy to appreciate how surveying effort has increased, likely due to both the spread of lanternfly and to a higher investment of resources.

About 40% of the total data points were obtained through citizen-science projects; the well-established PDA and NJDA public reporting tools provided over 250,000 individual data points since 2019, while iNaturalist added just over 10,000 points. While management and surveying efforts led by state and federal agencies often focus on the leading edge of the invasion, where control actions are more effective, public reporting provides a constant and consistent source of data at the core. This helps the monitoring of these areas to be consistent and protracted in time, without subtracting important resources and work hours from managing the edge. In addition, iNaturalist provides constant, yet scattered, observations in areas where the surveying effort is not focused, as they are far from the invasion range. Those observations can then be confirmed by specialists during spatially-targeted surveys. The reliability of individually-reported records might vary with the experience and knowledge of the reporter. For this reason, in the dataset, records collected through citizen-science efforts are clearly distinct from records collected through expert-led surveys through the use of different categories under the variable "collection_method". This allows users of the data to only focus on records deriving from management and control actions, if necessary.

Data sets collection and processing

The goal for **lydemapr** is to update the dataset as new data become available and funding for the package is sustained. The plan is to request individual datasets periodically from federal and state sources, often coinciding with the termination of the biological season for spotted lanternfly (late spring, after eggs from the previous season are detected) or the temporary suspension of field operations (autumn-winter). Openly-available data (iNaturalist) are downloaded directly from the source at any time. To ensure we consider only agreed-upon, research-grade entries, the data are downloaded using the following query:

"search_on=names&quality_grade=research&identifications=most_agree&captive=false&place_id=1&taxon_id=324726".

Individual datasets pertaining to one-off collection efforts (e.g. the citizen-science project run by the Virginia Polytechnic Institute and State University) were obtained by contacting directly the data maintainer and are not updated unless the project itself is conducted again.

Individual datasets were processed in batches according to the data source. Each source had unique data collection methods which were generally consistent within a source although they did vary between years and across different data collection types (e.g. between visual surveys, control actions and trapping). Processing the data in batches first allowed us to harmonise individual datasets that shared similar, yet not identical, data structures, producing intermediate data tables that then were combined seamlessly into the final comprehensive dataset provided with **lydemapr**. There were five batches, corresponding to the five categories of the variable "source" (see section "Variables included"): PDA data, State data (consisting of data collected before 2020 from Delaware, Indiana, Maryland, New York and Virginia), public-reporting tool data, iNaturalist data and USDA data. Within each batch, the first step was to homogenise shared variables. This entailed the following steps:

Table 1. Data points by biological year and state (abbreviated).

State	2014	2015	2016	2017	2018	2019	2020	2021
AZ	-	-	-	-	-	10	139	100
CT	-	-	-	-	-	3	2081	1269
DC	-	-	-	-	8	21	10	4
DE	-	-	-	-	1075	2207	4545	5354
IN	-	-	1	-	79	101	102	352
KS	-	-	-	-	-	-	-	21
KY	-	-	-	-	-	3	2	18
MA	-	-	-	-	-	-	893	1835
MD	-	-	-	-	39	2404	17408	4600
ME	-	-	-	-	-	-	-	20
MI	-	-	-	-	-	-	1	133
MO	-	-	-	-	-	15	18	-
NC	-	-	-	-	-	14067	5	86
NJ	-	-	-	-	2443	9528	13066	83132
NM	-	-	-	-	-	-	10	28
NY	-	-	-	-	18474	27046	18255	4033
OH	-	-	-	-	-	-	731	406
OR	-	-	-	-	-	-	92	15
PA	370	7677	9269	9229	77047	150109	90390	61802
RI	-	-	-	-	-	-	45	18
SC	-	-	-	-	-	2	7	33
UT	-	-	-	-	-	-	1	-
VA	-	-	-	2	1523	4353	4099	1209
VT	-	-	-	-	-	-	-	2
WV	-	-	-	-	3	995	2367	1550

- ensuring coordinates are collected using the same projection or transforming them accordingly;
- homogenising date formats for all date variables;
- extracting year information and transforming it into "bio_year" (see the section "Variables included");
- tracking the source agency when merging individual datasets in batches;
- aggregating count data (where present), separately for eggs and nymphs/adult (necessary for a more accurate estimation of density);
- combining variables containing information on detection results (where present) and the aggregated count data into three final variables: "lyde_present", "lyde_established", "lyde_density". These variables define whether any sign of spotted lanternfly was detected, whether an established population was found and what the estimated population density at the site was, respectively (see the section "Variables included" for details on these variables). Some datasets (e.g. iNaturalist) only allow for the extraction of the presence of spotted lanternfly, omitting an assessment of establishment status and population density.

Once the shared variables were homogenised, they were renamed as they appear in the final version of the comprehensive dataset. We then generated an intermediate dataset from each batch, that contained only the shared variables (latitude, longitude, year, biological year, source agency, presence of spotted lanternfly, establishment status, population density), thus excluding all variables relating traceable information (personal names, business names, comments, addresses). Intermediate datasets were then combined together. During this step, the source was tracked through the appropriate variable. In addition, state information was added by intersecting point coordinates for each survey with state polygons (obtained through the package **tigris**) (Walker and Rudis 2023).

During a final cleaning step, we removed all data points not associated with a precise geolocation, a collection date (at least year) or a reference to the presence of the spotted lanternfly. After this, we shared the results as a high-resolution map with agency collaborators for a final check before distribution. Through this process, we were warned directly by the data providing agencies of potential mistakes, conflicts or suspicious data points. These problematic data points were vetted and corrected or removed.

The final step was the anonymisation process, where the precise location was summarised at a coarser 1 km^2 scale. This was done by creating a 1 km^2 grid over the spatial extent of the contiguous United States and intersecting this grid with the precise geolocation of each data point in the dataset. The coordinates of each point were replaced with the coordinates of the centroid of the 1 km^2 grid cell the point fell under. The process was repeated with an even coarser 10 km^2 grid, producing two additional variables added to the combined dataset, "rounded_latitude_10k" and "rounded_longitude_10k", which can be used to summarise and rarefy the dataset, if necessary, when visualising the data. After the anonymisation step, the resulting dataset **lyde** was saved and stored within the package.

Variables included

- **source:** character variable defining in broad terms the source of the data. "inat" for data obtained from iNaturalist, "PA" from data originating from the Pennsylvania

Dept. of Agriculture's surveying and management effort, "prt" for data collected through public reporting platforms, "states" for data collected by state-level agencies other than PDA, "USDA" for data provided by the United States Dept. of Agriculture. Note: the data originating from the Pennsylvania Dept. of Agriculture are kept separate from data collected by other states, as Pennsylvania was the state where the first introduction was detected. As a result of this, initial surveying efforts were led by this state, which collected the largest share of data early on;

- **source_agency:** character variable refining the definition of the source by indicating the agency/institution/project from which the data point was obtained: possible values are "iNaturalist", "PDA" (Pennsylvania Dept. of Agriculture), "NJDA_Public_reporting" (New Jersey Dept. of Agriculture's Public Reporting tool), "PDA_Public_reporting" (Pennsylvania Dept. of Agriculture's Public Reporting tool), "DDA" (Delaware Dept. of Agriculture), "ISDA" (Indiana State Dept. of Agriculture), "MDA" (Maryland Dept. of Agriculture), "NYSDAM" (New York State Dept. of Agriculture and Markets), "VDA" (Virginia Department of Agriculture and Consumer Services), "VA_Tech_Coop_Ext" (Virginia Polythecnic and State University/Cooperative Extension), "USDA";
- **collection_method:** character string defining the method used to collect data: "individual_reporting" for data collected through iNaturalist and public reporting tools and "field_survey/management" for data collected by agencies in the field. The accuracy and reliability of self-reporting data might be lower than that collected by field surveyors.
- **year:** integer value defining the calendar year when the information was collected;
- **bio_year:** integer defining the biological year when the information was collected. The biological year follows the species' development schedule and starts around the time of the emergence of first–instar nymphs (1 May–30 April);
- **latitude:** expressed in decimal degrees (WSG84 coordinate system);
- **longitude:** expressed in decimal degrees (WSG84 coordinate system);
- **state:** character defining the state where the data was collected (two-letter abbreviation, https://www.faa.gov/air_traffic/publications/atpubs/cnt_html/appendix_a.html);
- **lyde_present:** logical value defining whether records are present for spotted lanternfly at the site at the time of survey. These might include regulatory incidents where a single live individual or a small number of dead individuals were observed at the site, but no signs of established population could be detected;
- **lyde_established:** logical value defining whether signs of an established population are present at the site at the time of survey. These include a minimum of two alive individuals or the presence of an egg mass as per the working definition of establishment provided by the USDA;
- **lyde_density:** ordinal variable defining the population density of spotted lanternfly at the site, estimated directly as an ordinal category by the data collector or derived from count data. The categories include: "Unpopulated", indicating the absence of an established population at the site (but not excluding the presence of spotted lanternfly in the form of regulatory incidents); "Low", indicating an established population is present, but at low densities, reflecting at most about 30 individuals or a single egg mass; "Medium", indicating the population is established and at higher densities, but still at low enough population size to allow for a counting of the individuals during

a survey visit (a few hundred at most); "High", indicating the population is established and thriving and the area is generally infested, to a degree where a count of individuals would be unfeasible within a survey visit;

- **pointID:** character string uniquely identifying each data point;
- **rounded_longitude_10k:** longitude of the centroid of the closest 10 km^2 grid cell, expressed in decimal degrees (WSG84 coordinate system), used to rarefy the dataset at a coarser resolution;
- **rounded_latitude_10k:** longitude of the centroid of the closest 10 km^2 grid cell, expressed in decimal degrees (WSG84 coordinate system), used to rarefy the dataset at a coarser resolution.

Package installation and data access

The **lydemapr** package can be installed in two different ways. The public repository allows the user to install the package directly from GitHub, by executing the following command in a local R or RStudio instance: **devtools::install_github("ieco-lab/lydemapr", build_vignette = TRUE)**. This requires the package **devtools** (Wickham et al. 2022) and its dependencies to be installed locally. Alternatively, the package can be obtained by cloning the repository from the GitHub page https://github.com/ieco-lab/lydemapr. The package can then be installed locally by opening the file **lydemapr.Rproj** in RStudio and clicking "Install package" in the Build tab (or by executing the command **devtools::install()**). Once the package is installed, the user has access to the complete dataset, which can be loaded by typing **lydemapr::lyde** in the R console. In addition, the package contains a rarefied and summarised version of the same dataset at a lower spatial resolution (10 km^2), which can be accessed by typing **lydemapr::lyde_10k** instead. All information concerning package installation and data access is also available at the front page of the GitHub repository.

The R package structure allows us to update the dataset regularly as more data become available and if funding is obtained to support this initiative. In addition, a live GitHub repository grants us the ability to add functionalities and to improve the visualisation and summary tools included.

If the user is only interested in accessing the data without using the R package or is unfamiliar with R, all datasets contained in **lydemapr** are available for download through Zenodo (DOI: 10.5281/zenodo.7976229), where the user can download the data (in .csv format) and Metadata associated with it.

Package functions

For a summary overview of the data, the function **lyde_summary()** provides a breakdown of the dataset, showing the number of data points collected each year in each state where data have been collected (Table 1). The package contains two customisable

functions that can be used to visualise the data spatially. The function **map_spread()** provides an up-to-date map displaying the progression of the established invasion range through time, in addition to the locations of surveys which did not detect established populations (Fig. 2). Function arguments allow the user to select the spatial resolution at which the data should be mapped (choosing between 1 and 10 km^2) and the spatial extent of the figure produced. A second function included in the package, **map_yearly()** maps the findings of the survey efforts in terms of the species' population density. The visualisation is broken down by the year the surveys were conducted (Fig. 3). Through this visual depiction, it is possible to observe where survey efforts have been focusing on each year, as the invasion front progressed.

Conclusion

The dataset we provide on the spread of the spotted lanternfly, a high-impact forest and grapevine pest, will be useful in a variety of current and future efforts. Several models have been developed to forecast the future spread and establishment potential of spotted lanternfly in the United States and globally (Jung et al. 2017; Wakie et al. 2020; Huron et al. 2022; Jones et al. 2022; Lewkiewicz et al. 2022; Maino et al. 2022). Statistical forecasting models (e.g. Wakie et al. 2020; Huron et al. 2022; Jones et al. 2022) heavily rely on high resolution spatial data to derive future predictions. Leveraging this big data-set will allow new models to be developed and current ones to be refined and improved. On the other hand, mechanistic mathematical models (Lewkiewicz et al. 2022; Maino et al. 2022), despite building their predictions through a bottom-up approach that involves a deeper understanding of the species' own biology and ecology, require spatial data for validation and model tuning. To ensure future models can be compared and combined through ensemble procedures, these models should be based on the same historic and present spread data of spotted lanternfly, reaffirming the importance of a unified and readily available dataset.

From a management standpoint, a comprehensive data-set can provide additional information on population trends through time in specific areas, allowing for the expansion of current studies (Cook et al. 2021), as well as offering insight on the efficacy of control actions over time. In addition, our openly-accessible and comprehensive dataset has broad applications in education, to promote citizen-science initiatives in under-surveyed areas, but also to provide an opportunity for data science projects for students. As the issues related to the spread of invasive species are often issues students experience first-hand, working on this dataset can represent an engaging learning opportunity.

There were two unexpected challenges to creating the **lydemapr** dataset. One of the main challenges we encountered was the heterogeneity in the data collection methods. This challenge greatly inflated the time, effort and eco-informatic data-coding skills required to aggregate the data. The heterogeneity was greater in the first few years (until about 2019), when more and more agencies were becoming involved, but the coordination between them was low. To solve conflicts encounters when harmonising

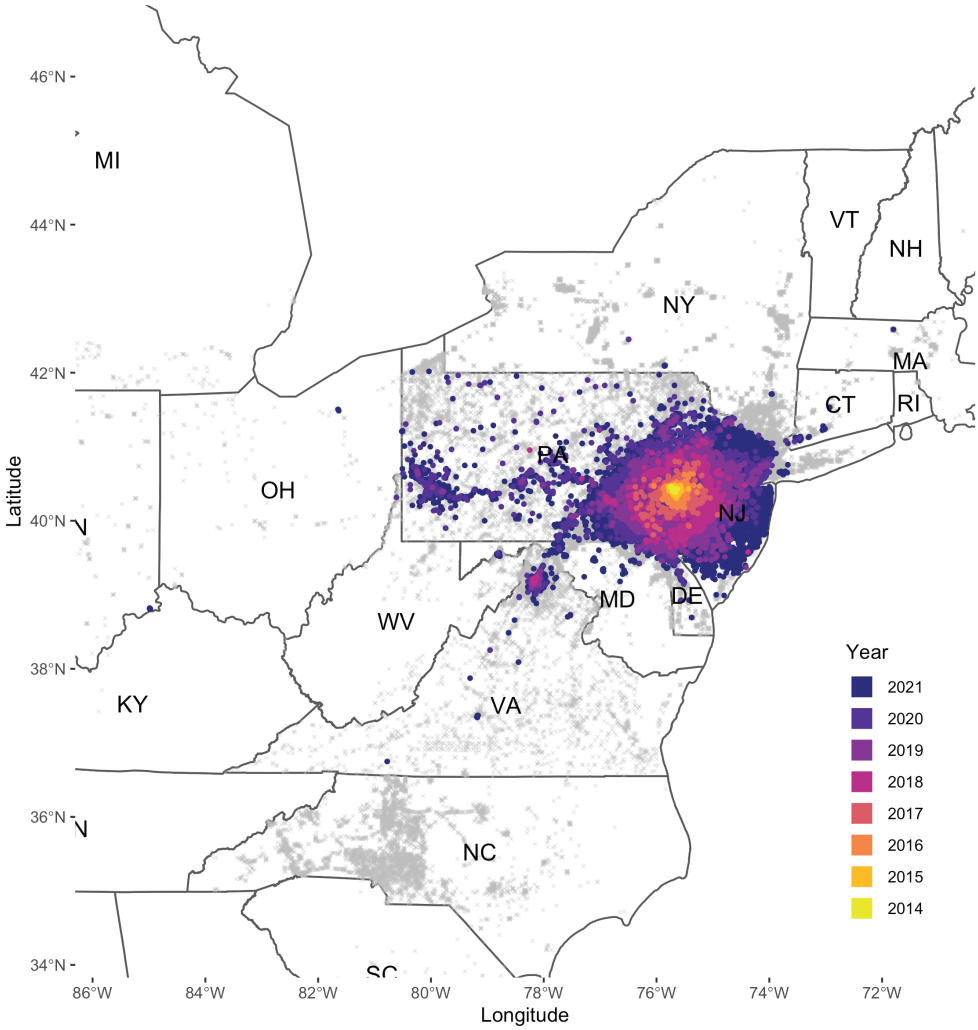

Figure 2. Map produced through the package function **map_spread**(). The map shows the year of first discovery of established populations of the spotted lanternfly (coloured points) in 1-km^2 grid cells across the eastern United States, as well as the location of negative survey records for the establishment of the species (grey crosses).

the data, which occurred, in particular, when combining different methods to score population density of spotted lanternfly, we contacted directly the maintainers of the individual datasets for insight. An additional challenge we faced was reaching a compromise between safeguarding the privacy of stakeholders while providing a high-resolution dataset to allow accurate forecasting and management planning. Protecting individual interests while allowing data to be shared openly is a topic of current relevance (Zipper et al. 2019). The resolution of 1 km^2 used in our dataset was reached after thorough discussions with the agencies involved, to ensure no breach of privacy

Figure 3. Map produced through the package function **map_yearly()**, showing the population density of spotted lanternfly assessed yearly in 10-km^2 grid cells across the eastern United States (red tiles).

occurred. Paramount to overcome both challenges was a tight collaboration with the agencies. We contacted data maintainers soon after a new agency was becoming involved in data collection, to start developing a relationship of trust and cooperation. This created an open line of communication with the agencies collecting the data from the field and curating the individual datasets and produced a feedback loop that we believe strengthens the quality and reliability of our dataset.

Author's contribution

SDB and MRH conceived the paper, gathered the data, produced the comprehensive dataset and wrote the code for the package. LB, PK, JL and GRP provided survey data and helped harmonise it across sources. All authors contributed with the writing of the manuscript.

Data availability

The package, containing the open access data, is stored as a public repository at https://github.com/ieco-lab/lydemapr. Additionally, versions of the 1 km^2 and 10 km^2 datasets are stored on Zenodo DOI: 10.5281/zenodo.7976229.

Acknowledgements

We would like to thank Eric Day for providing data on a citizen-science project run by the Virginia Polytechnic Institute and State University and the Virginia Cooperative Extension. We thank Jocelyn Behm, Stefani Cannon, Anna Carlson, Jason Gleditsch, Stephanie Lewkiewicz, Sam Owens, Payton Phillips and Timothy Swartz for their insightful comments on early drafts. This work was funded by the United States Department of Agriculture Animal and Plant Health Inspection Service Plant Protection and Quarantine under agreements AP19PPQS&T00C251, AP20PPQS&T00C136, AP20PPQS&T00C118, AP22PPQS&T00C146 and AP22PPQS&T00C097; the United States Department of Agriculture National Institute of Food and Agriculture Specialty Crop Research Initiative Coordinated Agricultural Project Award 2019-51181-30014; the Pennsylvania Department of Agriculture under agreements 44176768, 44187342, C9400000036, C94000833 and C940000835; and the California Department of Food and Agriculture under agreement A20-0850-000-SA.

References

Araújo MB, New M (2007) Ensemble forecasting of species distributions. Trends in Ecology & Evolution 22(1): 42–47. https://doi.org/10.1016/j.tree.2006.09.010

Barringer LE, Ciafré CM (2020) Worldwide feeding host plants of spotted lanternfly, with significant additions from North America. Environmental Entomology 49: 999–1011. https://doi.org/10.1093/ee/nvaa093

Barringer LE, Donovall LR, Spichiger S-E, Lynch D, Henry D (2015) The first New World record of *Lycorma delicatula* (Insecta: Hemiptera: Fulgoridae). Entomological News 125(1): 20–23. https://doi.org/10.3157/021.125.0105

Catlin-Groves CL (2012) The citizen science landscape: from volunteers to citizen sensors and beyond. International Journal of Zoology 2012: e349630. https://doi.org/10.1155/2012/349630

Clark JS, Lewis M, McLachlan JS, HilleRisLambers J (2003) Estimating Population Spread: What Can We Forecast and How Well? Ecology 84(8): 1979–1988. https://doi.org/10.1890/01-0618

Cook RT, Ward SF, Liebhold AM, Fei S (2021) Spatial dynamics of spotted lanternfly, *Lycorma delicatula*, invasion of the Northeastern United States. NeoBiota 70: 23–42. https://doi.org/10.3897/neobiota.70.67950

Dara SK, Barringer L, Arthurs SP (2015) *Lycorma delicatula* (Hemiptera: Fulgoridae): A new invasive pest in the United States. Journal of Integrated Pest Management 6(1): 20. https://doi.org/10.1093/jipm/pmv021

Diagne C, Catford JA, Essl F, Nuñez MA, Courchamp F (2020) What are the economic costs of biological invasions? A complex topic requiring international and interdisciplinary expertise. NeoBiota 63: 25–37. https://doi.org/10.3897/neobiota.63.55260

Dickinson JL, Zuckerberg B, Bonter DN (2010) Citizen science as an ecological research tool: challenges and benefits. Annual Review of Ecology, Evolution, and Systematics 41(1): 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Fantle-Lepczyk JE, Haubrock PJ, Kramer AM, Cuthbert RN, Turbelin AJ, Crystal-Ornelas R, Diagne C, Courchamp F (2022) Economic costs of biological invasions in the United States. The Science of the Total Environment 806: 151318. https://doi.org/10.1016/j.scitotenv.2021.151318

Fisher RA (1937) The wave of advance of advantageous genes. Annals of Human Genetics 7: 355–369. https://doi.org/10.1111/j.1469-1809.1937.tb02153.x

Hampton SE, Strasser CA, Tewksbury JJ, Gram WK, Budden AE, Batcheller AL, Duke CS, Porter JH (2013) Big data and the future of ecology. Frontiers in Ecology and the Environment 11(3): 156–162. https://doi.org/10.1890/120103

Higgins SI, Richardson DM (1996) A review of models of alien plant spread. Ecological Modelling 87(1–3): 249–265. https://doi.org/10.1016/0304-3800(95)00022-4

Hudgins EJ, Liebhold AM, Leung B (2017) Predicting the spread of all invasive forest pests in the United States. Ecology Letters 20: 426–435. https://doi.org/10.1111/ele.12741

Hulme PE (2009) Trade, transport and trouble: Managing invasive species pathways in an era of globalization. Journal of Applied Ecology 46(1): 10–18. https://doi.org/10.1111/j.1365-2664.2008.01600.x

Huron NA, Helmus MR (2022) Predicting host associations of the invasive spotted lanternfly on trees across the USA. bioRxiv: 2022.09.12.507604. https://doi.org/10.1101/2022.09.12.507604

Huron NA, Behm JE, Helmus MR (2022) Paninvasion severity assessment of a U.S. grape pest to disrupt the global wine market. Communications Biology 5(1): 655. https://doi.org/10.1038/s42003-022-03580-w

Johnson BA, Mader AD, Dasgupta R, Kumar P (2020) Citizen science and invasive alien species: An analysis of citizen science initiatives using information and communications technology (ICT) to collect invasive alien species observations. Global Ecology and Conservation 21: e00812. https://doi.org/10.1016/j.gecco.2019.e00812

Jones M, O'Brien M, Mecum B, Boettiger C, Schildhauer M, Maier M, Whiteaker T, Earl S, Chong S (2019) Ecological Metadata Language version 2.2.0. https://doi.org/10.5063/f11834t2

Jones C, Skrip MM, Seliger BJ, Jones S, Wakie T, Takeuchi Y, Petras V, Petrasova A, Meentemeyer RK (2022) Spotted lanternfly predicted to establish in California by 2033 without preventative management. Communications Biology 5(1): 558. https://doi.org/10.1038/s42003-022-03447-0

Jongejans E, Shea K, Skarpaas O, Kelly D, Ellner SP (2011) Importance of individual and environmental variation for invasive species spread: A spatial integral projection model. Ecology 92(1): 86–97. https://doi.org/10.1890/09-2226.1

Jung J-M, Jung S, Byeon D, Lee W-H (2017) Model-based prediction of potential distribution of the invasive insect pest, spotted lanternfly *Lycorma delicatula* (Hemiptera: Fulgoridae), by using CLIMEX. Journal of Asia-Pacific Biodiversity 10(4): 532–538. https://doi.org/10.1016/j.japb.2017.07.001

Kelling S, Hochachka WM, Fink D, Riedewald M, Caruana R, Ballard G, Hooker G (2009) Data-intensive science: A new paradigm for biodiversity studies. Bioscience 59(7): 613–620. https://doi.org/10.1525/bio.2009.59.7.12

Kobori H, Dickinson JL, Washitani I, Sakurai R, Amano T, Komatsu N, Kitamura W, Takagawa S, Koyama K, Ogawara T, Miller-Rushing AJ (2016) Citizen science: A new approach to advance ecology, education, and conservation. Ecological Research 31(1): 1–19. https://doi.org/10.1007/s11284-015-1314-y

Kot M, Lewis MA, van den Driessche P (1996) Dispersal data and the spread of invading organisms. Ecology 77(7): 2027–2042. https://doi.org/10.2307/2265698

Kovacs K, Václavík T, Haight RG, Pang A, Cunniffe NJ, Gilligan CA, Meentemeyer RK (2011) Predicting the economic costs and property value losses attributed to sudden oak death damage in California (2010–2020). Journal of Environmental Management 92(4): 1292–1302. https://doi.org/10.1016/j.jenvman.2010.12.018

Leroy B, Kramer AM, Vaissière A-C, Kourantidou M, Courchamp F, Diagne C (2022) Analysing economic costs of invasive alien species with the invacost r package. Methods in Ecology and Evolution 13(9): 1930–1937. https://doi.org/10.1111/2041-210X.13929

Lewkiewicz SM, De Bona S, Helmus MR, Seibold B (2022) Temperature sensitivity of pest reproductive numbers in age-structured PDE models, with a focus on the invasive spotted lanternfly. Journal of Mathematical Biology 85(3): 29. https://doi.org/10.1007/s00285-022-01800-9

Lindenmayer D, Scheele B (2017) Do not publish. Science 356(6340): 800–801. https://doi.org/10.1126/science.aan1362

Lunghi E, Corti C, Manenti R, Ficetola GF (2019) Consider species specialism when publishing datasets. Nature Ecology & Evolution 3(3): 319–319. https://doi.org/10.1038/s41559-019-0803-8

Maino JL, Schouten R, Lye JC, Umina PA, Reynolds OL (2022) Mapping the life-history, development, and survival of spotted lantern fly in occupied and uninvaded ranges. Biological Invasions 24(7): 2155–2167. https://doi.org/10.1007/s10530-022-02764-z

Michener WK, Jones MB (2012) Ecoinformatics: Supporting ecology as a data-intensive science. Trends in Ecology & Evolution 27(2): 85–93. https://doi.org/10.1016/j.tree.2011.11.016

Murman K, Setliff GP, Pugh CV, Toolan MJ, Canlas I, Cannon S, Abreu L, Fetchen M, Zhang L, Warden ML, Wallace M, Wickham J, Spichiger S-E, Swackhamer E, Carrillo D, Cornell A, Derstine NT, Barringer L, Cooperband MF (2020) Distribution, survival, and development of spotted lanternfly on host plants found in North America. Environmental Entomology 49(6): 1270–1281. https://doi.org/10.1093/ee/nvaa126

Neubert MG, Kot M, Lewis MA (2000) Invasion speeds in fluctuating environments. Proceedings. Biological Sciences 267(1453): 1603–1610. https://doi.org/10.1098/rspb.2000.1185

Norberg A, Abrego N, Blanchet FG, Adler FR, Anderson BJ, Anttila J, Araújo MB, Dallas T, Dunson D, Elith J, Foster SD, Fox R, Franklin J, Godsoe W, Guisan A, O'Hara B, Hill NA, Holt RD, Hui FKC, Husby M, Kålås JA, Lehikoinen A, Luoto M, Mod HK, Newell G, Renner I, Roslin T, Soininen J, Thuiller W, Vanhatalo J, Warton D, White M, Zimmermann NE, Gravel D, Ovaskainen O (2019) A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecological Monographs 89(3): e01370. https://doi.org/10.1002/ecm.1370

Norman-Burgdolf H, Rieske LK (2021) Healthy trees – healthy people: A model for engaging citizen scientists in exotic pest detection in urban parks. Urban Forestry & Urban Greening 60: 127067. https://doi.org/10.1016/j.ufug.2021.127067

NYIPM (2023) Spotted Lanternfly | New York State Integrated Pest Management. New York State Integrated Pest Management: Spotted Lanternfly. https://nysipm.cornell.edu/environment/invasive-species-exotic-pests/spotted-lanternfly/ [May 25, 2022]

R Core Team (2021) R: A language and environment for statistical computing. https://www.R-project.org/

Ramirez VA, Bona SD, Helmus MR, Behm JE (2023) Multiscale assessment of oviposition habitat associations and implications for management in the spotted lanternfly (*Lycorma delicatula*), an emerging invasive pest. Journal of Applied Ecology 60(3): 411–420. https://doi.org/10.1111/1365-2664.14365

Reichman OJ, Jones MB, Schildhauer MP (2011) Challenges and opportunities of open data in ecology. Science 331(6018): 703–705. https://doi.org/10.1126/science.1197962

Robertson PA, Mill A, Novoa A, Jeschke JM, Essl F, Gallardo B, Geist J, Jarić I, Lambin X, Musseau C, Pergl J, Pyšek P, Rabitsch W, von Schmalensee M, Shirley M, Strayer DL, Stefansson RA, Smith K, Booy O (2020) A proposed unified framework to describe the management of biological invasions. Biological Invasions 22(9): 2633–2645. https://doi.org/10.1007/s10530-020-02298-2

Rodrigues AMM, Johnstone RA (2014) Evolution of positive and negative density-dependent dispersal. Proceedings of the Royal Society B: Biological Sciences 281: 20141226–20141226. https://doi.org/10.1098/rspb.2014.1226

Sakai AK, Allendorf FW, Holt JS, Lodge DM, Molofsky J, With KA, Baughman S, Cabin RJ, Cohen JE, Ellstrand NC, McCauley DE, O'Neil P, Parker IM, Thompson JN, Weller SG (2001) The population biology of invasive species. Annual Review of Ecology and Systematics 32(1): 305–332. https://doi.org/10.1146/annurev.ecolsys.32.081501.114037

Santaoja M (2022) Insect affects: A study on the motivations of amateur entomologists and implications for citizen science. Science & Technology Studies 35: 58–79. https://doi.org/10.23987/sts.107703

Skellam JG (1951) Random dispersal in theoretical populations. Biometrika 38(1–2): 196–218. https://doi.org/10.1093/biomet/38.1-2.196

Sullivan BL, Aycrigg JL, Barry JH, Bonney RE, Bruns N, Cooper CB, Damoulas T, Dhondt AA, Dietterich T, Farnsworth A, Fink D, Fitzpatrick JW, Fredericks T, Gerbracht J, Gomes C, Hochachka WM, Iliff MJ, Lagoze C, La Sorte FA, Merrifield M, Morris W, Phillips TB, Reynolds M, Rodewald AD, Rosenberg KV, Trautmann NM, Wiggins A, Winkler DW, Wong W-K, Wood CL, Yu J, Kelling S (2014) The eBird enterprise: An integrated approach to development and application of citizen science. Biological Conservation 169: 31–40. https://doi.org/10.1016/j.biocon.2013.11.003

Travis JMJ, Dytham C (2002) Dispersal evolution during invasions. Evolutionary Ecology Research 4: 1119–1129.

Urban JM (2020) Perspective: Shedding light on spotted lanternfly impacts in the USA. Pest Management Science 76(1): 10–17. https://doi.org/10.1002/ps.5619

Urban JM, Calvin D, Hills-Stevenson J (2021) Early response (2018–2020) to the threat of spotted lanternfly, *Lycorma delicatula* (Hemiptera: Fulgoridae) in Pennsylvania. Annals of the Entomological Society of America 114: 709–718. https://doi.org/10.1093/aesa/saab030

Wakie TT, Neven LG, Yee WL, Lu Z (2020) The establishment risk of *Lycorma delicatula* (Hemiptera: Fulgoridae) in the United States and Globally. Journal of Economic Entomology 113: 306–314. https://doi.org/10.1093/jee/toz259

Walker K, Rudis B (2023) tigris: Load Census TIGER/Line Shapefiles. https://cran.r-project.org/web/packages/tigris/index.html [May 26, 2023]

Wickham H, Hester J, Chang W, Bryan J (2022) devtools: Tools to Make Developing R Packages Easier. https://devtools.r-lib.org/ https://github.com/r-lib/devtools [December 15, 2022]

Zhang C, Boyle KJ (2010) The effect of an aquatic invasive species (Eurasian watermilfoil) on lakefront property values. Ecological Economics 70(2): 394–404. https://doi.org/10.1016/j.ecolecon.2010.09.011

Zipper SC, Stack Whitney K, Deines JM, Befus KM, Bhatia U, Albers SJ, Beecher J, Brelsford C, Garcia M, Gleeson T, O'Donnell F, Resnik D, Schlager E (2019) Balancing open science and data privacy in the water sciences. Water Resources Research 55(7): 5202–5211. https://doi.org/10.1029/2019WR025080