Predicting the Presence of Invasive Species

Nick Romano, Jared Rosenberger, Kirin Sharma, Ferdinand Yeke

CS 430-01: Machine Learning in the Cloud

December 9, 2024

Predicting the Presence of Invasive Species

Invasive species pose a problem when it comes to a given environment. Invasive species can hurt the growth of neighboring species as they compete for the same resources to survive or more importantly impact the health of species that are native to the area. As a result, it is important to be proactive in the identification of when invasive species arise. For this project, we used data provided from Olmsted Park Conservatory to predict presence levels of invasive species within Cherokee, Seneca, and Shawnee Parks. The invasive species we are concerned with include: Lonicera japonica, Microstegium vimineum, Euonymus alatus, Ampelopsis brevipedunculata, Morus alba, Celastrus orbiculatus, Vinca minor, Euonymus fortunei, Akebia quinata, Fallopia japonica, Achyranthes, Hederea helix, L. maackii, Privet/Ligustrum, Ailanthus, and Canopy gap and/or downed trees present.

In this project, we explore the data pertaining to the 2022 Invasive Survey data provided to use from the Olmsted Park Conservatory. Then we narrow our focus down to Cherokee, Seneca, and Shawnee Parks and compare the Invasive Survey data for the years 2022 and 2024 to look out for any trends. Finally, we built K-Nearest Neighbor and Random Forest classifier models, that will predict whether a given trimble station has an invasive species with a concerning presence level, measured by area of the trimble station. We deemed that a concerning level for an invasive species to be 1-5% or more of the total area of a trimble station. The 1-5% lower threshold was decided on since it is somewhat early to reverse its spread but if left untreated can lead to further problems down the road. The data used for the models were data that comes from Cherokee, Seneca, and Shawnee parks using data that combines their 2022 and 2024 Invasive Survey data. The motivating questions that we seek to answer are: Whether or not we can accurately predict invasive species with concerning or high presence levels by using the trimble station characteristics, excluding the information on invasive species? Also, what non-invasive species trimble station characteristics contribute to higher presence levels of invasive species?

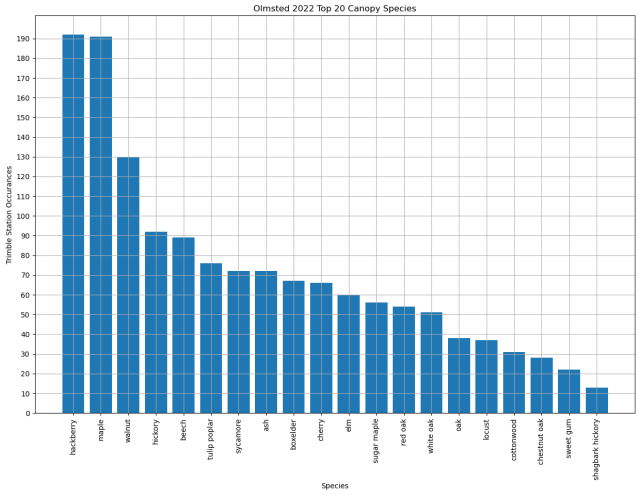
Data Cleaning

For this project, the first steps including cleaning the 2022 and 2024 Invasive Species Survey Excel sheets provided by the Olmsted Park Conservatory. During this process, there were a few inconsistencies in recorded observations when it came to the Canopy and Understory species columns. Specifically, inconsistent delimiting with a comma and an occasional typo. After fixing these two columns in Excel, we imported the data into Python. In Python, there were a handful of trimble stations that were scratched or didn't have data recorded for them, as a result these trimble stations were removed. When it came to the invasive species columns there missing observations. For this we assumed that a missing observation for the presence level of an invasive species indicated an absence within a trimble station. So, we filled these missing values with 0. Finally, we encoded the canopy and understory columns. With help of the US Department of Agriculture's plant database, we searched for the common names of the invasive species and removed any that were within the canopy and understory columns. This was due to the fact that since we are planning to use the canopy and understory species to determine whether an invasive species has a concerning presence level, having the invasive species in the canopy and understory columns would introduce some bias. Encoding the columns resulted in columns being created for the common canopy species and common understory species based on presence in the canopy and understory respectfully. Once we got the data cleaned, we moved onto exploring the data.

Exploratory Data Analysis 1: 2022 Invasive Species Survey

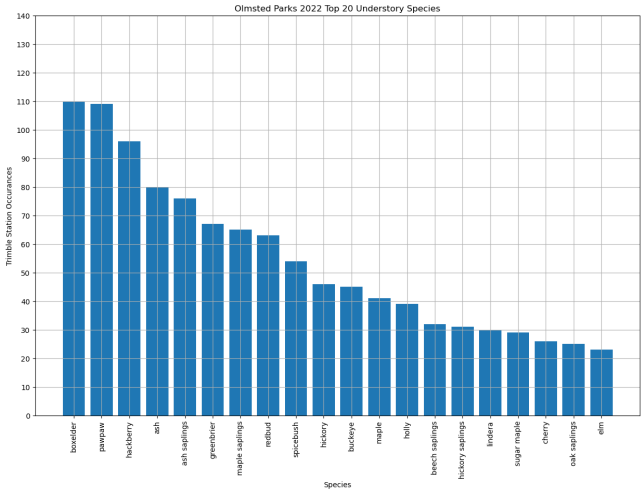
Our first exploration into the data is with the 2022 Invasive Species Survey. This survey consists of data pertaining to Bingham, Chickasaw, Cherokee, Iroquois, Seneca, Shawnee, Tyler, Seminary Parks. We wanted to gain insight to understanding the data related to the makeup of the Olmsted parks to see park system wide patterns. Within this exploratory data analysis, we try to gain insight into which common species occur within more trimble stations across the Olmsted Parks. We also investigate the how the invasive species differ in their presence levels across the park system. Finally, we want to see if there is any connection between the invasive species.

Graph 1: Top 20 Common Species Across Parks in 2022



When it comes to the more common canopy species, we looked at the number of trimble stations that were noted as being the among the common species of a trimble station. Across the parks included in the 2022 Invasive Species Survey, hackberry, maple, and walnut were noted as being the most common as each of these species were noted to be prevalent in over 120 trimble stations, which is about a quarter of the number of trimble stations within the 2022 Invasive Species Survey. This indicates that the hackberry, maple, and walnut species have been able to be succeed within the parks. After these species, there is a slight drop-off down to the third more common species hickory, which is noted as being a more common species in just over 90 trimble stations. The rest of the top 20 common canopy species can be observed in the Graph 1. In creating this graph, we gain the insight of being able to visualize the species of the canopy. Although it only concerns the more occurring species, it ties into our goal of predicting invasive species in that we have a better understanding of the canopy species, which is about of the makeup of the trimble stations.

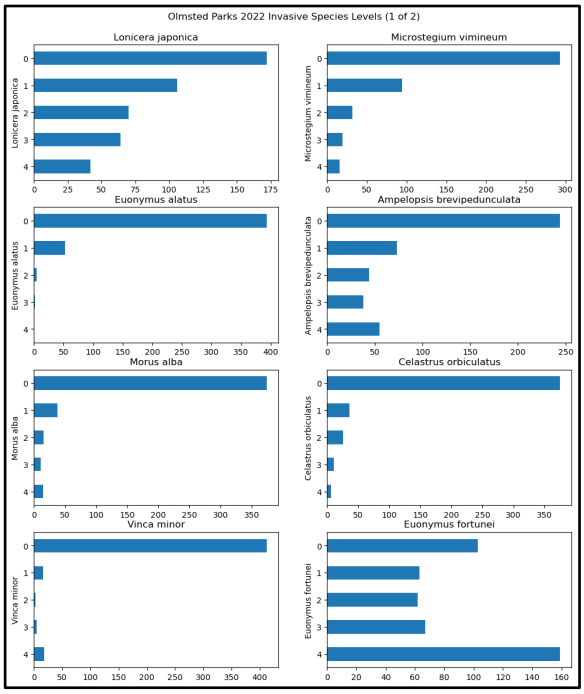
Graph 2: Top 20 Understory Species



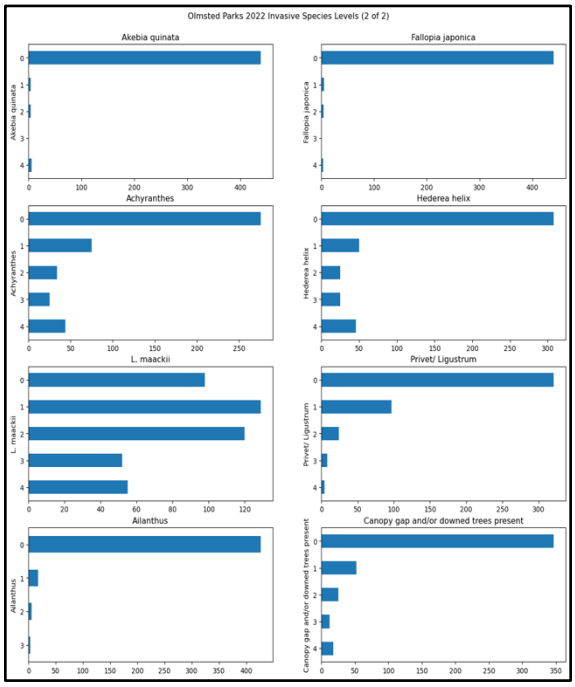
After getting a sense of the more common canopy species we shifted to getting an understanding of the more common understory species. Like the common canopy graph, Graph 2 contains the 20 understory species that were noted to be the most common across the trimble stations. The species noted to be as being the most common are Boxelder, pawpaw, hackberry, ash, and ash saplings, all of which are noted be prevalent in 70 trimble station. Something unique about the understory data is that there are saplings of species along with species. We believe that understory species that were not listed as saplings to be more matured understory species. This can be seen in Graph 2 as ash saplings, maple saplings, beech saplings, hickory saplings, and oak saplings make an appearance in the more common prevalent species across the parks in 2022.

With a general understanding of the makeup for the canopy and understory species, we move over to investigate the invasive species. Since in the end we are wanting to predict whether an invasive species is at or past a certain area of the total area of the trimble station area, we created bar charts to visualize the presence levels for each of the invasive species we are predicting. To note, the presence level ordinal values are as follows: 0 (meaning absence or no observation), 1 (indicating 0-1% of total area of a trimble station), 2 (1-5% of the total area of a trimble station), 3 (5-25% of the total area of a trimble station), 4 (25% or more). Graphs 3 and 4 show the

Graph 3: Invasive Species Levels Across Parks in 2022

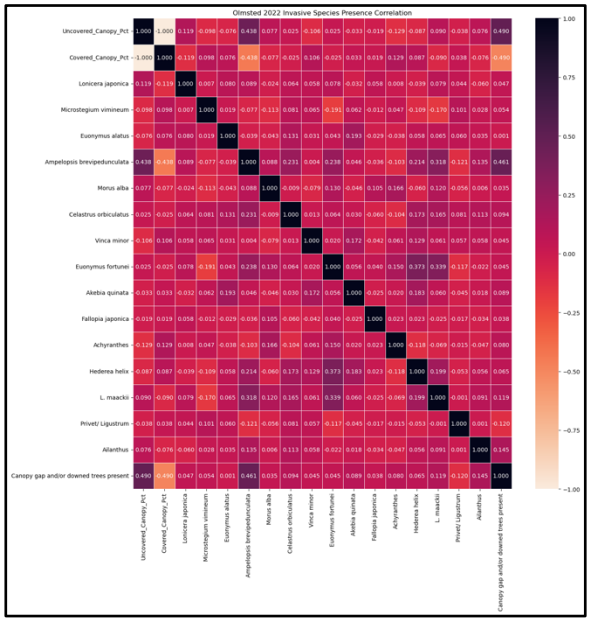


Graph 4: Invasive Species Levels Across Parks in 2022



number of trimble stations that have a particular level of each invasive species. For the most part, it appears that the invasive species have a skewed distribution in terms of their presence levels. An absence present, indicated by a 0, appears to be the most common level presence for the invasive species, then you see a gradual or steep decline in the number of trimble stations with higher presence levels. However, there are a few invasive species that don't follow this shape/distribution. Euonymus fortunei, Lonicera japonica, and L. maackii all have higher presence levels in significantly more trimble stations compared to other invasive species. This provides an insight into the fact that these three invasive species require more focus in dealing with. On top of knowing the number of trimble stations in which the degree of invasive species presence occurs in; we were wondering if there are relationship between the presence levels of the invasive species. To do this, we created a correlation matrix heatmap which is represented by Graph 5. The correlation heatmap calculates the correlation between each set of invasive species across the trimble stations surveyed in 2022, along with the trimble station's canopy coverage percentage and uncovered canopy percentage. The darker the value in the heatmap indicates a stronger positive correlation, while a lighter color indicates a stronger inverse correlation. The dark line running down the middle diagonal is the correlation between an invasive species and itself, which provides no information since everything has a perfection correlation with itself. The only relatively strong relationships are between the

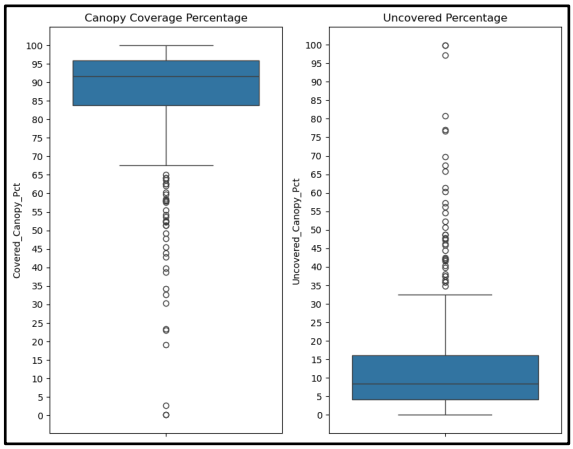
Graph 5: Invasive Correlation Heatmap



canopy coverage and canopy uncovered percentage, the canopy gap and the canopy coverage statistics, between Ampelopsis brevipedunculata and canopy gaps, and between Euonymus fortunei and Hederea helix and L. maacki. Other than that there aren't really any relative strong relationships, positive or negative.

Finally, to wrap up the EDA on the 2022 Invasive Species Survey across all the parks, we looked at the canopy coverage percentage values. Doing so allows us to gain insight on the distribution of the coverage caused by the canopy. The distribution was visualized using a box plot which visualizes the minimum, 25-50-75 percentiles, and the maximum values. We decided to use a boxplot mainly to evaluate the presence of outliers. Graph 6 illustrates the distributions for the canopy coverage and uncovered percentage, which are inverses of each other. But of interest are the large number of outliers within the canopy coverage values. From the graph, we can determine that the majority of the canopy coverage percentage falls between 85% or so and 95%, as indicated by the interquartile range, represented by the box in the left subplot of Graph 6.

Graph 6: Canopy Coverage Boxplot



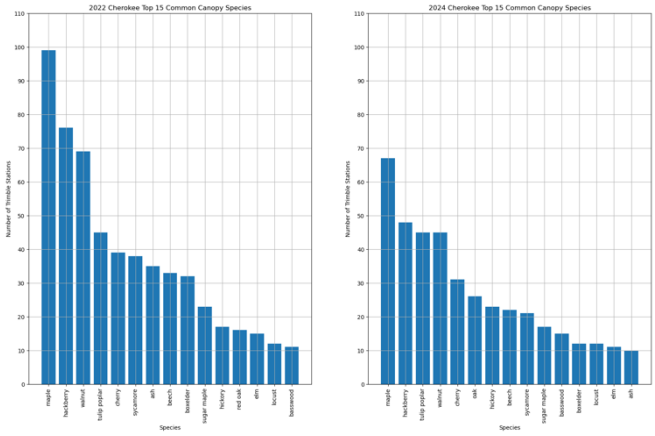
With some insight into the trimble stations statistics of the trimble stations in Bingham, Chickasaw, Cherokee, Iroquois, Seneca, Shawnee, Tyler, Seminary Parks in the 2022 Invasive Species Survey, we narrow our focus to Cherokee, Seneca, and Shawnee Parks to get a better understanding of individual parks. Furthermore, we compare the data that was collected for the 2022 and 2024 Invasive Species Surveys for these parks.

Exploratory Data Analysis 2: Comparing 2022 and 2024 Data for Cherokee, Seneca, & Shawnee

Initially, considering the data for all the parks we have is extensive; so, as a result, we elected to narrow our view to a few parks. These parks being Cherokee, Seneca, and Shawnee. For this EDA, we compare the surveyed information for these parks that took place in 2022 and 2024. Like the first EDA,

we investigate the same trimble station characteristics but are comparing differences in the data for the two invasive surveys.

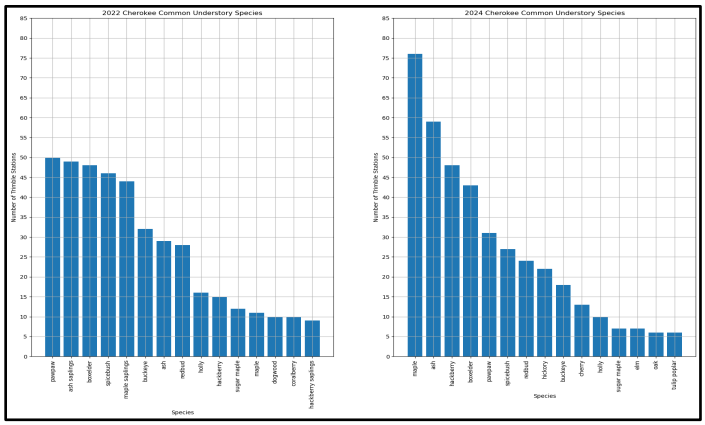
Graph 7: 2022 & 2024 Cherokee Common Canopy



Cherokee Park. When it comes to the number of trimble stations surveyed for Cherokee Park in the 2022 and 2024 Invasive Species Surveys, the number of trimble stations are close. Accounting for scratched trimble stations or trimble stations without any data, there are 164 trimble stations we have data on for Cherokee Park in 2022 and 146 trimble stations we have data on for Cherokee Park in 2024.

In comparing the more common canopy species for 2022 and 2024 (Graph 7), maple, hackberry, walnut, tulip poplar, and cherry are still at the top of the more occurring species as they are top 5 in both years. This indicates that the more thriving species in Cherokee Park are still succeeding and thriving when the survey was conducted again in 2024. This is supported by the fact that in 2022, maple, hackberry, and walnut occurred in about half of the trimble stations that were surveyed and occurred in just under half of the trimble stations surveyed in 2024. Although there is a drop-off in the number of occurrences for the canopy species for 2024, due to having fewer trimble stations recorded. Furthermore, there is a drop-off in the occurrence of ash from 35 down to 10 trimble stations, which indicates that ash is occurring in less trimble stations or other species are taking its place as the more common species in a handful of trimble stations. This provides insight that the common canopy species appears to be consistent for both surveyed years.

Graph 8: 2022 & 2024 Cherokee Common Understory

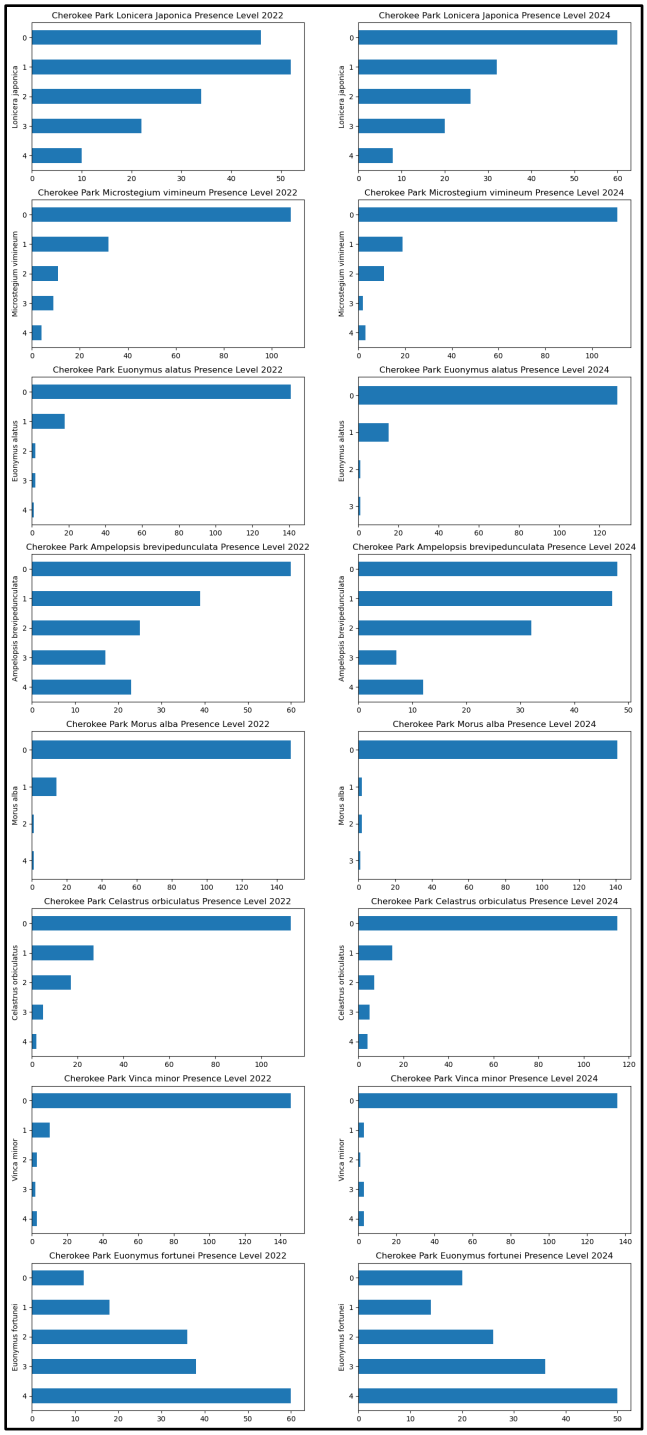


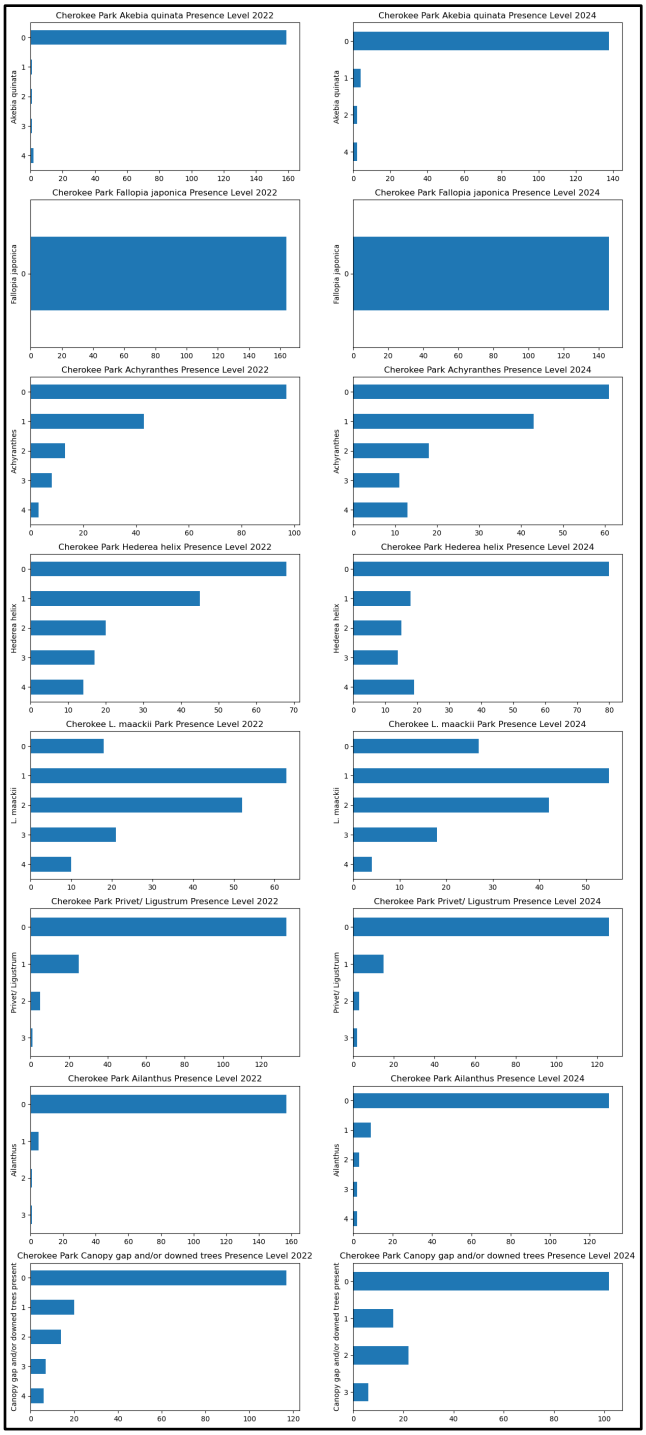
In comparing the more common understory species (Graph 8), the first thing that noticed was the jump in occurrences for

both ash and maple in the 2024 survey. This is because in the 2022 survey, ash saplings were the second most occurring species with (presumably matured) ash at 7th. During the time between surveys, the saplings developed and matured, resulting in an increase in ash. The same can be said for maple as maple saplings were the 5th most common species in 2022 with (presumably mature) Maples at 12th. So in between surveys the maple saplings developed resulting in an increase in maple. Other than that, pawpaw, boxelder, spicebush, and redbud still remained consistently healthy as they are among the more occurring species in both surveys. Also, there was in increase in hackberry occurring in around 15 trimble stations in 2022 but increasing to just under 50 in 2024. This insight provides us the knowledge that even if there isn't much change among canopy species, understory species can quickly change as either new species arrive, or saplings mature.

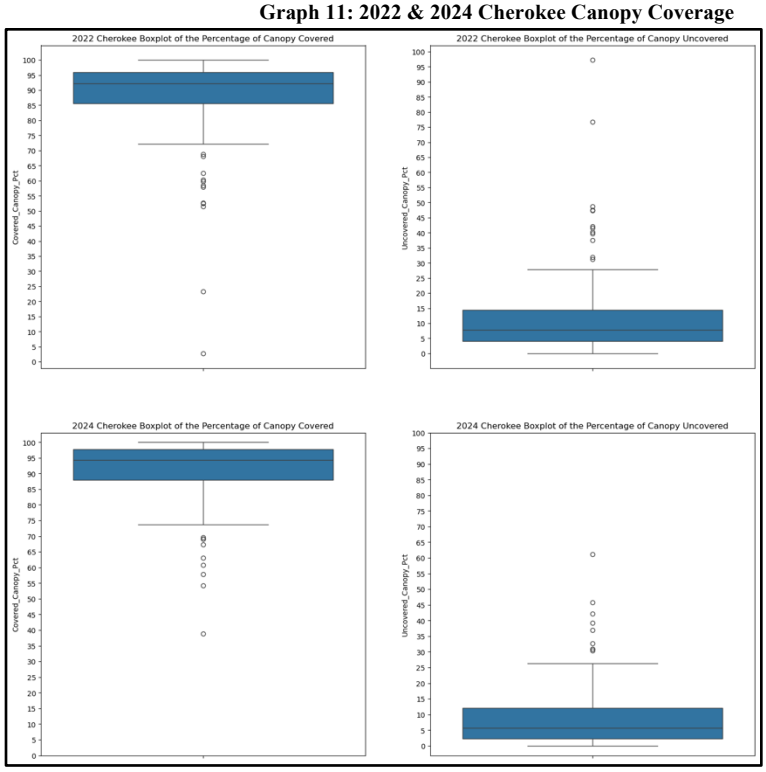
When it comes to the presence levels of the invasive species (Graphs 9 and 10), it appears that the distribution of the presence level is right skewed, if the distributions were positioned vertically instead of vertically, where the peak of the distribution is further left than in the middle for both surveyed years. Most invasive species have a large portion of their measurements being an absence in the trimble stations surveyed then very few instances of the higher instance levels. This indicates that there are more trimble stations with a small presence of the invasive species, if at all. However, there are a few invasive species for which they don't follow this pattern, as they have more of their data in the scarce (0-1%), few (1-5%), some (5-25%), and many (25%) levels than in the absence 0% level. These include: Lonicera Japonica, Ampelopsis Brevipedunculata, Euonymus Fortunei, and L. Maackii. The only invasive species to not be present is Fallopia Japonica with being absent in all trimble stations surveyed in both years.

Graphs 9 & 10: 2022 and 2024 Cherokee Invasive Species Levels





Graph 11: 2022 & 2024 Cherokee Canopy Coverage

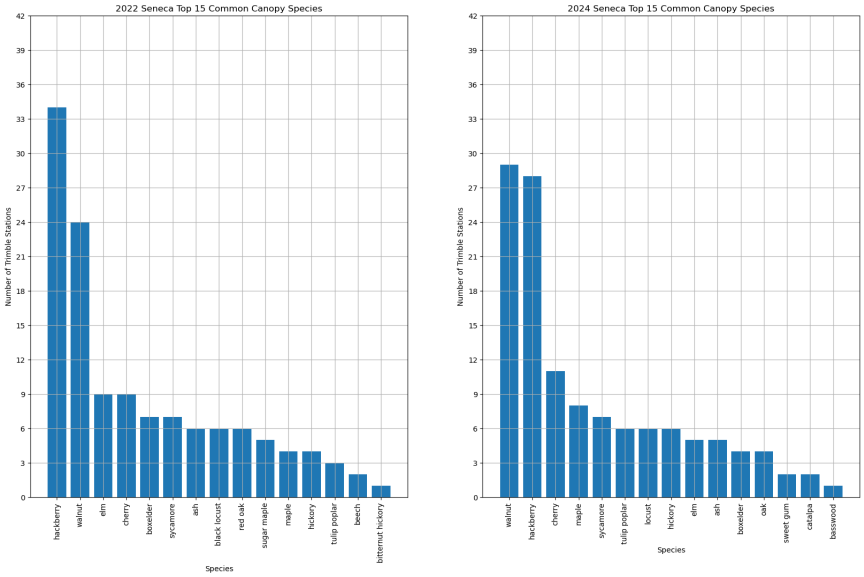


Lastly, we checked to see if there were any changes to the canopy coverage percentages for Cherokee Park trimble stations. We investigated this by created boxplots to investigate the distributions of canopy coverage statistics. Since the canopy coverage and uncovered canopy coverages is just calculated by subtracting one of the percentages by 100%, our attention is more on the covered canopy percent. Both are variables are plotted in Graph 11. The first thing that sticks out is the presence of outliers. This indicates that there are a few trimble stations for in both 2022 and 2024 that have differing canopy coverings as compared to the other trimble stations. Secondly, the quartile range values (min-25percentile-50percentile-75percentile-max) for the covered canopy precent are 2.76-85.635-92.2-95.905-100.00. Whereas the 2024 quantile ranges are 38.9-87.91-94.28-97.66-100.0. In comparing the two sets of values, the interquartile range (25 percentile-50 percentile-75percentile) has not changed much 85.635-92.2-95.905 to 87.91-94.28-97.66. This suggests that there has been more coverage gained within the trimble stations of Cherokee Park in 2024.

Seneca Park. When it comes to the number of trimble stations surveyed for Cherokee in the 2022 and 2024 Invasive Species Surveys, both surveys had information regarding 52 trimble stations, which were the same trimble stations in both surveys.

In comparing the more common canopy species for 2022 and 2024 (Graph12), Hackberry, Walnut, Cherry, and Sycamore have been able to stay successful in their environments as they are amongst the more prevalent of the more common species in both surveys. Unlike the Cherokee Park, the

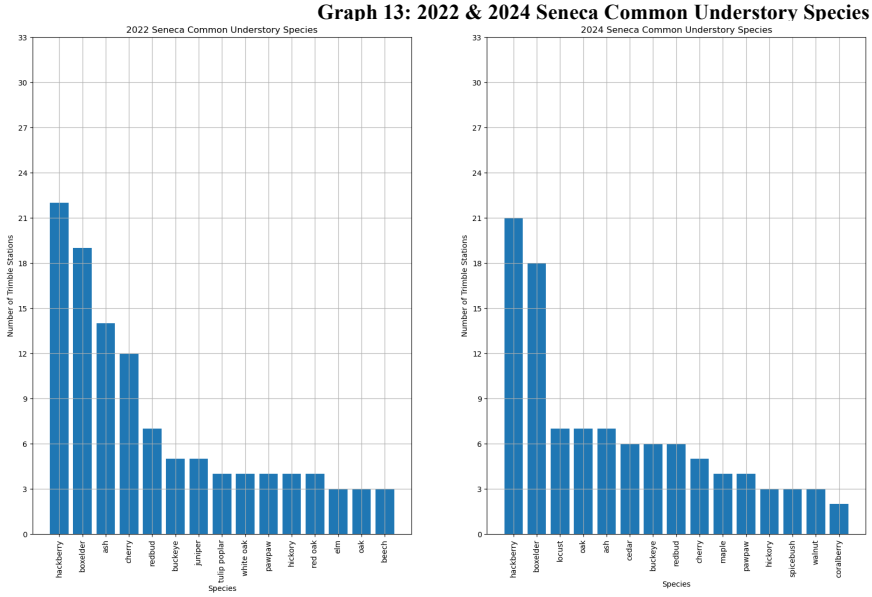
Graph 12: 2022 & 2024 Seneca Common Canopy Species



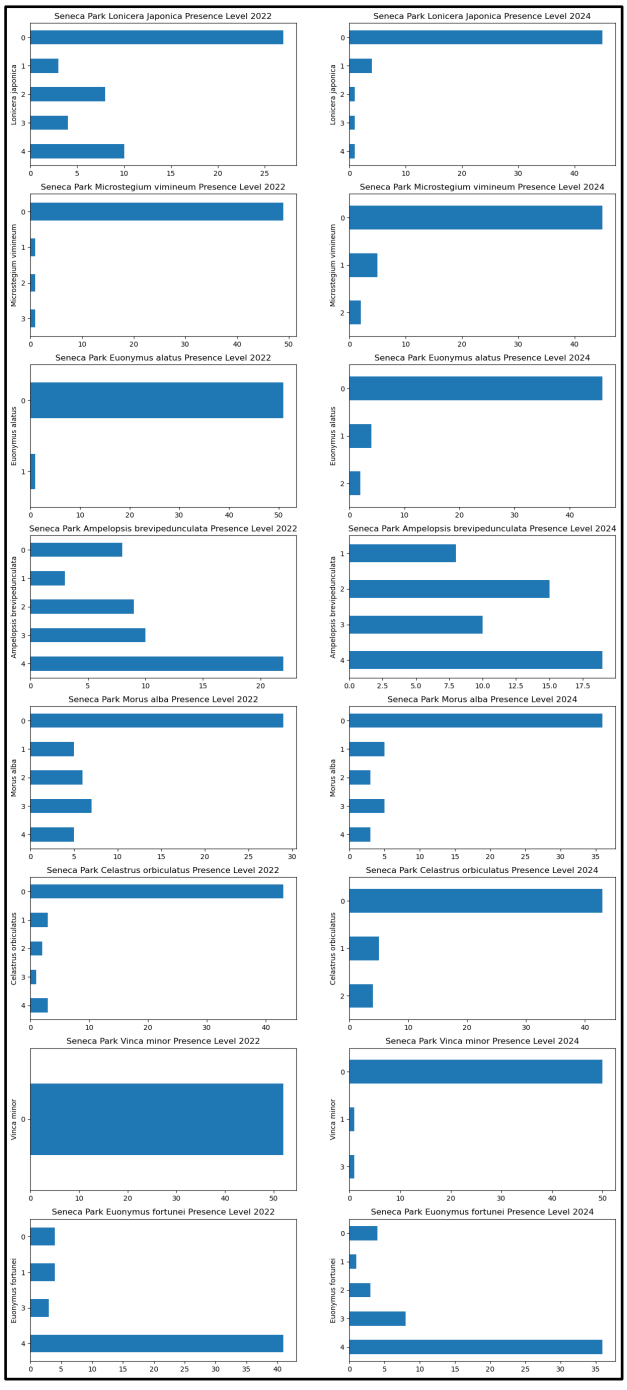
drop off between the two most common species with the rest isn't as smoothed out. The drop-off from second to third in both years is over 15 trimble station occurrences. This indicates that Walnut and Hackberry seem to be the more successful species, in terms of the number of trimble stations, within Seneca park.

When evaluating Seneca's understories over the two surveyed years (Graph 13), it appears that hackberry and boxelder are the more abundant species as they had a leg up in occurrences in 2022 and 2024, as they occur across more trimble stations as other species. This may indicate that other species maybe having trouble having success within the trimble stations within Seneca park. Like, the canopy species, there exists a significant drop off from the more abundant understory species. As you see a drop-off of 5 trimble stations between cherry and redbud in the 2022 survey and a drop-off of 11 between Boxelder and Ash in the 2024 survey. This provides insight into Seneca's trimble station characteristics in that the more abundant of the more common understory species are still abundant while some of the other more common understory species are still at similar levels.

Graph 13: 2022 & 2024 Seneca Common Understory Species



Graph 14: 2022 & 2024 Seneca Invasive Species Levels

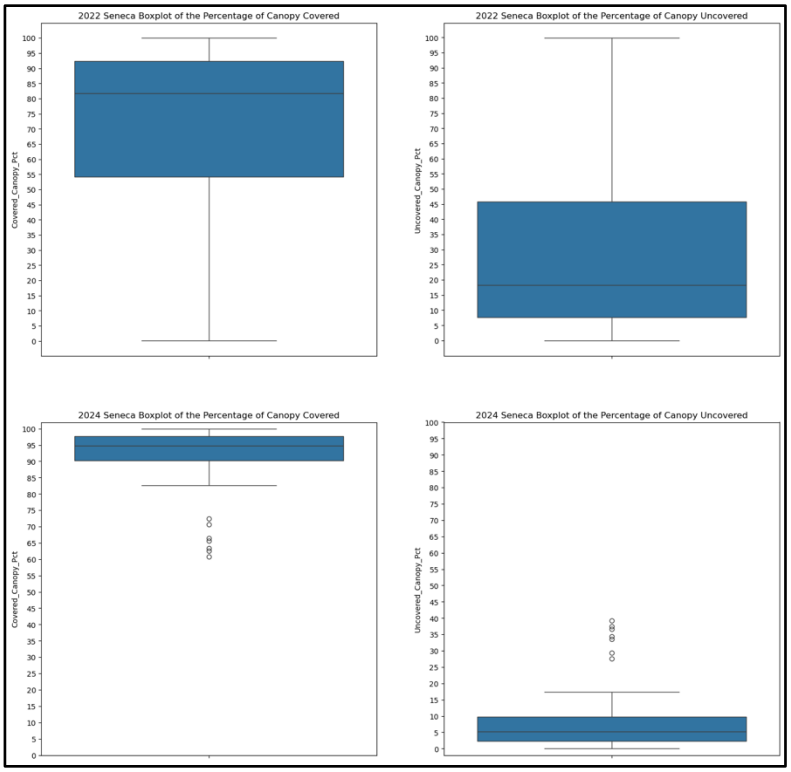


When it comes to the comparing the presence levels of the invasive species in 2022 and 2024 in Seneca Park (Graph 14 and 15), it appears that the distribution of the presence level is right skewed, where the peak of the distribution is further left than in the middle for both surveyed years. Most invasive species have a large portion of their measurements being an absence in the trimble stations surveyed, then having very few instances of higher instance levels. This is similar to the distributions in Cherokee park. However, there are a few invasive species with no presence in either survey: akebia quinta and fallopia japonica. Furthermore, there are a few invasive species whose presence level distributions does stand out from the others: L. Maackii, Euonymus fortunei, and Ampelopsis brevipedunculata. These invasive species appear to have a significant presence level within the park as their distributions are more skewed to higher levels of presence whereas other invasive species are skewed toward lower levels of presence. This gives us a better understanding of the invasive species presence for trimble stations in Seneca Park for the two years surveyed.

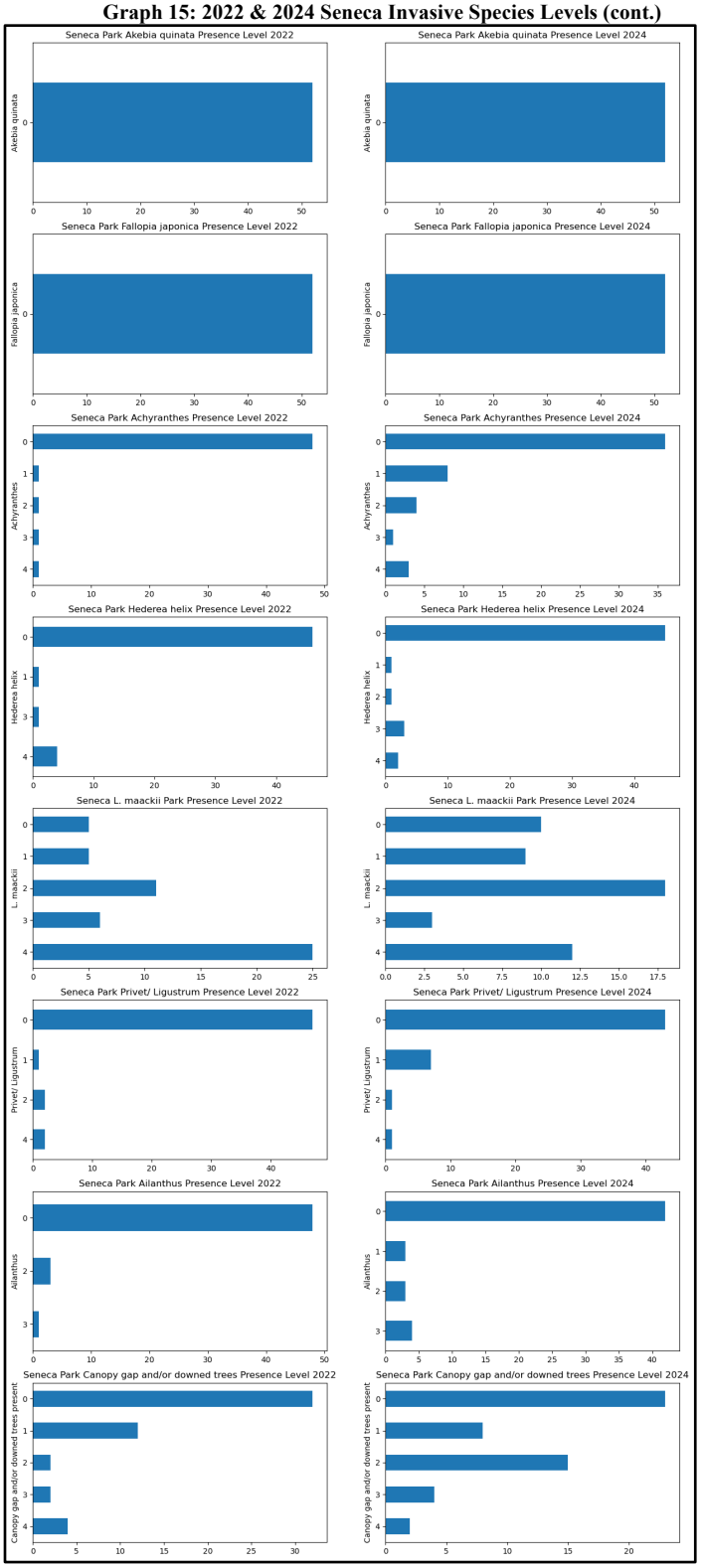
Finally, we compare the canopy percentage coverage for the trimble stations in Seneca Park in the two years surveyed by creating a boxplot (Graph 16). The boxplot gives insight into the distribution of the canopy coverage percentages with the trimble stations within Seneca Park for the two years surveyed. In

looking at the boxplot, there is a considerable difference between the distributions. The main difference comes in the size of the interquartile range of the boxplot of the covered canopy percent boxplots. In 2022, the interquartile range is 38.22 (92.33 – 54.11), whereas the 2024 interquartile range is 7.41 (97.66 – 90.25). This indicates that there was a few trimble stations in which the coverage canopy increased in a drastic way. Evaluating the canopy coverages gives us the insight that overtime the coverage values can change from survey to survey and in different ways from park to park. So, when we are using the canopy coverage value our models we will need to scale the value by park and by year.

Graph 16: 2022 & 2024 Seneca Canopy Coverage Boxplots

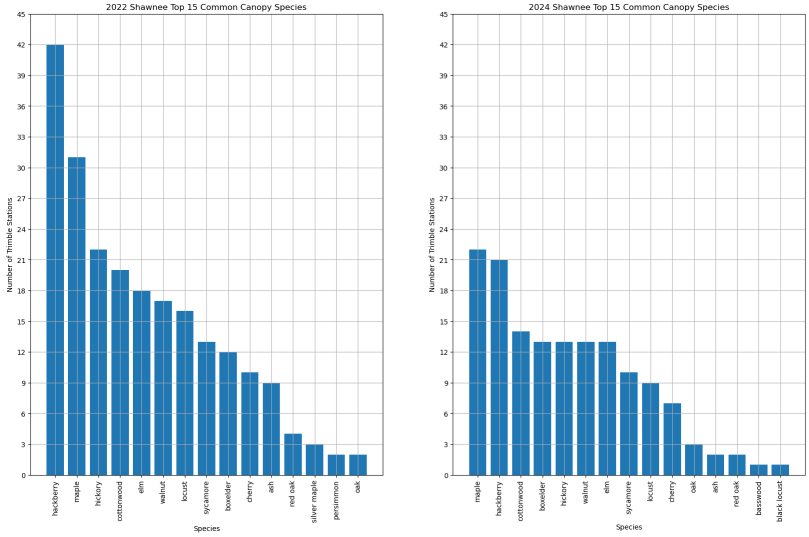


Graph 15: 2022 & 2024 Seneca Invasive Species Levels (cont.)



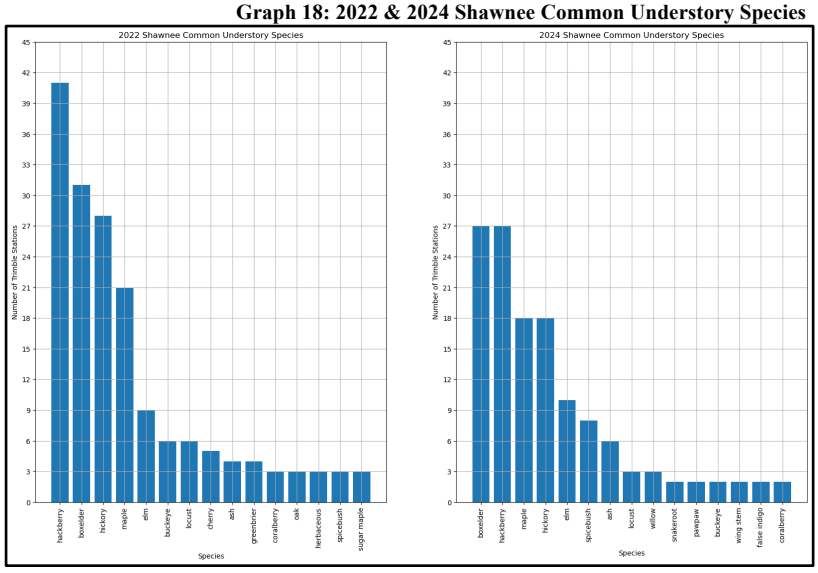
Shawnee Park. The last park that we compare overall trimble characteristics from 2022 and 2024 is Shawnee Park. In 2022, we have clean data for 72 trimble stations; whereas in 2024, we have clean data for 46 trimble stations.

Graph 17: 2022 & 2024 Shawnee Common Canopy Species



In comparing the data between the two surveys for Shawnee Park, we first looked at the common canopy and understory species. In comparing the more common canopy species for 2022 and 2024 for Shawnee Park (Graph 17), hackberry, maple, hickory, cottonwood, elm, walnut, locust, sycamore, boxelder, and cherry all appear in the top 10 of the more common canopy species. This indicates that since the last survey, that these species have managed to be successful in their respected environments as they are regularly occurring within the park. The drop off in the occurrences in trimble stations is due to the fact that in the 2024 survey, only 46 trimble stations were surveyed as in that survey there were a lot of trimble stations with missing entries.

Graph 18: 2022 & 2024 Shawnee Common Understory Species



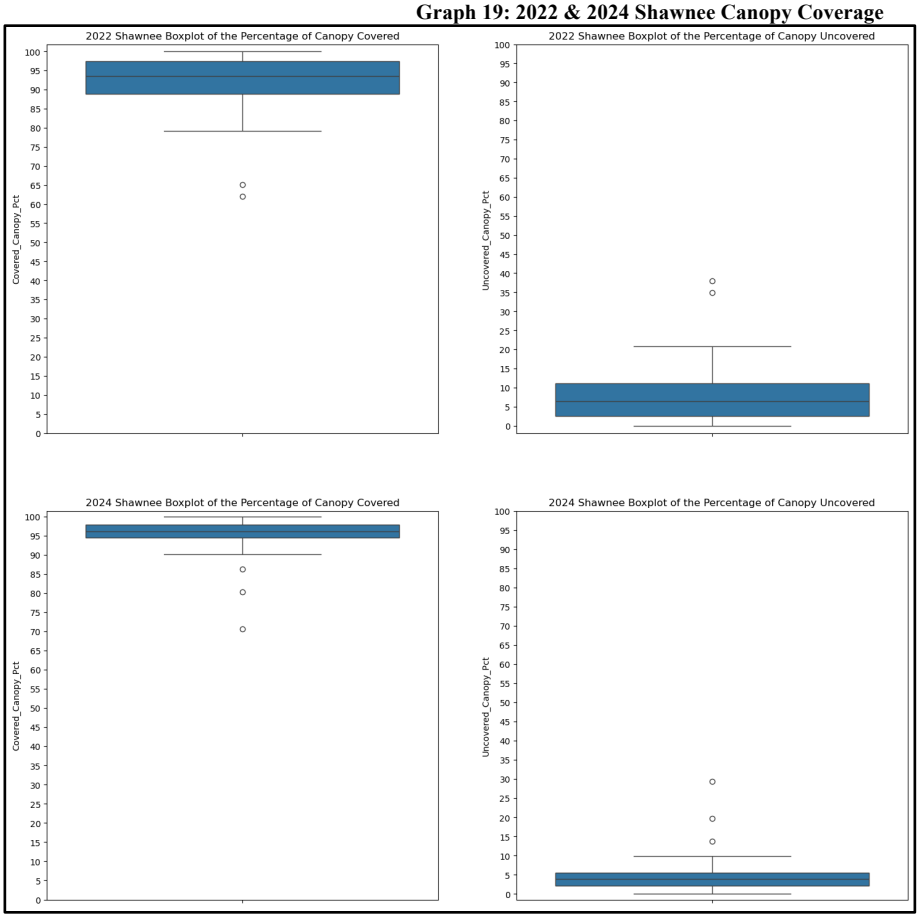
In comparing the more common understory species, boxelder, hackberry, hickory, maple, and elm are a step above the other species as in both surveys these species rounded out the top for occurring in the most trimble stations. Even with the drop off in surveyed trimble stations in 2024, these species occur in a relatively significant number of trimble stations in comparison to other species. This indicates that species

have managed to be successful in their respected environments as they are regularly occurring within the park.

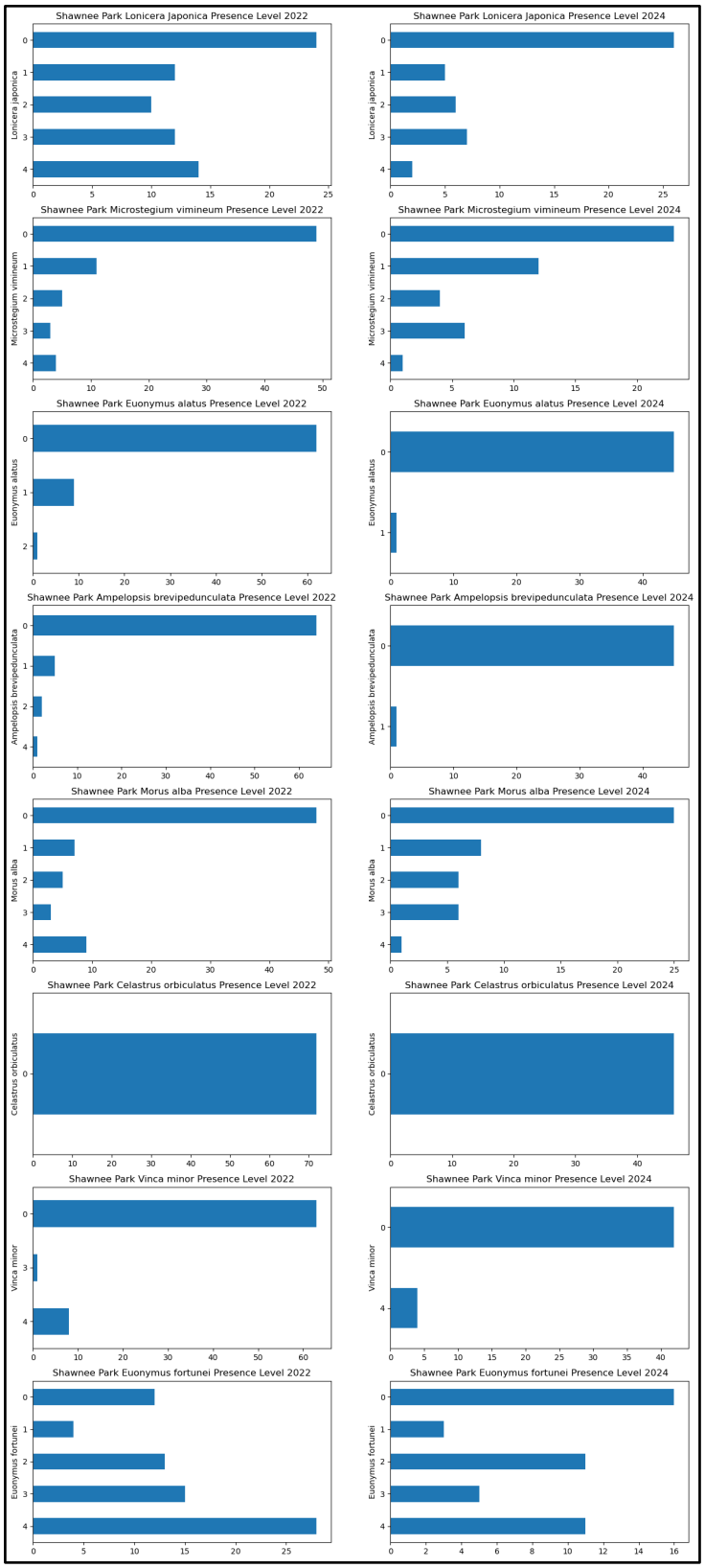
Next, we focused on the invasive species presence levels. Compared to Cherokee and Seneca, there appears to be more invasive species in which there is an absence of that invasive species within trimble stations across Shawnee Park. The trimble stations with no presence include Ailanthus, Akebia quinta, and Celastrus orbiculatus all of which have no presence in any trimble station in both surveys. Like Cherokee you see that a handful of the invasive species have a skewed distribution toward no presence. However, L. Maackii, Achyranthes, Euonymus fortunei and lonicera japonica have distributions in which they more trimbles in which they have a larger presence.

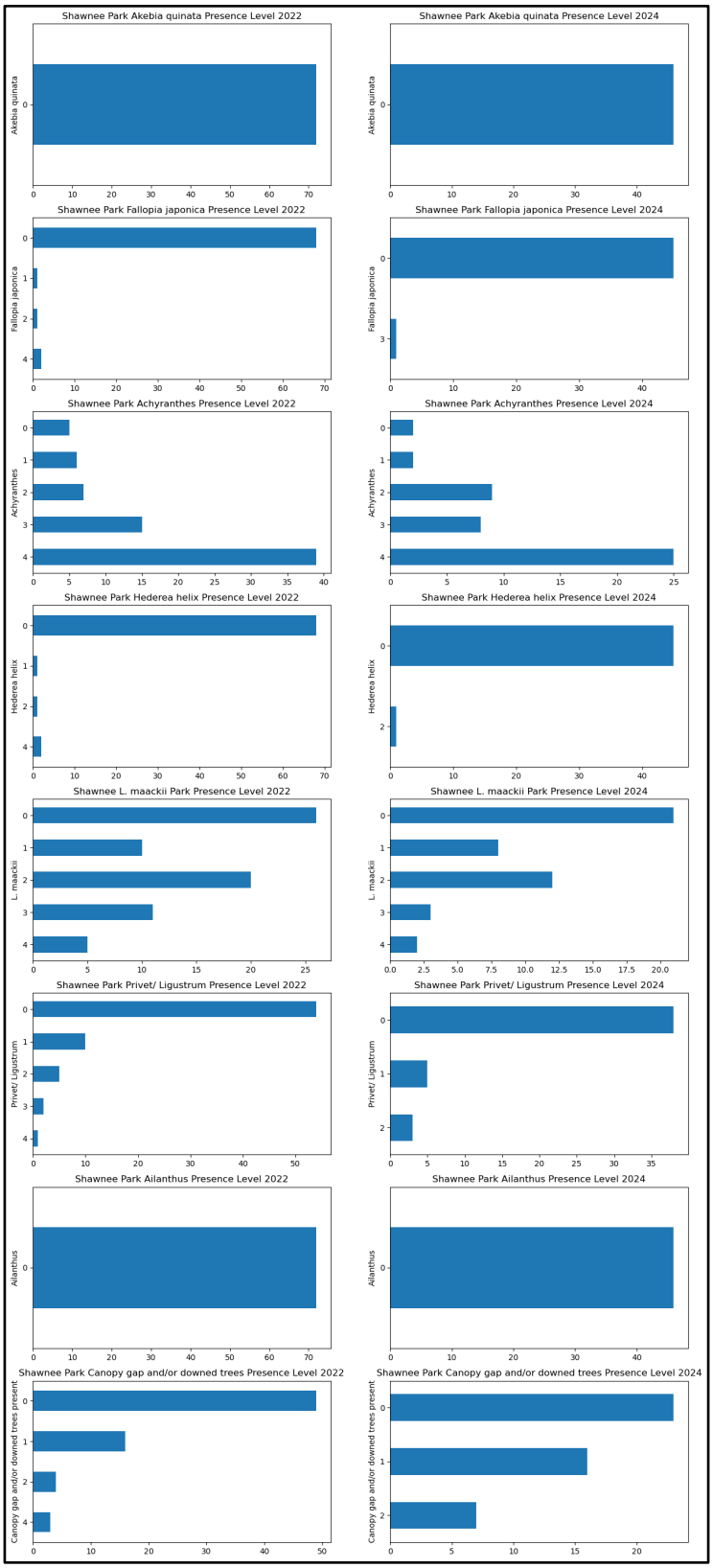
Finally, we focused on the canopy coverage. In comparing the canopy coverage for Shawnee park there spread of the canopy coverage percentages has shrunk in the 2024 survey. As the interquartile range in 2022 is 8.45, whereas the interquartile range for 2024 is 3.315. This could be due to the shortened available data for Shawnee park.

Graph 19: 2022 & 2024 Shawnee Canopy Coverage



Graphs 20 & 21: 2022 & 2024 Shawnee Invasive Species Levels





Machine Learning Models

Now with an understanding of the makeup and differences between the surveyed data pertaining to Cherokee, Seneca, and Shawnee Park, we now move on to the machine learning stage. The goal of this project was to be able to predict whether an invasive species has reached a concerning presence level among a given stations for each of these parks. In other words, given the trimble statistics of the most common species, that are not invasive, we try to predict whether the following invasive species have reached a concerning presence level: Lonicera japonica, Microstegium vimineum, Euonymus alatus, Ampelopsis brevipedunculata, Morus alba, Celastrus orbiculatus, Vinca minor, Euonymus fortunei, Akebia quinata, Fallopia japonica, Achyranthes, Hederea helix, L. maackii, Privet/Ligustrum, Ailanthus, and Canopy gap and/or downed trees present. As mentioned in the introduction, we deemed that a concerning level for an invasive species to be 1-5% or more of the total area of a trimble station. The 1-5% lower threshold was decided on since it is somewhat early to reverse its spread but if left untreated can lead to further problems down the road.

The models that we used were the K-Nearest Neighbor (KNN) Classifier model and the Random Forest models. How a KNN classifier model essential works is that the model is trained on a set of data that belong to certain classes. To make a prediction on a new datapoint, the features, or information, of the new datapoint is compared alongside the of features of N-datapoints that it is trained on. By looking at the N-number datapoints that are most similar to the new datapoint, the classification is made by what class/category that is in the majority of the N-nearest trained datapoints. A Random Forest model works by making various decisions on learning from training data on how to split the data up by class/category based on the features of the data. A Random Forest model iterates many times over to try different splits on the data. We selected KNN as a way of comparing similar trimble stations. If two trimble stations are similar within a park, then they might have similar invasive species presence. Random Forest was selected because of its running many models with the varying splits.

The data that we used for our models were the combined 2022 and 2024 data for Cherokee, combined 2022 and 2024 data for Seneca, and combined 2022 and 2024 data for Shawnee. Before we

could start modeling, we did some manipulations on the data. Which included encoding the canopy and understory columns after removing the invasive species that we are trying to predict from the columns since this would introduce bias, as mentioned in the data cleaning portion of the paper. We also scaled the canopy coverage statistic using the MinMaxScalar, which converts the values to a value between 0 and 1 by subtracting the value by the minimum then dividing the value by the maximum within the array of data. We scaled the covered canopy coverage variable by park and year, for example the covered canopy coverage variable for the 2022 Cherokee data was scaled by itself, the 2024 Cherokee data was scaled by itself, and so on for Seneca and Shawnee. Then, we encoded the invasive species levels to the following format [0:0, 1:0, 2:1, 3:1, 4:1]. This creates the predictions to binary since we want to predict concerning levels for an invasive species to be 1-5% or more of the total area of a trimble station, which are levels 2, 3, and 4. Another issue was that there is an imbalance in the encoded invasive presence levels, this was solved using SMOTE that essentially creates new datapoints for the minority encoded binary presence level.

We used for loops to find the best parameters to use for the KNN and Random Forest Models and then run the best parameter models for each invasive species within Cherokee, Seneca, and Shawnee. For KNN, the parameters that we changed and evaluated for the best fitting models were the number of neighbors (the number of nearby datapoints that are used in the classification of the new datapoint) and weight (whether the model gives equal weight to the nearby datapoints or more weight to the points that are immediately closer). For Random Forest, the parameters that we changed and evaluated for the best fitting models were the number of estimators (The number of decision paths iterated and evaluated) and max depth (the number of splits/decisions made in evaluating the data). Once the best parameters are found then an another for loop was used to create and run a KNN model or Random Forest model with those best parameters for a given species and park.

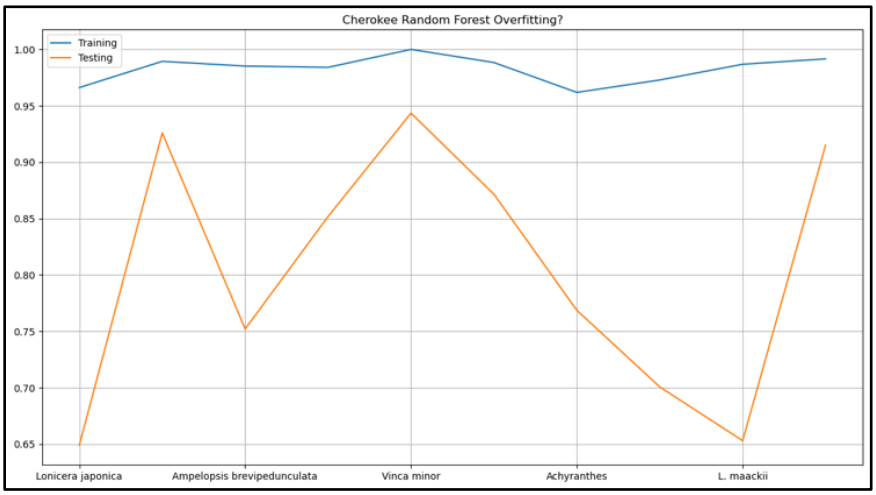
Model Results

Note: there were a few invasive species for each park were the imbalance was too big to use SMOTE to balance out the data. As a result, if an invasive species is not listed within a park model table then that invasive species wasn't modeled because there wasn't enough instances of that invasive species

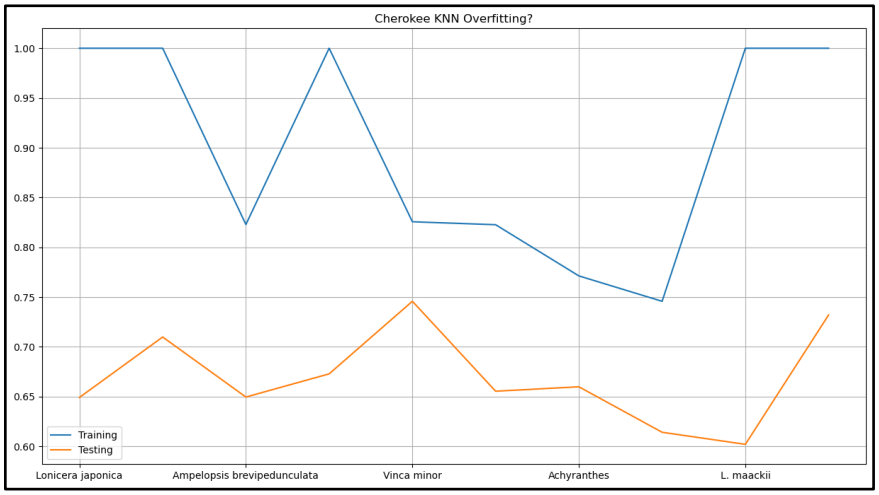
Cherokee Model Results. In comparing the testing accuracy scores of both models for all the invasive species, the random forest model was better at predicting higher/concerning levels of invasive species and downed trees. The lower KNN scores may be due to the lower precision scores indicating that the KNN models had difficulties in predicting more false positives, or predicting a high presence level when there actually isn't, which is seen in the low testing precision scores column for the Cherokee KNN portion of the table below. Despite having higher rates of false positives, the KNN model appeared to do a slightly better job of limiting making false negative predictions, as seen in the slightly higher testing recall score. Overall, the random forest classifier in the end seems to be the better model for the invasive species in Cherokee

Cherokee KNN							
Invasive Species	n_neighbors	weights	**Training Data Accuracy Score**	**Testing Data Accuracy Score**	**Testing Recall Score**	**Testing Precsion Score**	**Testing F1 Score**
Lonicera japonica	15	distance	1	0.649123	0.933333	0.608696	0.736842
Microstegium vimineum	3	distance	1	0.709877	1	0.632812	0.77512
Ampelopsis brevipedunculata	3	uniform	0.822878	0.649573	0.919355	0.612903	0.735484
Celastrus orbiculatus	3	distance	1	0.67284	0.9875	0.603053	0.748815
Vinca minor	3	uniform	0.825666	0.745763	1	0.656489	0.792627
Euonymus fortunei	3	uniform	0.822674	0.655405	0.268657	0.9	0.413793
Achyranthes	3	uniform	0.771261	0.659864	0.973684	0.606557	0.747475
Hederea helix	3	uniform	0.745763	0.614173	0.949153	0.54902	0.695652
L. maackii	7	distance	1	0.602041	0.645833	0.584906	0.613861
Canopy gap and/or downed trees present	3	distance	1	0.732026	0.953488	0.689076	0.8

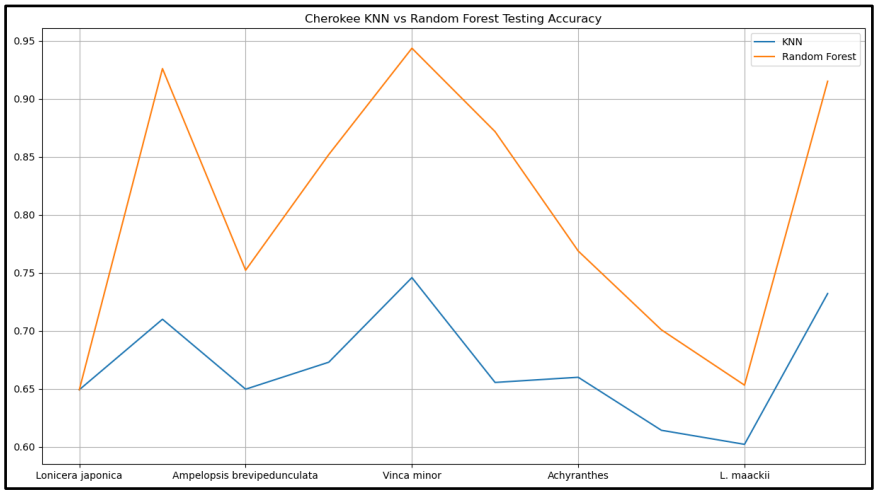
Cherokee Random Forest							
Invasive Species	max depth	n_estimators	**Training Data Accuracy Score**	**Testing Data Accuracy Score**	**Testing Recall Score**	**Testing Precsion Score**	**Testing F1 Score**
Lonicera japonica	15	150	0.966165	0.649123	0.733333	0.647059	0.6875
Microstegium vimineum	15	200	0.989418	0.925926	0.962963	0.896552	0.928571
Ampelopsis brevipedunculata	15	50	0.98524	0.752137	0.822581	0.73913	0.778626
Celastrus orbiculatus	15	150	0.984127	0.851852	0.9125	0.811111	0.858824
Vinca minor	15	200	1	0.943503	0.94186	0.94186	0.94186
Euonymus fortunei	15	200	0.988372	0.871622	0.895522	0.833333	0.863309
Achyranthes	11	100	0.961877	0.768707	0.881579	0.728261	0.797619
Hederea helix	11	250	0.972881	0.700787	0.830508	0.636364	0.720588
L. maackii	15	50	0.986842	0.653061	0.6875	0.634615	0.66
Canopy gap and/or downed trees present	13	150	0.991597	0.915033	0.906977	0.939759	0.923077



Visualization shows some overfitting for the Random Forest models for the invasive species with Cherokee Park as the models have better performance with the training dataset.



Visualization shows possible overfitting for the KNN models for invasive species within Cherokee park as the model has better performance with the training dataset



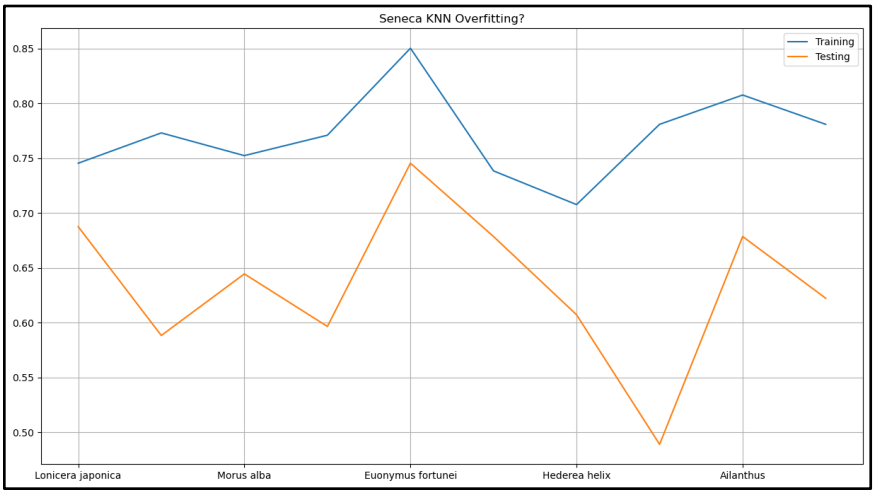
Visualization shows that the Random Forest model is more accurate in predicting invasive species levels across 2022 and 2024 Cherokee Park data

Seneca Model Results. When it comes to the models for Seneca, the problem of overfitting doesn't seem to be as bad of an issue in comparison to the models with Cherokee park. In comparing the testing accuracy scores of both models for all the invasive species, the random forest model was better at predicting higher/concerning levels of invasive species and downed trees. The lower KNN scores may be due to the lower precision scores indicating that the KNN models had difficulties in predicting more false positives, or predicting a high presence level when there actually isn't, which is seen in the low testing precision scores column for the Seneca KNN portion of the table below. The Random Forest models appears to have more consistency in limiting making false positive predictions as the KNN models have recall scores that fluctuate between high and lower recall scores, whereas the Random Forest models are steadily in the 80 and 90 percent ranges.

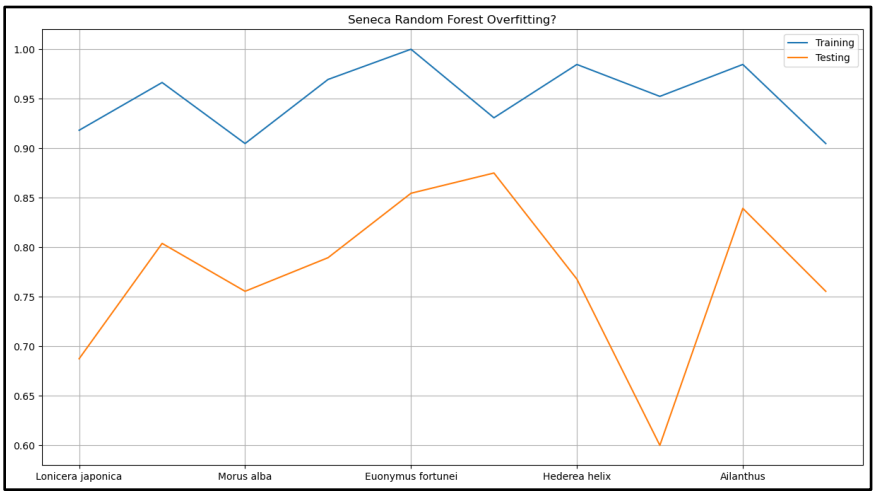
Seneca KNN							
Invasive Species	n_neighbors	weights	**Training Data Accuracy Score**	**Testing Data Accuracy Score**	**Testing Recall Score**	**Testing Precsion Score**	**Testing F1 Score**
Lonicera japonica	3	uniform	0.745455	0.6875	0.958333	0.621622	0.754098
Ampelopsis brevipedunculata	3	uniform	0.773109	0.588235	0.16	1	0.275862
Morus alba	3	uniform	0.752381	0.644444	1	0.578947	0.733333
Celastrus orbiculatus	3	uniform	0.770992	0.596491	1	0.530612	0.693333
Euonymus fortunei	3	uniform	0.850394	0.745455	0.571429	0.888889	0.695652
Achyranthes	3	uniform	0.738462	0.678571	1	0.617021	0.763158
Hederea helix	3	uniform	0.707692	0.607143	1	0.56	0.717949
L. maackii	3	uniform	0.780952	0.488889	0.153846	0.8	0.258065
Ailanthus	3	uniform	0.807692	0.678571	0.965517	0.622222	0.756757
Canopy gap and/or downed trees present	3	uniform	0.780952	0.622222	1	0.540541	0.701754

Seneca Random Forest							
Invasive Species	max depth	n_estimators	**Training Data Accuracy Score**	**Testing Data Accuracy Score**	**Testing Recall Score**	**Testing Precsion Score**	**Testing F1 Score**
Lonicera japonica	7	100	0.918182	0.6875	0.916667	0.628571	0.745763
Ampelopsis brevipedunculata	13	50	0.966387	0.803922	0.64	0.941176	0.761905
Morus alba	5	50	0.904762	0.755556	0.954545	0.677419	0.792453
Celastrus orbiculatus	9	150	0.969466	0.789474	0.961538	0.694444	0.806452
Euonymus fortunei	15	50	1	0.854545	0.892857	0.833333	0.862069
Achyranthes	5	50	0.930769	0.875	0.896552	0.866667	0.881356
Hederea helix	11	150	0.984615	0.767857	0.928571	0.702703	0.8
L. maackii	11	200	0.952381	0.6	0.423077	0.785714	0.55
Ailanthus	9	50	0.984615	0.839286	0.965517	0.777778	0.861538
Canopy gap and/or downed trees present	5	50	0.904762	0.755556	0.9	0.666667	0.765957

Visualization shows that the Random Forest model is more accurate in predicting invasive species presence across trimble stations with the 2022 and 2024 Seneca Data



Visualization compares training and testing data accuracy scores for the Seneca KNN models



Visualization compares training and testing data accuracy scores for the Seneca Random Forest models

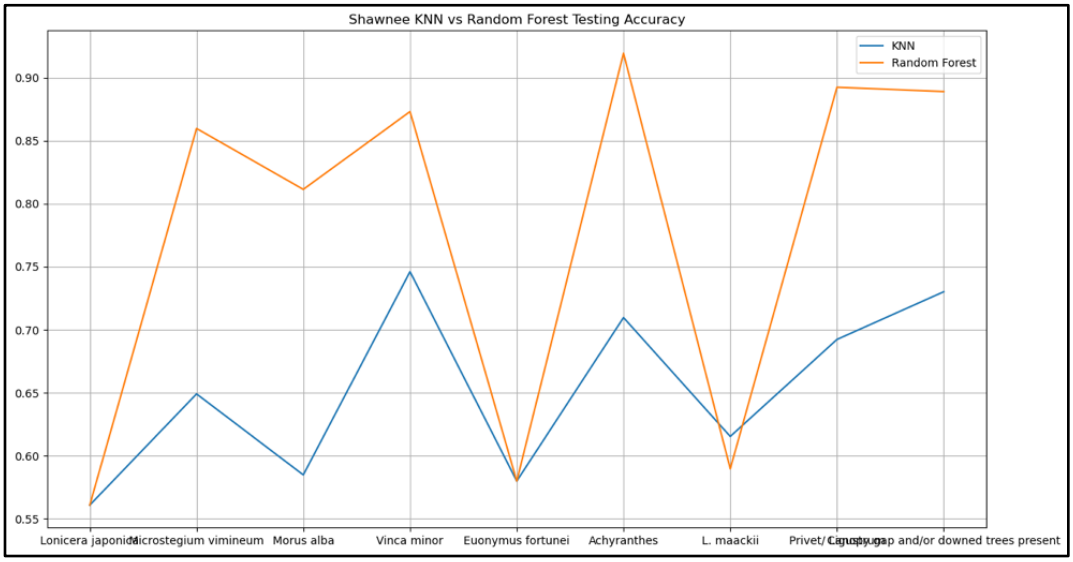
Shawnee Park. Unlike the Random Forest models of Cherokee and Seneca Parks, the Random Forest models for Shawnee appeared to have more difficulty with its precision scores for a few species, meaning that it made more false positive predictions. This species include Lonicera jaonica, Euonymus fortunei, and L. Maacki. Corresponding with the increased proportion of false negatives for these species, the Random Forest model for these species had lower accuracy scores. Other than those 3 species, the Random Forest model seemed to perfrom better than the KNN model, since the KNN model has the same trouble with its precsion scores.

Shawnee KNN

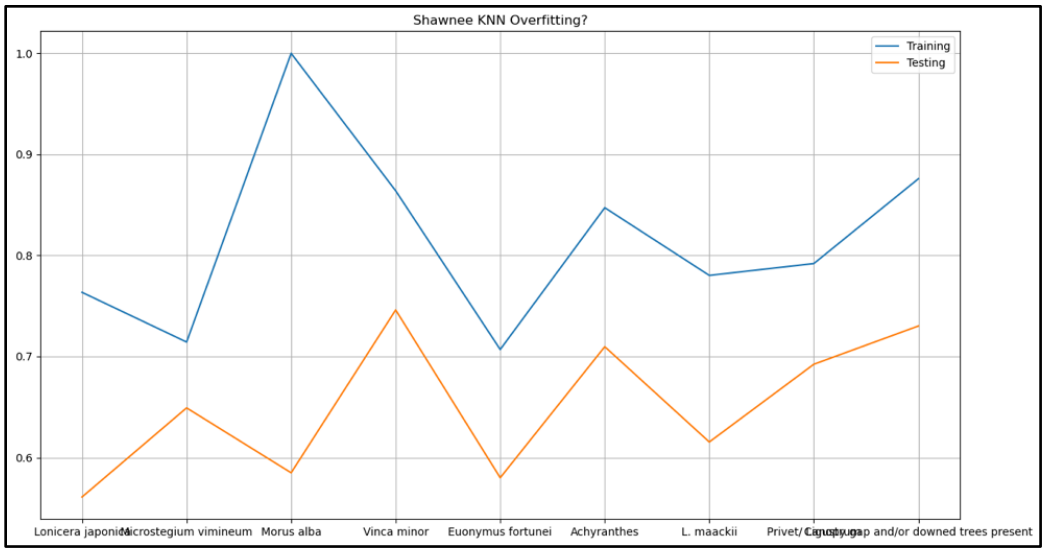
Invasive Species	n_neighbors	weights	Training Data Accuracy Score	Testing Data Accuracy Score	Testing Recall Score	Testing Precsion Score	Testing F1 Score
Lonicera japonica	7	distance	0.763441	0.560976	0.75	0.535714	0.625
Microstegium vimineum	7	uniform	0.714286	0.649123	1	0.565217	0.722222
Morus alba	3	distance	1	0.584906	0.88	0.536585	0.666667
Vinca minor	3	uniform	0.863946	0.746032	1	0.68	0.809524
Euonymus fortunei	13	uniform	0.706897	0.58	0.375	0.6	0.461538
Achyranthes	3	uniform	0.847222	0.709677	0.37931	1	0.55
L. maackii	3	uniform	0.78022	0.615385	0.705882	0.545455	0.615385
Privet/ Ligustrum	3	uniform	0.791946	0.692308	0.972222	0.648148	0.777778
Canopy gap and/or downed trees present	3	uniform	0.875862	0.730159	1	0.645833	0.78481

Shawnee Random Forest

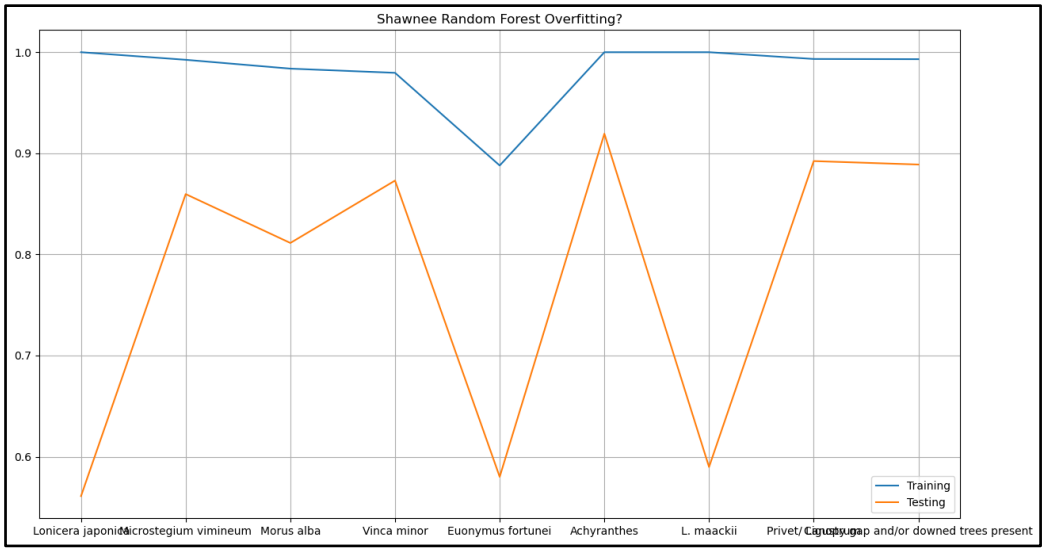
Invasive Species	max depth	n_estimators	Training Data Accuracy Score	Testing Data Accuracy Score	Testing Recall Score	Testing Precsion Score	Testing F1 Score
Lonicera japonica	13	50	1	0.560976	0.65	0.541667	0.590909
Microstegium vimineum	13	200	0.992481	0.859649	0.923077	0.8	0.857143
Morus alba	9	150	0.98374	0.811321	1	0.714286	0.833333
Vinca minor	7	200	0.979592	0.873016	0.911765	0.861111	0.885714
Euonymus fortunei	3	100	0.887931	0.58	0.625	0.555556	0.588235
Achyranthes	13	50	1	0.919355	0.862069	0.961538	0.909091
L. maackii	15	50	1	0.589744	0.764706	0.52	0.619048
Privet/ Ligustrum	15	50	0.993289	0.892308	0.916667	0.891892	0.90411
Canopy gap and/or downed trees present	13	100	0.993103	0.888889	1	0.815789	0.898551



Visualization shows that the Random Forest model is more accurate in predicting invasive species presence across trimble stations with the 2022 and 2024 Shawnee Data except for Lonicera jaonica, Euonymus fortunei, and L. Maacki.



Visualization compares training and testing data accuracy scores for the Shawnee KNN models.



Visualization compares training and testing data accuracy scores for the Shawnee Random Forest models.

Possible overfitting due to noise in testing scores

Conclusions and Recommendations

For our project, we were able to answer are questions. When it come to whether we could acccurately predict the presence of high presence levels of invasive species, the Random Forest models appeared to perfrom better across the three parks that we investigated. Possible applications of this model include forecasting ahead for a given trimble station, with canopy coverage and canopy and understory species that you expect to find within said trimble station and predict whether you will find an invasive species that is present in a large area of that trimble station. Being able to predict high presense levels of invasive species is important because oen has to be on the look out for invasive species and properly handle them earlier on in their growth so that the native species and enviroment is not threatened.

In terms of our second question of whether there are any features that can help determine high persence levels of invasive species, we have attached the list to the end of the paper of the feature importances of the Random Forest models. Feature importances from Random Forest models are the features that we consistenly important in the various decision trees that were generated.

Future plans for this project might include expanding the data that is being looked at in our model by either expanding to parks that we already have data for or possibly collecting future data to spot future trends in canopy, understory, and invasive species growth.

Resources

Natural Resources Conservation Service. Plants Database. U.S. Department of Agriculture. 2024. Accessed December 5, 2024. https://plants.usda.gov/.

Data provided by Olmsted Parks Conservatory.

Feature Importances for Invasive Species by Park

Cherokee: Lonicera japonica important features

Covered_Canopy_Pct
boxelder understory presence
boxelder canopy presence
pawpaw understory presence
tulip poplar canopy presence

Cherokee: Microstegium vimineum important features

Covered_Canopy_Pct
ash understory presence
hackberry canopy presence
buckeye understory presence
boxelder understory presence

Cherokee: Ampelopsis brevipedunculata important features

Covered_Canopy_Pct
spicebush understory presence
maple canopy presence
cherry canopy presence
maple saplings understory presence

Cherokee: Celastrus orbiculatus important features

Covered_Canopy_Pct
hackberry canopy presence
hickory canopy presence
ash understory presence
hickory understory presence

Cherokee: Vinca minor important features

boxelder understory presence
Covered_Canopy_Pct
ash understory presence
cherry canopy presence
buckeye understory presence

Cherokee: Euonymus fortunei important features

Covered_Canopy_Pct
hackberry canopy presence
pawpaw understory presence
walnut canopy presence
maple canopy presence

Cherokee: Achyranthes important features

Covered_Canopy_Pct
beech canopy presence
buckeye understory presence
oak canopy presence
ash canopy presence

Cherokee: Hederea helix important features

Covered_Canopy_Pct
boxelder understory presence
hickory canopy presence
hackberry canopy presence
maple understory presence

Cherokee: L. maackii important features

Covered_Canopy_Pct
boxelder canopy presence
hackberry canopy presence
spicebush understory presence
cherry canopy presence

Cherokee: Canopy gap and/or downed trees present important features

Covered_Canopy_Pct
pawpaw understory presence
walnut canopy presence
maple understory presence
sugar maple canopy presence

Seneca: Lonicera japonica important features

Covered_Canopy_Pct
ash understory presence
hickory canopy presence
walnut canopy presence
locust canopy presence

Seneca: Ampelopsis brevipedunculata important features

Covered_Canopy_Pct
sycamore canopy presence
maple canopy presence
cherry understory presence
tulip poplar canopy presence

Seneca: Morus alba important features

Covered_Canopy_Pct
hickory canopy presence
ash canopy presence
sycamore canopy presence
walnut canopy presence

Seneca: Celastrus orbiculatus important features

Covered_Canopy_Pct
ash understory presence
cherry canopy presence
sycamore canopy presence
elm canopy presence

Seneca: Euonymus fortunei important features

Covered_Canopy_Pct

hackberry canopy presence
cherry canopy presence
hackberry understory presence
ash understory presence

Seneca: Achyranthes important features

hackberry understory presence
Covered_Canopy_Pct
cherry canopy presence
hackberry canopy presence
walnut canopy presence

Seneca: Hederea helix important features

Covered_Canopy_Pct
hackberry canopy presence
cherry understory presence
elm canopy presence
sycamore canopy presence

Seneca: L. maackii important features

Covered_Canopy_Pct
ash canopy presence
cherry canopy presence
buckeye understory presence
ash understory presence

Seneca: Ailanthus important features

boxelder understory presence
Covered_Canopy_Pct
cherry canopy presence
hackberry canopy presence
ash understory presence

Seneca: Canopy gap and/or downed trees present important features

Covered_Canopy_Pct
boxelder understory presence
ash canopy presence
redbud understory presence
walnut canopy presence

Shawnee Lonicera japonica important features

Covered_Canopy_Pct
elm understory presence
hackberry understory presence
maple canopy presence
sycamore canopy presence

Shawnee Microstegium vimineum important features

Covered_Canopy_Pct
hackberry canopy presence
boxelder understory presence

locust canopy presence
hickory canopy presence

Shawnee Morus alba important features

Covered_Canopy_Pct
hackberry canopy presence
cottonwood canopy presence
maple canopy presence
locust canopy presence

Shawnee Vinca minor important features

Covered_Canopy_Pct
boxelder understory presence
hickory understory presence
cottonwood canopy presence
hackberry canopy presence

Shawnee Euonymus fortunei important features

hackberry canopy presence
walnut canopy presence
locust canopy presence
hickory canopy presence
Covered_Canopy_Pct

Shawnee Achyranthes important features

Covered_Canopy_Pct
hackberry canopy presence
hickory understory presence
elm canopy presence
cottonwood canopy presence

Shawnee L. maackii important features

Covered_Canopy_Pct
maple understory presence
hackberry understory presence
cottonwood canopy presence
hickory canopy presence

Shawnee Privet/ Ligustrum important features

Covered_Canopy_Pct
boxelder understory presence
cottonwood canopy presence
hickory understory presence
locust canopy presence

Shawnee Canopy gap and/or downed trees present important features

Covered_Canopy_Pct
walnut canopy presence
hackberry canopy presence
maple understory presence
hickory canopy presence