Spatiotemporal Prediction of Spotted Lanternfly Spread

Jackson Bauer

Abstract

The Spotted Lanternfly (*Lycorma delicatula*), an invasive insect first detected in Pennsylvania in 2014, has quickly spread across the eastern United States, causing immense agricultural and economic damage. Accurate prediction of its spread is critical for proactive defense and mitigation strategies. We present a machine learning approach for predicting county level SLF infestation using XGBoost. The key design relies on the iterative calculation of spatial lag, the proportion of infested neighboring counties, each year. Capturing the dispersal of invasive species. We compare three models: environmental features only (AUC=0.8670), environmental plus geographic location (AUC=0.9676), and environmental plus spatial lag (AUC=9452). The spatial lag model achieves the most realistic performance and, when combined with infestation constraints (prediction threshold=0.7, neighbor threshold=0.4), produces realistic yearly forecasts. Our approach shows that integrating spatial design through continuously changing neighbor relationships improves prediction accuracy for invasive species dispersal.

1 Introduction

1.1 The Spotted Lanternfly Crisis

The Spotted Lanternfly (SLF, *Lycorma delicatula*) is an invasive planthopper native to China and south east Asia. It was first reported in Berks County, Pennsylvania in 2014 [1]. Since then, it has spread across the mid and northeastern United States, with infestations in over 200 counties as of 2025. SLF nymphs feed on a wide variety of common plants and trees, while SLF adults have a preference for Tree of Heaven (*Ailanthus altissima*). The economic impact is vast. Estimated annual damages are hundreds of millions in Pennsylvania alone.

The rapid spread of the SLF infestation is a significant challenge for agricultural security. Current mitigation strategies are reactive. Simply responding to new reports rather than anticipating the spread. This reactive approach is inefficient and ineffective, because an infestation is always well developed and stable before being detected. There is a need for predictive tools that can forecast SLF dispersal to allow for proactive mitigation.

1.2 Research Gap and Contributions

Several studies have documented SLF biology and dispersal behaviors. SLF's show an impressive host adaptability, with nymphs being able to survive on over 65 known host species [3]. Adults have a smaller range of host species, with strong preference for Tree of Heaven [4]. The insects adaptability and wide host range make tracking and predicting spread extremely challenging [1].

Recent research has uncovered an important pattern in SLF behavior. Owen et al. [6] showed that SLF's group in urban areas in colder climates. This is likely due to the urban heat island phenomenon which causes a warmer microclimate around cities and other dense, urban areas that ease survival and increase activity in cold, northern regions. This finding has a critical impact on predictive modeling, causing urban characteristics to be weighted differently across the country.

Our work addresses these research gaps with the following contributions:

1. **Spatial lag feature**: We introduce a spatial lag variable, the proportion of infested neighboring counties, which is calculated iteratively throughout each step in our forecasting. This captures the spreading front of an invasion.
2. **Ecological constraints**: We implement spread thresholds (prediction probability and neighbor infestation) that slow down the spread, prevent unrealistic over prediction, and produce forecasts more consistent with previously observed spread rates.
3. **10-year forecasts**: We show county level infestation predictions for the next 10 years, giving counties warning well in advance.

2 Methodology

2.1 Data Sources and Preprocessing

2.1.1 SLF Observation Data

We obtained 26,132 SLF presence reports from the Early Detection & Distribution Mapping System (EDDMapS) from 2014-2025. Reports were organized into yearly county level presence/absence, creating a binary classification for every year, county combination. With the constraint that once a county becomes infested, it remains infested for all subsequent years. This means that the models do not account for any type of mitigation or eradication efforts.

2.1.2 Environmental Features

Tree of Heaven Distribution: We obtained county level presence/absence report data for Tree of Heaven

from the EDDMapS. This creates a binary indicator for the presence of SLF's preferred host plant. Tree of Heaven is especially vital because it is the primary diet for adult SLF and plays a critical role in reproduction [4].

Urban-Rural Classification: We used the National Center for Health Statistics urban-rural classification dataset, which classifies counties on a 1-6 scale from most urban (1, Large Central Metro) to most rural (6, Noncore). Research by Owen et al. [6] has shown that SLF prefer urban areas when in colder climates. This is due to the urban heat island phenomenon which causes a warmer microclimate that aid in SLF survival in northern regions.

We created an "urban_effective" binary feature equal to 1 for counties north of 38°N latitude, where the urban heat islands provide survival advantage, and 0 for counties south of this line. This threshold approximates the separation line between regions where the urban level influences SLF survival from southern areas where urban heat islands provide no additional benefit.

2.1.3 Geographic Features

County centroids by latitude and longitude were calculated from US Census Bureau TIGER/Line shapefiles. County neighbors were determined using the shapefile topology, finding all counties whose boundaries physically touch each other. From this we created a network of 3,109 counties in the continental United States.

2.1.4 Feature Engineering

For each year, county combination, we calculated:

- **Year**: Numeric year (2014-2023 for training, 2024-20205 for testing, 2026-2035 for forecasting)
- **Tree_of_heaven**: Binary indicator which was constant across all years
- **Urban_code**: NCHS value 1-6
- **Urban_effective**: Binary indicator based on latitude threshold
- **Centroid_lat, Centroid_lon**: Geographic coordinates
- **Spatial_lag**: Proportion of infested neighboring counties in a year which was calculated yearly

The spatial lag is defined as:

$$\text{spatial\_lag}_{i,t} = \frac{\sum_{j \in N(i)} \text{infested}_{j,t}}{|N(i)|} \tag{1}$$

where $|N(i)|$ is the number of neighbors.

2.2 Model Structure

We employed XGBoost, a gradient boosting library, because of its performance on non linear relationships, and integrated assistance with class imbalance. We trained three models with progressively increasing feature sets:

Model 1 (Environmental): year, tree_of_heaven, urban_code, urban_effective

Model 2 (Environmental + Centroid): Model 1 features + centroid_lat, centroid_lon

Model 3 (Environmental + Spatial Lag): Model 1 features + spatial_lag

2.3 Hyperparameter Optimization

We performed hyperparameter optimization using random search. 50 hyperparameter iterations were sampled. Each configuration was evaluated using five fold cross validation to retain class proportions across folds.

Model performance was tested using the Receiver Operating Characteristic Area Under Curve (ROC AUC). The mean ROC-AUC from the five folds was used as the optimization metric for choosing the best hyperparameters.

Table 1: Hyperparameter search space for the XGBoost classifier

Hyperparameter	Range	Distribution
n_estimators	100 – 500	Discrete uniform
max_depth	3 – 10	Discrete uniform
learning_rate	0.01 – 0.30	Continuous uniform
subsample	0.60 – 1.00	Continuous uniform
colsample_bytree	0.60 – 1.00	Continuous uniform
gamma	0 – 5	Continuous uniform
min_child_weight	1 – 10	Discrete uniform
reg_alpha	0 – 1	Continuous uniform
reg_lambda	0 – 2	Continuous uniform

2.4 Training and Evaluation

2.4.1 Train/Test

We used a yearly train test split:

- **Training**: 2014-2023 (10 years, 31,090 year, county observations)
- **Testing**: 2024-2025 (2 years, 6,218 year, county observations)

This yearly split avoids data leakage that would happen if we used typical random cross-validation.

2.4.2 Prediction Thresholds

Standard XGBoost classification defines a 0.5 probability threshold. Initial experiments showed this threshold produced extremely aggressive and unrealistic forecasts. We implemented two constraints to slow the spread to a rate more realistic to observed spread rates:

1. **Probability threshold**: 0.7 (higher confidence required to classify a county as infested)
2. **Spatial lag threshold** (Model 3 only): spatial_lag ≥ 0.4 (at least 40% of neighboring counties must be infested for the target county to be infested)

Table 2: Model performance comparison on 2024-2025 test set. All metrics computed with prediction threshold=0.7 and spatial lag threshold=0.4 for Model 3 only.

Metric	(Env)	(Env+Loc)	(Env+Spatial)
ROC-AUC	0.8670	0.9676	0.9452
Accuracy	0.6665	0.9788	0.9598
Precision	0.1963	0.9905	0.7782
Recall	0.9048	0.7656	0.7582
F1-Score	0.3227	0.8636	0.7681
MCC	0.3161	0.8607	0.7462

Both conditions must be satisfied for a prediction to be made. These thresholds significantly reduce false positives and give the model more realism.

2.4.3 Evaluation Metrics

Each model was evaluated using the following classification metrics:

- **ROC-AUC**: Area under the receiver operating characteristic curve
- **Accuracy, Precision, Recall**
- **F1-Score**: mean of precision and recall
- **Matthews Correlation Coefficient (MCC)**: Balanced measure for imbalanced data
- **Confusion Matrix**: True/false positives and negatives

2.5 Iterative Forecasting

We implemented an iterative prediction loop for forecasting future dispersal:

1. Initialize: Use 2025 observed infestation as baseline
2. For each forecast year t (2026-2035):
   (a) Calculate spatial_lag for all counties at year $t-1$
   (b) Predict infestation for year t using Model 3
   (c) Apply thresholds (probability ≥ 0.7 and spatial_lag ≥ 0.4)
   (d) Update county infestation status: $\text{infested}_t = \max(\text{infested}_{t-1}, \text{predicted}_t)$
   (e) Store predictions and iterate to next year

This method allows the spatial lag feature to change in the forecast as each year progresses, capturing the spreading infestation. The max operation in step 2d implements our constraint that counties remain infested for all consecutive years following their initial infestation.

3 Results

3.1 Model Performance

Table 2 shows evaluation metrics for all three models on the test set (2024-2025).

The massive improvement from Model 1 to Model 2 (ΔAUC = 0.1006) shows why geographic location is so important in modeling spread. The improvement from Model 1 to Model 3 (ΔAUC = 0.0782) shows that the spatial relationship between neighbors also carries predictive power, though less than raw geographic coordinates.

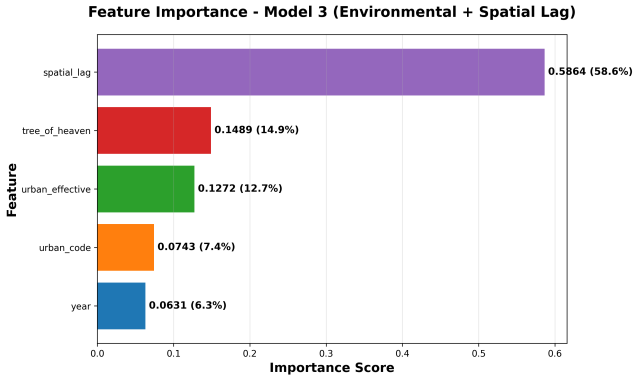


Figure 1: Feature importance for Model 3 (Environmental + Spatial Lag).

3.2 Feature Importance

Figure 1 shows the feature importance scores for Model 3. It makes sense that spatial lag dominates because SLF's can only spread to regions which are nearby. This confirms that the neighbor infestation status feature is the most informative predictor of future spread.

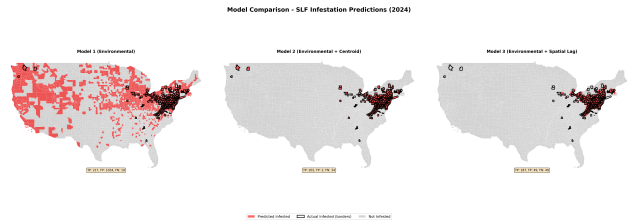
3.3 Spatial Predictions

Figure 2 displays predicted versus actual infestation patterns for 2024 and 2025 test years. The model successfully captures the geographic spread pattern, with predictions concentrated along the expanding front in Pennsylvania, New Jersey, Maryland, Virginia, and emerging hotspots in Ohio and Indiana. False positives (predicted but not actually infested) occur primarily in counties adjacent to known infestations, representing reasonable risk predictions. False negatives are rare, indicating high sensitivity to actual spread. Figure 2 displays predicted versus actual infestation patterns for 2024 and 2025 test years. The model successfully captures the geographic spread pattern, with predictions concentrated along the expanding front in Pennsylvania, New Jersey, Maryland, Virginia, and emerging hotspots in Ohio and Indiana. False positives (predicted but not actually infested) occur primarily in counties adjacent to known infestations, representing reasonable risk predictions. False negatives are rare, indicating high sensitivity to actual spread.

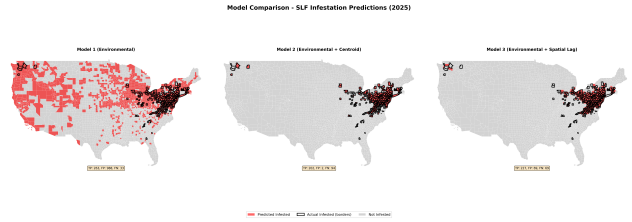
3.4 Ten Year Forecast

Figure 3 presents the 10-year iterative forecast (2026-2035) starting from the 2025 baseline. The forecast shows:

- Continued outward spread from the Mid-Atlantic center



(a) 2024 predictions



(b) 2025 predictions

Figure 2: Model predictions on test years. Red shading shows the predicted infestation; highlighted borders show actual infestations. Models 2 and 3 capture the spread pattern with high accuracy. Most errors occur in counties adjacent to known infested counties. Model 1 predicts a large amount of false positives, essentially just predicting infestations in counties with the Tree of Heaven.

Table 3: Sensitivity analysis of prediction thresholds on Model 3 test set.

Prob. Thresh.	Spatial Lag Thresh.	FP	FN	F1
0.5	0.0	559	62	0.6092
0.6	0.0	408	74	0.6620
0.7	0.0	300	81	0.7094
0.7	0.3	137	124	0.7638
0.7	0.4	118	132	0.7681
0.7	0.5	91	155	0.7607

- Gradual westward expansion
- Limited northward spread

By 2035, the model forecasts 573 infested counties, which is almost a 100% increase from 2025.

3.5 Ablation Study: Threshold Sensitivity

To evaluate the impact of our dual-threshold constraints, we tested Model 3 with different threshold combinations (Table 3).

Biological Containment Systems: For biological containment systems, false negatives are much more costly than false positives. A missed infestation allows the SLF's to have an unregulated and untracked spread, making prevention and eradication more difficult and expensive. A false positive causes an uninfested county to needlessly begin SLF mitigation and eradication, an acceptable trade off when compared to unchecked spread.

Challenges in iterative forecasting: For multiyear forecasts, completely unconstrained thresholds create a separate problem. The default threshold (0.5, 0.0) produces 559 false positives in the 2 year test set. In an iterative 10 year forecast, these false positives snow ball into more false positives, which is a huge problem. Without constraints, this snow ball produces wildly unrealistic forecasts where most of the United States becomes infested. Completely inconsistent with reality and observed spread rates.

Significance of probability threshold: Increasing the probability threshold from 0.5 to 0.7 reduces false positives from 559 to 300 while increasing false negatives from 62 to 81. This increase in false negatives is concerning when forecasting a few years, but is less of a problem over longer time periods where balance between FP and FN are more important.

Significance of spatial lag threshold: Adding the spatial lag constraint further reduces false positives but creates more false negatives. This tradeoff reflects competing objectives:

- **For 1-2 year forecasts**: Low thresholds (0.7, 0.0) with 81 FN is preferable, accepting high FP to minimize missed predictions.
- **For 10-year forecasts**: High thresholds (0.7, 0.4) with 132 FN are preferable to prevent snowballing from over predicting.

Selected thresholds (0.7, 0.4): We select these thresholds for our 10-year forecasts to keep a balance between false positives and false negatives.

Diminishing returns beyond (0.7, 0.4): Increasing the spatial lag threshold to 0.5 provides slight FP reduction (118 $\rightarrow$ 91) while significantly increasing FN (132 $\rightarrow$ 155). The extra 23 false negatives are not justified by the modest false positive reduction.

Recommendation for real life use: Organizations should select threshold levels based on the length of the desired forecast. Additionally the cost of false positives should be investigated to discover how expensive it is to mitigate and eradicate a non infested county.

4 Discussion

4.1 Findings and Interpretation

Our results establish two key findings for invasive species dispersal prediction:

1. Geographic information is highly predictive but may overfit. Model 2's impressive performance (AUC=0.968) with geographic coordinates demonstrates that proximity to previously observed infestations is the strongest predictor. However, this model may just capture location patterns that would not generalize to new regions. The extremely high precision (0.991) suggests the model is very conservative, but could be learning the

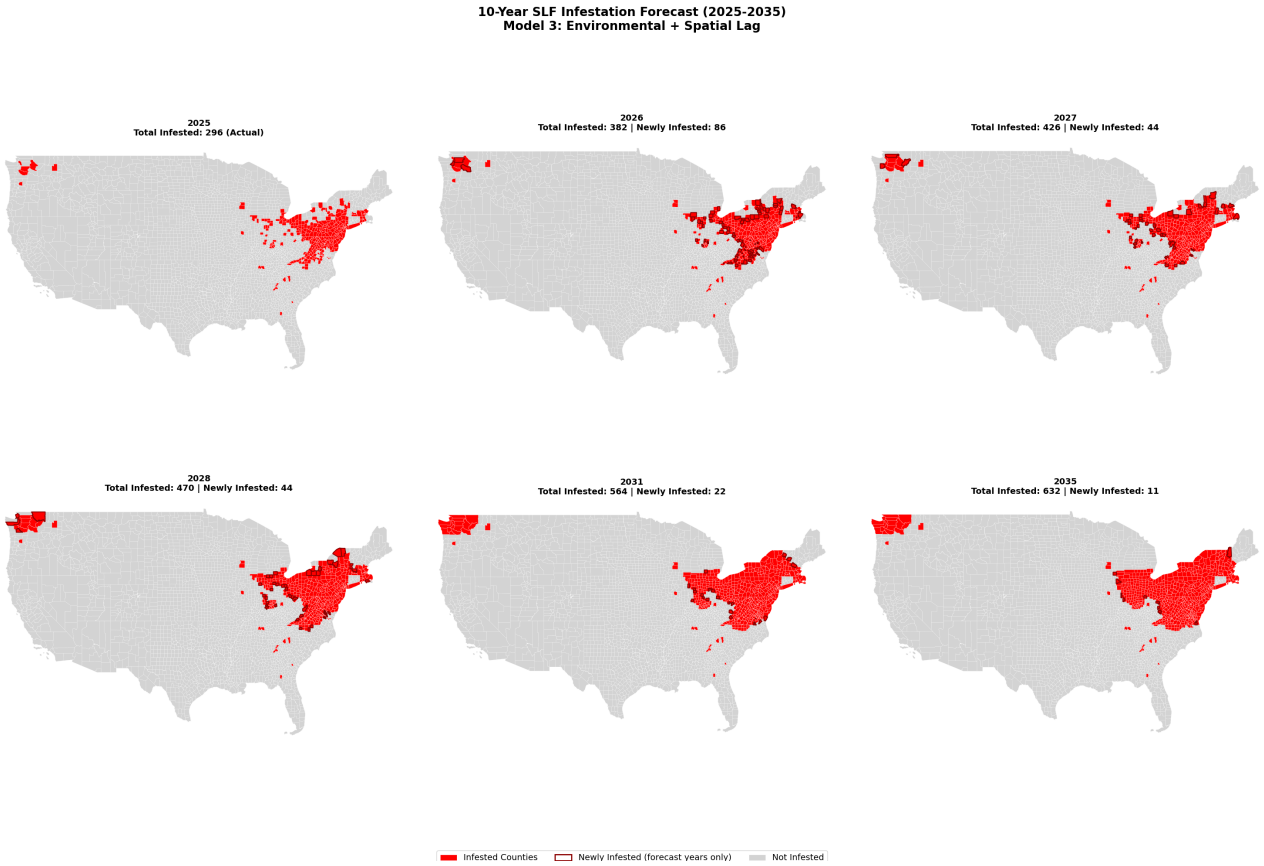


Figure 3: Ten-year forecast (2026-2035) with Year 0 (2025) baseline. Maps show iterative predictions with spatial lag recalculated annually using dual thresholds (probability ≥ 0.7, spatial lag ≥ 0.4). Red shading indicates cumulative infestations; dark red borders highlight newly infested counties each year.

geographic pattern of current infestations rather than spread behaviors.

2. Spatial lag captures spread behavior with better balance. Model 3's performance (AUC=0.945) by using the proportion of infested neighboring counties provides a better understanding of dispersal. The balance between precision (0.755) and recall (0.773) shows this model captures the biological components of dispersal more effectively than geographic coordinates.

The key tradeoff is between forecasting accuracy and behavorial understanding. Model 2's coordinate features achieve the best measured performance (AUC=0.968), but completely relies on current dispersal information. Model 3's spatial lag feature drops some accuracy (AUC=0.945), but gains an understanding of the core dispersal factors and may be better at generalizing to longer forecasting periods.

4.2 Limitations and Future Work

We acknowledge several limitations: **Data bias**: Our data is made up of EDDMapS user reports. Counties with more active users will report more SLF's making the county appear 'more' infested. Rural areas with no active users may appear to have no SLF's when they actually do. Counties may have an infestation remain unreported if users fail to spot or report. If county wide accurate surveys are taken with no reporting bias, their data should be used in future models.

Limited environmental features: SLF's are extremely adaptable to the United States because they can survive in a wide variety of climates and their preferred host trees are very common. Future models should implement more habitat and diet features as they are researched.

Static environmental features: We assume that the tree of heaven population and urbanization levels remain constant. Climate change may alter the distribution of the Tree of Heaven and the habitat of SLF's over time. Future work should implement environmental features which change over time.

No human aided dispersal: SLF spreads through natural dispersal and by hitching rides on human transports such as trucks, trains, and cargo. Our model captures general spread, but does not explicitly model human aided travel due to its complexity. In future work, human aided dispersal features should be investigated and implemented into the model.

No prevention or mitigation: Our predictions assume no active mitigation programs. In real life, local government, industry, and private land owners would constantly attempt to slow or prevent the SLF infestation. As infestation mitigation strategies are researched and tested, they should be implemented into the model to better reflect real life.

Single species: Our framework is generalizable, but we have only tested it for SLF. Future work should be used to make the model applicable for a wide range of invasive species.

Test period: Our test set only spans 2 years, 2024 - 2025. Our validation window increases as additional data becomes available each year.

4.3 Practical Implications

Our forecast provides actionable information for information policy makers and the agricultural industry:

Early detection: Counties predicted to be infested within 1-5 years should begin prevention and surveillance strategies. They should also begin to develop mitigation and eradication plans.

Long term prevention: Counties and regions predicted to be infested in 5-10 years should begin to allocate resources and prepare for a potential future infestation.

5 Conclusion

We have presented a spatiotemporal machine learning model for predicting invasive species dispersal. We compared environmental only, coordinate based, and neighbor based approaches. Applied to the Spotted Lanternfly invasion across the continental United States, our models achieved strong performance on test data (AUC ranging from 0.867 to 0.968) and created a realistic 10 year forecasts at the county level.

Our key findings revealed trade offs which must be considered for real life use. Coordinate based models (Model 2, AUC=0.968) achieve the highest accuracy for short term predictions, spatial lag models (Model 3, AUC=0.945) provide a greater understanding of mechanics, and a more balanced prediction which may better generalize to long term forecasts and unknown regions. The significant performance improvement of both geographic models over environmental features alone shows that geographic features are critical for predicting the dispersal of invasive species.

For multiyear forecasts, we use a spatial lag feature that calculates neighboring infested counties at each time step, which captures the expanding infestation. Combined with spread constraints (prediction threshold=0.7, neighbor threshold=0.4), our model creates realistic forecasts that avoid over prediction.

This project supports the use of proactive invasive species prevention over reactive mitigation. Enabling resources to be allocated to counties most at risk. As global trade continues to cause invasions, predictive tools will become increasingly useful for protecting agriculture and natural ecosystems.

Acknowledgments

References

[1] Lawrence Barringer, Claire M Ciafré, Worldwide Feeding Host Plants of Spotted Lanternfly, With Significant Additions From North America, Environmental Entomology, Volume 49, Issue 5, October 2020, Pages 999–1011, https://doi.org/10.1093/ee/nvaa093

[2] Andrew C Dechaine, Mark Sutphin, Tracy C Leskey, Scott M Salom, Thomas P Kuhar, Douglas G Pfeiffer, Phenology of Lycorma delicatula (Hemiptera: Fulgoridae) in Virginia, USA, Environmental Entomology, Volume 50, Issue 6, December 2021, Pages 1267–1275, https://doi.org/10.1093/ee/nvab107

[3] D D Calvin, J Keller, J Rost, B Walsh, D Biddinger, K Hoover, B Treichler, A Johnson, R T Roush, Spotted Lanternfly (Hemiptera: Fulgoridae) Nymphal Dispersion Patterns and Their Influence on Field Experiments, Environmental Entomology, Volume 50, Issue 6, December 2021, Pages 1490–1504, https://doi.org/10.1093/ee/nvab104

[4] Houping Liu, Seasonal Development, Cumulative Growing Degree-Days, and Population Density of Spotted Lanternfly (Hemiptera: Fulgoridae) on Selected Hosts and Substrates, Environmental Entomology, Volume 49, Issue 5, October 2020, Pages 1171–1184, https://doi.org/10.1093/ee/nvaa074

[5] Laura J Nixon, Sharon K Jones, Lisa Tang, Julie Urban, Karen Felton, Tracy C Leskey, Survivorship and Development of the Invasive Lycorma delicatula (Hemiptera: Fulgoridae) on Wild and Cultivated Temperate Host Plants, Environmental Entomology, Volume 51, Issue 1, February 2022, Pages 222–228, https://doi.org/10.1093/ee/nvab137

[6] Hannah L Owen, Fang Meng, Kristin M Winchell, Urbanization and environmental variation drive phenological changes in the spotted lanternfly, Lycorma delicatula (Hemiptera: Fulgoridae), Biological Journal of the Linnean Society, Volume 143, Issue 4, December 2024, blae099, https://doi.org/10.1093/biolinnean/blae099

A Use of AI Tools

This research utilized AI assistance for:

- Code visualization assistance
- Debugging

B Data Availability

All data sources are publicly available:

- SLF and Tree of Heaven reports: EDDMapS (https://www.eddmaps.org/)
- County boundaries: US Census TIGER/Line (https://catalog.data.gov/dataset/tiger-line-shapefile-current-nation-u-s-state-and-equiv
- Urban-Rural codes: USDA ERS (https://www.ers.usda.gov/data-products/rural-urban-continuum-codes)

C Code Repository

Code, models, and data can be found at: https://github.com/jacksonbauer2024S/CS4824Capstone/