Description

The Spotted Lanternfly (*Lycorma delicatula*) is an invasive pest species in the North Eastern USA, first discovered in Berks County, PA, in 2014. Since its first discovery, several agencies (with the Pennsylvania Dept. of Agriculture, and the US Dept. of Agriculture in a leading role) have taken up the task to monitor and control SLF populations.

Format

A dataframe with 658,390 observations and 14 variables

source

Character variable defining the source of the data.

year

Integer value defining the calendar year when the information was collected.

bio_year

Integer value definig the biological year when the information was collected based on SLF life cycle. Biological year starts on May 1st and ends on April 30th.

latitude

Expressed in decimal degrees in WSG84 coordinate system

longitude

Expressed in decimal degrees in WSG84 coordinate system

state

Character defining the state where the data was collected, abbreviated with Census-official 2-letter code

lyde_present

Logical value (`TRUE`/`FALSE`) 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 (`TRUE`/`FALSE`) defining whether signs of an established population are present at the site at the time of survey. These include a minimum of 2 alive individuals or the presence of an egg mass.

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 hundreds 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.

source_agency

Agency/organization/platform responsible for data collection. `DDA`: Delaware Dept of Agriculture; `iNaturalist`: inaturalist.org; `ISDA`: Indiana State Dept of Agriculture; `MDA`: Maryland Dept of Agriculture; `NJDA_Public_reporting`: New Jersey Dept of Agriculture public reporting tool platform; `NYSDAM`: New York State Dept of Agriculture and Markets; `PDA_Public_reporting`: Pennsylvania Dept of Agriculture public reporting tool platform; `USDA`: United States Dept of Agriculture; `VA_Tech_Coop_Ext`: Virginia Polytechnic Institute and State University, and Virginia Cooperative Extension; `VDA`: Virginia Dept of Agricuture.

collection_method

Character string defining the method used for data collection. `field_survey/management` for data points collected by professionals during field operations; `individual_reporting` for data points collected by individuals through public reporting tools, inatualist observations, or citizen science projects.

pointID

Character string uniquely identifying each data point.

rounded_longitude_10k

longitude of the centroid of the closest 10 km2 grid cell, expressed in decimal degrees (WSG84 coordinate system), used to rarefy the dataset at a coarser resolution.

rounded_latitude_10k

latitude of the centroid of the closest 10 km2 grid cell, expressed in decimal degrees (WSG84 coordinate system), used to rarefy the dataset at a coarser resolution.

Check these articles and reports for reference:

https://www.biorxiv.org/content/10.1101/2023.01.27.525992v1.full.pdf //this is the report we are following

https://www.dec.ny.gov/animals/113303.html

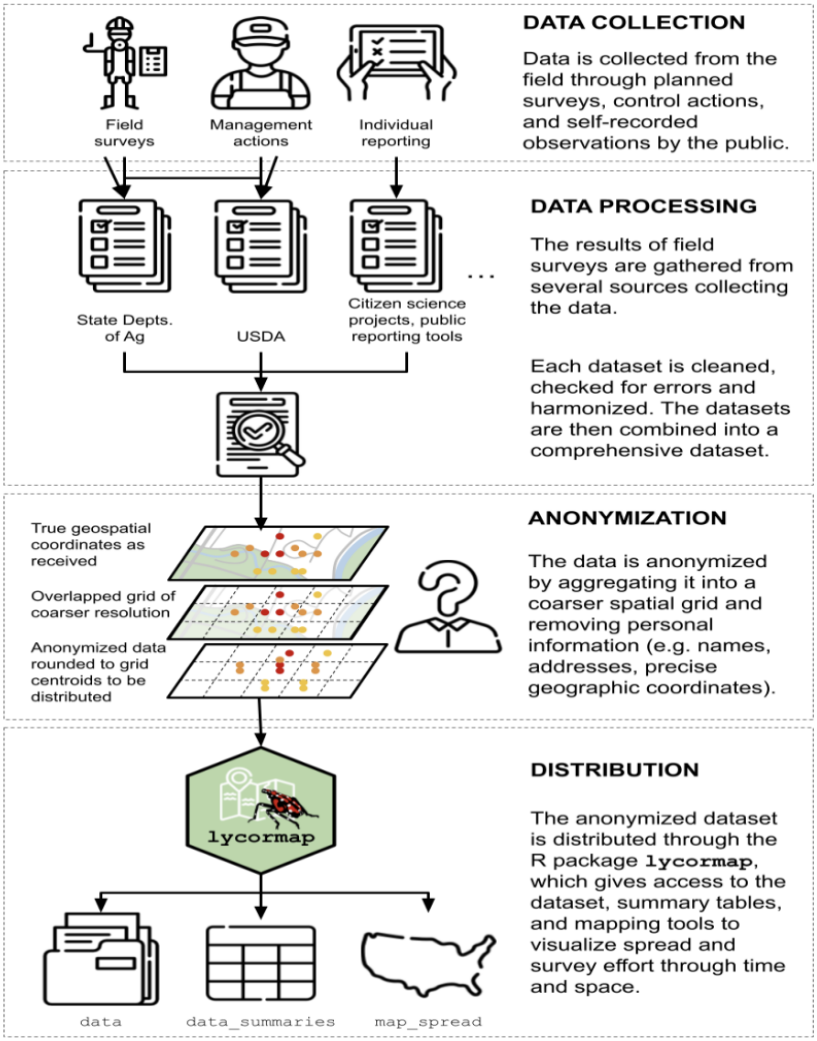
https://www.aphis.usda.gov/aphis/resources/pests-diseases/hungry-pests/the-threat/spotted-lanternfly/spotted-lanternfly

Much detailed description of variable:

The spotted lanternfly (Lycorma delicatula, White 1845; often referred to as SLF in the literature) was first discovered in the United States in Berks County, Pennsylvania, in 2014 (Barringer et al., 2015; Dara et al., 2015), and has since spread to 12 states across the Mid Atlantic and Midwestern United States (NYIPM, 2022; Urban et al., 2021). This phloem-feeding planthopper is native to China and was likely introduced accidentally via a shipment of landscaping materials (Urban, 2020). The spotted lanternfly is known to feed on over 100 species of plants (Barringer & Ciafré, 2020; Huron & Helmus, 2022; Murman et al., 2020) and poses a major economic burden on viticulture as it feeds on grapevine reducing total yield and plant vigor (Urban, 2020). State agencies and the United States Department of Agriculture (USDA) have collected large amounts of data on the spread of this pest through field surveys. In addition, given the species is easily recognized and hard to misidentify, a broad public campaign to educate the public has promoted the collection of citizen science data. Data is 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 is heavily heterogeneous. Currently, any research team analyzing the spread of the pest has to invest a significant amount of time processing the data before using it in model construction and validation.

Data and metadata

The dataset contained in the package represents an anonymized 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 spotted lanternfly in the United States



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 was 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, and citizen science data 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 publication, the aggregated and anonymized dataset contains 658,392 individual observations pertaining to 61,715 point-locations throughout the United States collected between 2014 and 2022. These 61,715 point-locations represent centroids of a 1 km2 146 grid at which the geospatial data was aggregated for anonymization. The exact latitude and longitude of each survey contained in the geospatial data collected by the sources are rounded to the coordinates of the centroids. This approach, while removing the ability to derive property-level information from the dataset, allows us to distribute survey-level information the data user can summarize as it best fits their needs. All variables containing traceable information regarding pesonal names, business names, contact information, and comments, were also removed from the dataset.

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 is kept separate from data collected by other states, as Pennsylvania was the state were the first introduction was detected. Because 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 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 (May 1st–April 30th).

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 d**efining whether signs of an established population are present at the site at the time of survey.** These include a minimum of **2 alive individuals or the presence of an egg mass a**s 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 km2 grid cell, expressed in decimal degrees (WSG84 coordinate system), used to rarefy the dataset at a coarser resolution.
rounded_latitude_10k
latitude of the centroid of the closest 10 km2 grid cell, expressed in decimal degrees (WSG84 coordinate system), used to rarefy the dataset at a coarser resolution.

The dataset you described contains information about spotted lanternfly (SLF) observations, including the source, year, location, and various attributes related to the presence and density of SLF. Depending on your goals and interests, you can perform various data analysis and visualization tasks with this dataset. Here are some potential analyses and tasks you can do:

1. **Descriptive Statistics**: Calculate summary statistics to get an overview of the dataset. For example, you can find the mean, median, standard deviation, and other statistics for the "lyde_density" variable to understand the distribution of SLF population density.
2. **Time Series Analysis**: Analyze how SLF observations vary over the years. You can create time series plots to visualize trends and patterns in SLF presence and density.
3. **Geospatial Analysis**: Use the latitude and longitude coordinates to create maps or perform geospatial analysis. Plot SLF observations on a map to see the geographical distribution. You can also cluster the data to identify hotspots of SLF activity.
4. **Categorical Analysis**: Explore the "lyde_density" variable by creating bar plots or pie charts to understand the distribution of SLF population density categories.

5. **Correlation Analysis**: Investigate relationships between different variables. For example, you can check if there is a correlation between the presence of SLF and the "year."
6. **Grouped Analysis**: Group the data by different attributes like "state," "source," or "source_agency" to compare SLF observations across different categories.
7. **Time Series Forecasting**: If you have enough data, you can build time series forecasting models to predict future SLF observations based on historical data.
8. **Anomaly Detection**: Identify anomalies or unusual patterns in the data, which might be of interest for further investigation.
9. **Data Cleaning and Preprocessing**: Check for missing data, outliers, and inconsistencies in the dataset. Clean and preprocess the data as needed for analysis.
10. **Data Visualization**: Create various plots and visualizations to communicate your findings effectively, such as line graphs, bar charts, heatmaps, and scatter plots.
11. **Machine Learning**: If you have specific predictive tasks in mind, you can train machine learning models to make predictions based on the dataset, such as predicting SLF presence or density.
12. **Statistical Testing**: Perform statistical tests to determine if there are significant differences between different groups in the dataset.

The specific analysis you choose to perform will depend on your research questions, objectives, and the insights you want to gain from the dataset. You can use Python libraries like pandas, matplotlib, seaborn, scikit-learn, and geospatial libraries (e.g., geopandas) to perform these tasks and create visualizations.

https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/statewide/time-series/36/tavg/ytd/9/2014-2023?base_prd=true&begbaseyear=1901&endbaseyear=2000