Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2117282

Prescriptions for the Afflicted Ecosystem in the Washington State

Summary

Since the first confirmed sighting of Asian giant hornet in December 2019, the ecosystem and residents of the Washington State have been materially adversely affected by this invasive pest. In our thesis, we established prediction and regression model to predict the spread of the hornets as well as the likelihood of a mistaken classification. Our analyses will to some extent help the government to classify the public reports and prioritize its limited resources to follow-up with additional investigations.

For task 1, we started by analyzing the data sets given to us, then we visualized the data by drawing two maps depicting the locations of reported sightings. Since the locations of positive reports are all within a specific region, we calculated the longest distance the queens will travel each year, which is 12km. Then we used this data to predict the spread of Asian giant hornets, and drew the maps showing the spread of the insects in 2020, 2021, and 2022. The accuracy of our model can be as high as 88.05%, and as the area of spread continues to increase, the level of precision of our model will continue to rise.

For task 2, we began with processing the data files and images. We first conducted image recognition based on Convolutional Neural Network (CNN) algorithm to classify the images into true or false categories, and the accuracy of the classification is 0.6. We then analyzed the detection time based on Logistic Growth model, and obtained the probability of discovery at every time point. After that, we used Logistic Regression model to classify and predict the lab status of each public report, and the accuracy of our prediction is 91.18%.

For task 3 and 4, we proposed two criteria to prioritize investigation of the report, first is the predicted result of the Logistic Regression Model and second is whether the location of the report is within the predicted regions in task 1. Then, we discussed how we can update our model from three aspects: images, time, and locations, and specified the frequency of updates is once per year.

For task 5, we referred to the Logistic Growth model, and analyzed the trend in changes of number of positive reports. Then we proposed a scenario in which we can be sure that the pest has been eradicated in the Washington State, and used methods in time series analysis to show the reliability of this approach.

Eventually, we gave our advice on testing and improving the accuracy of public reports from a psychological perspective. Our analyses and suggestions may help the government to better interpret the public reports and prioritize these reports for further investigations.

Keywords: Vespa Mandarinia, Convolutional Neural Network, KNN, Logistic Regression, Logistic Growth Model, Signal Detection Theory

Contents

1 Introduction

1.1 Problem Background

Invasion of alien species has always been known as detrimental to a region's agricultural system and biodiversity. Without natural enemies, alien species can reproduce and spread quickly. In September 2019, a nest of Asian giant hornets was discovered and destroyed on Vancouver Island, British Columbia and in December the first appearance of this species in the United States had been confirmed by the Washington State Department of Agriculture[1].

Asian giant hornet, or Vespa mandarinia, is the largest species of hornet in the world. Its occurrence is especially alarming due to its ability to attack and destroy honey bee hives[2] as well as its possible negative impacts on human health[3]. Under this circumstance, a prompt and feasible plan to detect and eradicate the hornets is in urgent need.

1.2 Restatement of Problem

Due to the severity of the invasion of Asian giant hornets, the State of Washington has requested people to report possible sightings of the hornets. The problem mainly requires us to come up with (1) the ways to interpret the data provided by the public reports, and (2) strategies we should adopt to prioritize different public reports for additional investigation.

First, we should analyze the data sets and create a model to predict the spread of Asian giant hornets, and discuss the level of precision for this prediction.

Second, since most reported sightings are mistaken, we need to carefully analyze the data sets and the image files provided to create a model for predicting the likelihood of a mistaken classification. After establishing the model, we should explain our prioritizing strategies of the public reports based on the classification analyses.

Further, we have to show the extensibility of our model by discussing how we can update it based additional new reports in the future, and we also need to specify the frequency for these updates.

Eventually, we should identify noticeable signals for the dying out of Asian giant hornets in the Washington State.

1.3 Our work

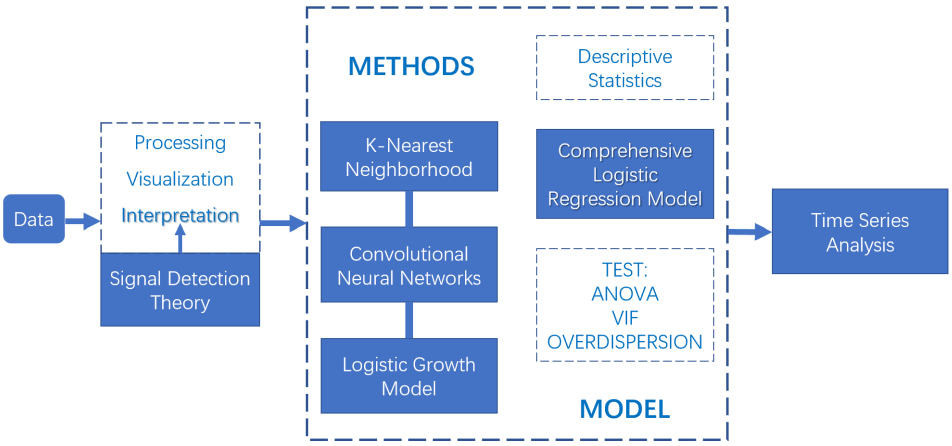


Figure 1: Framework of the Modeling Process

2 Preparation of the Models

2.1 Assumptions and Justifications

1. **We do not consider the photos that are entirely irrelevant or in which we cannot identify the hornets or the hives.**

   Photos that are irrelevant or unclear can barely provide useful information for us to recognize the traits of the hornets, yet the processing of which entails substantial time and efforts. Hence, we exclude these photos in the image recognition procedure.

2. **We do not take environmental factors into consideration when predicting the spread of the insects.**

   Since we can only use the data provided, in which the number of confirmed sightings is extremely small, it's hard for us to find the general pattern of the habitats of Asian giant hornets. In this case, we only consider the time of the migration and their flying distance.

3. **We assume one positive report of sighting in an area means there are around 100 Asian giant hornets.**

   According to the information we have found in the literature, queens will build their nests and raise the first broods of workers in the spring, then the nests will grow through the spring and summer and reach a peak population of around 100 workers[1]. Therefore, it's reasonable to assume there are 100 Asian giant hornets within a nest.

2.2 Notations

The primary notations used in this paper are listed in Table 1.

Table 1: Notations

Symbol	Definition
S	area of the spread of Asian giant hornets
x	time
y	population of the species
λ_y	change rate of population of the species
y_m	maximum possible number of population of the species in an area

3 Data Description

3.1 Data Pre-processing

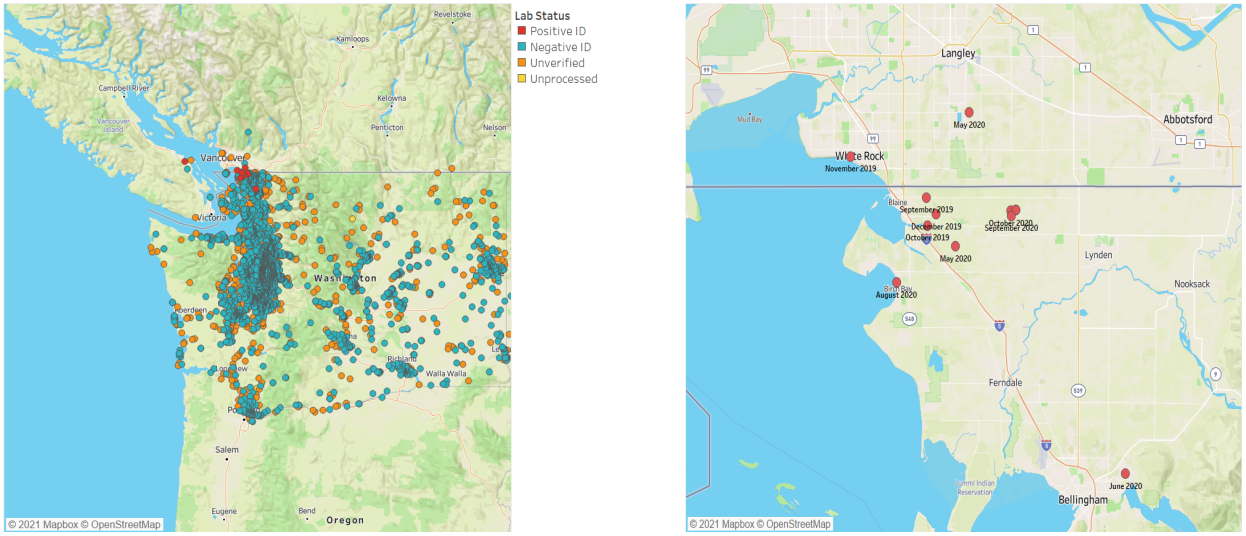
We are provided with three data files, in which are two data sets and one image file. All the data need to undergo pre-processing in order to be used in further analyses.

For two data sets, they present different aspects of the public reports, but they are interconnected by pairing global IDs with image file names. However, we observe that there are certain reports which have only one global ID but several or no corresponding image file names. In order to merge these two data sets into one file, we use global ID as reference. If there are several corresponding image file names, we divide them into separate pieces of data to reduce the loss of information.

For image files, we notice that there are different formats of images, such as video, jpg, and pdf. For videos, we use MATLAB to take screenshots of them every 30 frames; for pdfs, we directly convert them to jpg format. We then classified the photos into four types according to their lab status.

3.2 Data Visualization

Using the given data sets, we draw two maps depicting the locations of reported sightings of Asian giant hornets, figure 2a shows the locations of all public reported sightings with different colors representing different lab status. Figure 2b shows the locations and detection time of all the positive reports.



(a) All Public Reports (b) Positive Reports

Figure 2: Maps Depicting the Locations of Reported Sightings

3.3 Data Interpretation: a Psychological Perspective

Why the proportion of mistaken reports is so large? We refer to the signal detection theory in psychophysics and give our following explanations.

Since only a small number of Asian giant Hornets have been found, the public generally have limited knowledge of the traits of the hornets. If you ask them to make decisions on the basis of evidence which is far less than perfect[4], those decisions are very likely to be mistaken. Also, the sensory evidence on which perceptual decisions are made can be equivocal too[4]. There are primarily two aspects of an observer's decision, which are sensitivity and bias[4]. In this case, sensitivity refers to how well the public is able to identify correct Asian giant hornets and avoid incorrect ones, bias refers to the extent to which they favor one hypothesis over the other irrespective of the information they have been given. Therefore, there are two possible explanations for the high error rate of the public reports, first of which is sensitivity of the public is relatively low, second is that they favor the hypothesis that the insects they discover are Asian giant hornets.

To reduce the perceptual errors, we will give our suggestions later in the memorandum.

4 Model I: Predicting the Spread of Hornets

4.1 Background Information

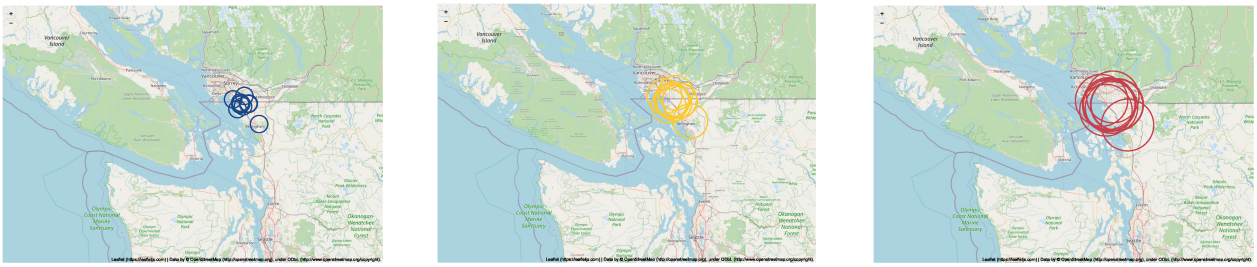
In September 2019, a nest of Asian giant hornets was found and destroyed in Nanaimo, British Columbia, Canada. In December 2019, another hornet was found in Blaine, Washing-

ton, which was the first confirmed sight of this pest in the United States[1]. The two sites are approximately 95 km far from each other, and it has been confirmed that Nanaimo, B.C., Canada and Blaine, WA detection result from two separate maternal lineages[5]. Since the nest in Blaine has been eradicated and no further relevant sightings have been reported, we only consider the spread of hornets starting from Blaine, WA.

According to the information given, all the current seasons' nests will die out in the winter and only overwintering queens can survive[1]. After the winter, the queens will go out and find suitable places for them to build their nests. They will produce around 100 workers until August and then begin to produce males and queens in September. Therefore, the species will have only one massive migration each year.

4.2 Results and Analysis: Solution to Task 1

Based on the current positive results, we calculated the longest distance the queens will travel in a single year, and it turned out this number is approximately 12 km, which is less than 30 km. Thus, we assume the longest distance the queens will travel each year is 12 km, and draw the maps showing the prediction of spread of Asian giant hornets in the following 3 years starting from 2019, which are 2020, 2021, and 2022, as figure 3 below:



(a) First Year : 2020 (b) Second Year : 2021 (c) Third Year : 2022

Figure 3: Maps Showing the Prediction of Spread of Asian Giant Hornets in the Following Three Years Beginning in 2019

As we have noticed in figure 2b, there is one point in the southeast which is unusually away from other points. Due to the limited resources to follow-up with additional investigation, we decided to exclude this point in our prediction, which results in the loss of accuracy of our model.

In our model, the prediction accuracy for the first year is 75.74%, for the second year is 57.78%, for the third year is 88.06%. As the area of spread continues to increase in the future, the level of precision of our model will also increase.

5 Model II: Comprehensive Logistic Regression Model

5.1 Data Processing

5.1.1 Image Recognition Based on CNN Algorithm

Convolutional Neural Network (CNN) is a kind of feed-forward Neural networks, it's default input is image, and can let us construct a specific Network structure as is shown in figure 4 . The nature of the coding input and artificial neurons can be part of the response within range of unit to make our feed-forward function more efficient. At the meantime, it doesn't require a large number of parameters, which can have great performance when processing large number of images[6].

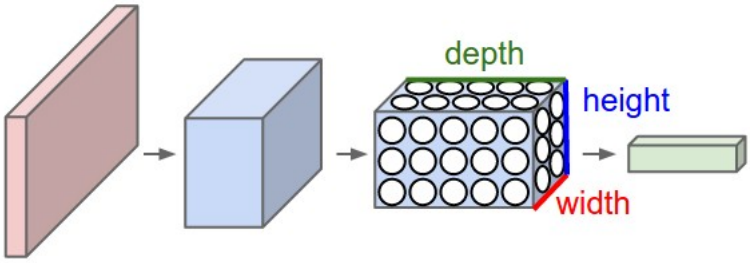


Figure 4: Diagram of Convolutional Neural Network[6]

Therefore, we used CNN to classify the images and distinguish Vespa Mandarinia from other insects. Since the number of positive IDs in the images provided is very small, which will affect the training and testing effect of our model, we then found nearly 100 additional Vespa Mandarinia images from the official website as supplements to the data set. In the implementation process, we use Tensorflow as a tool to build CNN, and lab status in the provided data set is used as the label of the picture. We used 1200 images as a training set and 250 images as a testing set (both contain images of Vespa Mandarinia and other insects).

Eventually, we use our model to classify the images provided by the public, assigning value of 1 to the correct images and value of 0 to the false images. The prediction accuracy of our model is 0.600, which is satisfactory, but can be improved with the input of more images in the future.

5.1.2 Analysis of Detection Time Based on Logistic Growth Model

Model Assumption: In an independent species, the change rate of the number of population is a variable that decreases linearly as the number of population increases.

Mathematical Deduction: Let x denote time, y denote the number of population of the species, λ_y the change rate of population of the species, y_m the maximum possible number

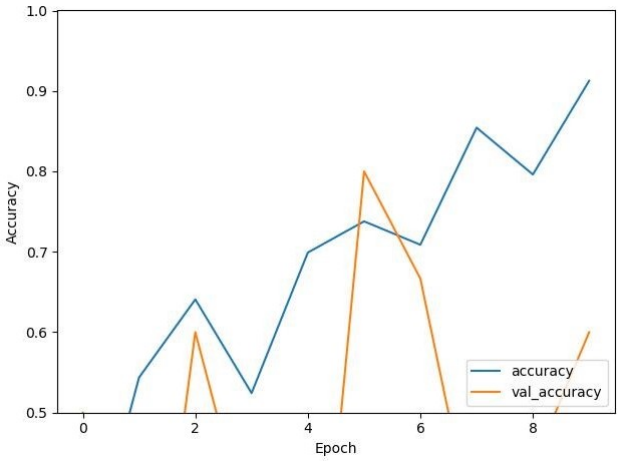


Figure 5: Accuracy of Image Recognition Based on CNN Algorithm

of population of the species in an area. According to the assumption of the Logistic model, given sufficiently short period of time Δx, the change in the number of population of the species can be expressed as $y(x+\Delta x)-y(x)=\lambda_y y(x)\Delta x$, where when $x=0$, $y=y_0$; when $y=0$, $\lambda_y=\lambda$; when $y=y_m$, $\lambda_y=0$.

$$\lambda_y=\begin{cases}\lambda, & y=0\\ 0, & y=y_m;\end{cases} \tag{1}$$

$$\lambda_y=-\frac{\lambda}{y_m}y+\lambda \tag{2}$$

Since $y(x+\Delta x)-y(x)=\lambda_y y(x)\Delta x$, we take $\lim_{\Delta x\to 0}\Delta x$ and obtain the following differential equation:

$$dy=\lambda_y y dx \tag{3}$$

When $y\in(0,y_m)$, combining (2) and (3) we have

$$\frac{dy}{y-\frac{y^2}{y_m}}=\lambda dx \tag{4}$$

$$(\frac{1}{y}+\frac{1}{y_m-y})dy=\lambda dx \tag{5}$$

Take integral of both sides of the above equation, then we have the anti-derivative of the function

$$y=\frac{y_m e^{\lambda x+c}}{1+e^{\lambda x+c}} \tag{6}$$

Substitute $x=0$, $y=y_0$, then we can solve the constant c as follows:

$$c=ln(\frac{y_0}{y_m-y_0}) \tag{7}$$

Replace the constant into the anti-derivative of the function, we obtain

$$y = \frac{y_m}{1 + (\frac{y_m}{y_0} - 1)e^{-\lambda x}} \tag{8}$$

Model Results: According to the background information, queens will hibernate during January to March, and the time period that people are most likely to discover the hornets is from April to December, which is identical to the data we got. The highest possibility of discovery appears on September since at this time the number of population reaches the peak. We also assume that the possibility of discovery is positively correlated with the number of the species, therefore, it's reasonable to assume that the possibility of spotting hornets during their hibernation period equals to the reciprocal of the peak population of the nest, connecting two logistic growth models both reach September as a top. Thus we can know the probability of discovery of every time point.

5.1.3 Notes and Location Analysis

Notes: After analyzing the note of each report, we find that the information provided by the note is generally uncorrelated to its lab status, thus we only use notes as supplements to the images.

Location: We judge whether the location of the sighting is within our prediction for the spread of the hornets, if yes, we assign it a value of 1; if no, we assign it a value of 0.

5.2 Establishment of the Model

The Principle of Logistic Regression

As our response variable Lab.Status is categorical, we use generalized linear regression (GLM) to fit the data. Since it also only has dichotomous (0/1) outcomes, so choosing logistic regression to be our model is reasonable.

In logistic regression, we assume $\pi = \mu_Y$ is the **mean** of Y (also equals variance), the **probability distribution** to be $Y_i \ \ Binomial(1, \mu_Y)$, $X\beta = ln(\frac{\pi}{1-\pi})$ is the **link function** (named **logit**) and $\mu = \frac{e^{X\beta}}{1+e^{X\beta}}$ is the **mean function**. And Y fit the model of the form

$$ln(\frac{\pi}{1-\pi}) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \tag{9}$$

After we do all these assumptions, we go through an **iterative maximum likelihood estimation procedure** to get the minimum deviance and derive all parameters. The **minimum**

deviance is denoted as

$$d_i = s_i \sqrt{-2[y_i ln\mu_i + (1 - y_i) ln(1 - \mu_i)]} \tag{10}$$

Predictors:

1. environmental suitability (referred to [7]) (continuous)
2. time normalization (continuous)
3. location (categorical)
4. image (categorical)

Dependent variable:

1. Lab.Status (categorical)

5.3 Results and Analysis: Solution to Task 2 & 3

	Estimate	Standard Error	z value	Pr(>|z|)	Signif.
(Intercept)	-4.85908	2.89372	-1.68	0.093116	.
suitability	0.00539	0.00514	1.05	0.294294	
time_normalization	-0.47049	2.13894	-0.22	0.825900	
Location	4.27831	1.50618	2.84	0.004504	**
Image	1.45026	1.39365	1.04	0.298051	

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < '' < 1*

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 47.13401 on 33 degrees of freedom

Residual deviance: 17.37167 on 29 degrees of freedom

Figure 6: Result of the Original Model

From the result of figure 6, if we set the criterion $\alpha = 0.3$, we can get the suitability, location, and image variables are significant, with value of 0.294, 0.004, 0.298 respectively. Then, under this criterion, we delete the time factor and construct the reduced model, the result of which is shown in figure 7. Our reduced model is: glm (formula = Lab.Status ~ suitability + Location + Image). Figure 7 shows that all three variables are significant under the criterion.

	Estimate	Standard Error	z value	Pr(>|z|)	Signif.
(Intercept)	-5.285	2.315	-2.283	0.0224	*
suitability	0.006	0.005	1.107	0.2684	
Location	4.162	1.374	3.030	0.0024	**
Image	1.546	1.333	1.160	0.2460	

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < ' ' < 1*

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 47.13 on 33 degrees of freedom

Residual deviance: 17.42 on 30 degrees of freedom

Figure 7: Result of the Reduced Model

```
               Accuracy : 0.9118
                 95% CI : (0.7632, 0.9814)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 3.83e-07

                  Kappa : 0.8235
```

Figure 8: Accuracy of the Logistic Regression Model

Table 2: Results of Coefficient Analysis

Predictor	Coefficient
(Intercept)	0.005067892
suitability	1.005676
location	64.219794741
image	4.692913804

From the result shown in table 2, we can get that the initial probability of Positive ID is 0.005, then by adding one degree of suitability, the result times 1.006 (which means increasing suitability can increase the probability). Location change from 0 to 1, the result is multiplied by 64.220 times and the change from 0 to 1 of Image leads to 4.693 times change.

In short, three coefficients are greater than one means the change is positive while the coefficients' number increase.

Solution to task 2: When a report is received, we can use our model to classify whether it is positive or negative. The result of confussion matrix shows that the accuracy of the logistic regression model is 91.18% as is shown in 8, which is satisfactory.

Solution to task 3: As is shown in figure 7, location is the most significant predictor. So when a report is received, we can first use our model to predict its lab status, if our model shows it's a positive report, we can then judge whether the location of the reported sighting is within the predicted area. Due to the limited resources of government agencies, we should give top priority to the report whose predicted lab status is positive and whose location is within the predicted spreading area.

5.4 Updates of the Model: Solution to Task 4

As more public reports are submitted over time, we can update the model from the following three aspects:

Images: As more public reports are submitted, the images provided by the public will help us to better train the image recognition algorithm, and hopefully will raise the accuracy of prediction.

Time: Due to the limited number of positive reports we have now, we can only assume the probability of seeing the hornets at every time point follows normal distribution. With more input of the data, we can obtain a more accurate fitting curve.

Location: Over time, the area of the spread of the hornets will increase, thus we will need to rejudge whether the location of the sighting is within our prediction.

The number of positive sightings have a seasonal trend with lag equals to 12 months, thus we only need to update our model once a year.

5.5 Time Series Analysis of Seasonal Model: Solution to Task 5

According to the logistic growth model, the number of species within a region will first increase and reach the peak. Although the number of the species exhibits seasonality, we can take the seasonal difference of the data to erase the seasonality, and examine the general trend.

Since the government of the Washington State is finding various ways to eradicate the pests, if those methods are effective, we will first see a downward trend in the number of hornets found, and eventually no positive reports for several years, as the trend exhibiting in the following diagram:

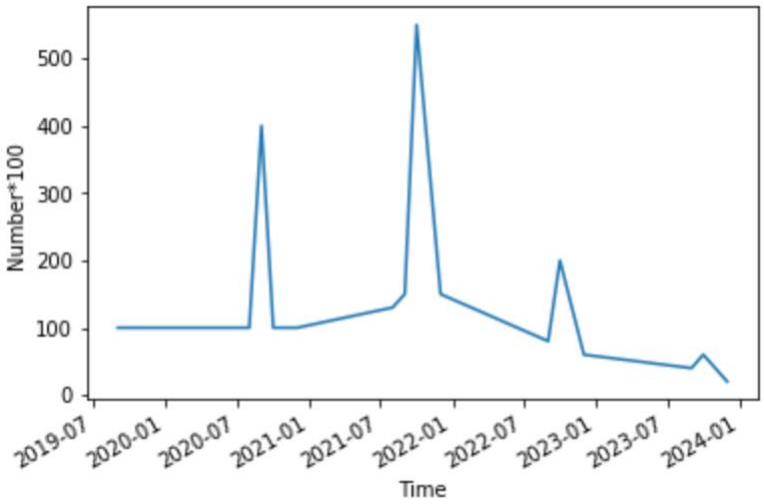


Figure 9: The number of Asian Giant Hornets Found in the Future

6 Statistical Analysis of Model II

Test VIF

Here we test the multicollinearity among the variables. Variance inflation factor (VIF) denote as

$$(VIF)_i = \frac{1}{1 - R_i^2} \tag{11}$$

where R_i^2 is the multiple coefficient of determination for regression model which delete the term i. In statistical area, if VIF is less than 2, we regard it as not having multicollinearity.

From the output of VIF, we can see that each VIF do not have significant difference compared to 1, meaning that no variable which can be linearly predicted from the others with a

Table 3: Result of VIF Test

Predictor	vif(fit red)
suitability	1.052674
location	1.069763
image	1.069931

substantial degree of accuracy.

ANOVA to test differences between full and reduced model

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
30	17.42059			
29	17.37167	1	0.04892092	0.8249517

Figure 10: The Result of ANOVA Test

We use the Chi-Square Comparison, do the ANOVA test to compare two models.

As our null hypothesis H_0 is these two models don't have significant difference. And P-value is 0.82 which is greater than 0.05, meaning that we can't reject the fact that these two models have same impact on predicting the result, we can set the reduced model to be our final model.

Confidence Interval of Coefficients

Table 4: Confidence Interval of Coefficients

Predictor	2.5%	97.5%
(intercept)	1.387779e-05	0.1968238
suitability	9.962931e-01	1.0181026
location	6.333491e+00	2068.4642070
image	3.738583e-01	117.2616176

From the result shown in table 4, for example, the 95% confidence interval for location is (6.333, 2068.464). With 95 percent confidence level, we estimate the multiples of the percentage growth from 6.333 times to 2068.464 times if the condition change from 0 to 1.

Test Overdispersion

Overdispersion often happens in the omission of an important predictor variable, clustering inherent in repeated measures data or known as state dependence. In logistic regression, overdispersion is suggested if ϕ is much larger than 1.

$$\phi = \frac{Residual\ deviance}{Residual\ df} \tag{12}$$

H_0 is $\phi = 1$, as the result is 0.4231 > 0.05, so we can not reject the null hypothesis, this model don't have overdispersion.

7 Model Evaluation

7.1 Strengths

- Our analysis is innovative. We referred to the signal detection theory in psychophysics to explain the high error rate of public reports, and suggested a relatively easy method to estimate the reliability of the reports after they were submitted.
- Using CNN to do image classification, whose artificial neurons can respond to surrounding units in a partial coverage area and perform well for large image processing.
- We have conducted many statistical tests during our analysis process, such as VIF, ANOVA, overdispersion, which let our result more convincing.

7.2 Weaknesses

- Since we ignore the environmental factors in predicting the spread of Asian giant hornets, the prediction results might be too general.
- We exclude some ambiguous photos in the image recognition step, and these photos may leave experts to analyze, which will result in extra manual work.
- Due to factors such as data set size and hardware limitations, the prediction effect of CNN is limited.

8 Conclusion

In this paper, we build a model to predict the spread of Vespa Mandarinia, find the method to calculate the potential rate of positive sighting, and show reasons of the eradication for the pest in Washington State.

We use circles with a radius of 12 km, which is the longest distance the queens can travel in one year, to describe the spread. The accuracy for the first year is 75.74%, for the second year is 88.06% and for the third year is 88.06%.

After using CNN, KNN, and Logistic Growth Model for the data processing, we form a Comprehensive Logistic Regression Model to predict the probability of the true sighting. This model's accuracy is 91.18% and under the criterion p-value = 0.3, we can get three significant variables, location, environment suitability, and image message, to be three factors impacting the probability. The higher the environmental suitability is, the bigger the percentage is. Also, the related distance to the existing positive ID or the judgment result of CNN can multiply the result to the corresponding times.

According to the result of our logistic growth model, we find that the number of places where the pest was discovered will finally approach zero even though it may increase in next year, which means that the Ves Mandarinia has lost their ability to multiply in large numbers and has been eradicated in Washington State.

Pest control is a very important part of the ecological environment and production activities. More and more people pay much attention to this field and participate in the research. Although our models still need to be improved in some aspects, we believe that they can offer useful information for fighting against Vespa Mandarinia. If possible, we hope that we can dig deeper into the control of the pest.

Memorandum

To: the Washington State Department of Agriculture
From: Team 2117282
Date: February 9th, 2021
Subject: Solutions from Team 2117282

This memo is presented to provide you with an overview of our models on predicting the spread of as well as recognizing the Asian giant hornets.This invasive species has had overwhelmingly negative impacts on local people and ecosystems. Our team has collected wide range of background information, and employed models that we think are the most suitable for the tasks. We even referred to the **Signal Detection Theory** in Psychophysics in interpretation of the high error rate of the public reports, and proposed two strategies that may be useful in increasing the accuracy of people's judgements.

First of all, based on the time and locations where Vespa Mandarinia were found offered by the data sets, we build a model using circle to simulate the spread of this kind of insects. We then calculated the longest distance that queens will travel in a single year, which is 12 km, since we have found that the migration of Vespa Mandarinia tends to follow the queens' life span. Due to the limited resources of government agencies, we exclude a location unusually away from other locations. The prediction accuracy for the first year is 75.74%, for the second year is 88.06% and for the third year is 88.06%. The larger the spread area is, the more accurate our model is.According to the signal detection theory, there are primarily two ways to increase the accuracy of public reports or judge the accuracy of the reporters, which are

1. **The government should provide the public with more information of the Asian giant hornets, including but not limited to their photos, possible locations, and habits.**

2. **The government can request the reporters to perform a short yes-or-no task after their reports. The task is about judging whether the insects in three photos given are Asian giant hornets or not. If the accuracy of the task is low, then it's very likely that their reports are not reliable.**

In Model II, we focused on the probability of the mistaken classification by the local people and proposed the measures you can take to do the preferential investigations on people's

knowledge of the Asian giant hornets. In this case, we do a thorough data cleaning including image identification and test processing. We use Convolutional Neural Network(CNN), which has outstanding performance in image classification. With additional nearly 100 images of Vespa mandarinia, we trained a CNN model having the test accuracy of 0.600 using 1100 images as training set and 250 images as test set together with their labels, which is obviously much higher than the ratio of Positive IDs in the data set files offered.

After that, we divide the test into parts. For the latitude and longitude numbers, we compare them with the given positive data, taking the threshold value to be 8 miles and change it into categorical and we use a comprehensive logistic growth model to fit the Vespa mandarinia's life history. We also find a document describing the environment suitability, so we use the environmental data, applying the KNN model to find the most appropriate suitability number for each point. Up to now, we explain all data we need to deal with your concern.

Basing on the processed data, we can build a model to find the influential variables which can do a difference in the probability of the positive finding. As our response variable is categorical, we choose to use logistic regression in general linear regression(GLM) algorithm. In order to reduce the unbiased impact, we randomly pick some data from Negative ID to form the training set. The result shows that the environment suitability, location and image message are significant to the model, while the discovery of time is more random. Also to be more scientific and statistical, we test VIF and find no multicollinearity as well as no overdispersion.

From this model, we take the exponential and see that the coefficient is big and the 95% interval is wide, which means the small area around the existing positive sightings has a much higher probability of reporting real Vespa mandarinia. As our model has accuracy up to 91.18%, by using this model, we can predict data by fitting this model to classify the lab status with small chances of incorrect, and give top priority to the report whose predicted lab status is positive and whose location is within the predicted spreading area. Also as the Vespa mandarinia has its life circle which is 1 year, so we can update all three factors once a year.

To find the evidence that the pest has been eradicated in Washington State, a trend in the number of sites where Vespa Mandarinia was found is fitted out under the seasonality according to the logistic growth model and assuming that the government's approach to pest control is effective, and the trend tells us that the amount of sites may experience a short period of increase, but finally it will approach zero in two or three years.

References

[1] Skvarla, Michael. "Asian Giant Hornets". *Penn State Extension*, 2020, https://extension.psu.edu/asian-giant-hornets. Accessed 5 Feb 2021.

[2] "Hornets | Washington State Department Of Agriculture". *Agr.Wa.Gov*, 2021, https://agr.wa.gov/departments/insects-pests-and-weeds/insects/hornets. Accessed 5 Feb 2021.

[3] Alaniz, Alberto J et al. "Giants Are Coming? Predicting The Potential Spread And Impacts Of The Giant Asian Hornet (Vespa Mandarinia , Hymenoptera:Vespidae) In The USA". *Pest Management Science*, vol 77, no. 1, 2020, pp. 104-112. *Wiley*, doi:10.1002/ps.6063.

[4] McNicol, Don. *Primer Of Signal Detection*. Lawrence Erlbaum Associates, 2005, pp. 1-13.

[5] Wilson, Telissa M et al. "First Reports Of Vespa Mandarinia (Hymenoptera: Vespidae) In North America Represent Two Separate Maternal Lineages In Washington State, United States, And British Columbia, Canada". *Annals Of The Entomological Society Of America*, 2020.*Oxford University Press (OUP)*, doi:10.1093/aesa/saaa024.

[6] Liu, He. *Detailed Introduction Of Convolutional Neural Network*. 2016, https://blog.csdn.net/qq 25762497/article/details/51052861. Accessed 8 Feb 2021.

[7] Zhu, Gengping et al. "Assessing The Ecological Niche And Invasion Potential Of The Asian Giant Hornet". *Proceedings Of The National Academy Of Sciences*, vol 117, no. 40, 2020, pp. 24646-24648. *Proceedings Of The National Academy Of Sciences*, doi:10.1073/pnas.2011441117.

[8] Kabacoff, Robert. *R In Action*. Manning Publications, 2015.

Appendix A: Code for Convolutional Neural Network

```
import tensorflow as tf
import pandas as pd
import numpy as np
from PIL import Image

from tensorflow.keras import layers, models
import matplotlib.pyplot as plt

if __name__ == '__main__':
    train_data_frame = pd.read_excel('train.xlsx', sheet_name='Sheet1')
    test_data_frame = pd.read_excel('test.xlsx', sheet_name='Sheet1')
    label_frame = pd.read_excel('label.xlsx', sheet_name='Sheet1')
    arr = np.array(train_data_frame)
    test_arr = np.array(test_data_frame)
    tmp = np.array(label_frame)
    labels = []
    for t in tmp:
        labels.append(t[0])
    labels = np.array(labels)
    train_img = []
    test_img = []
    for a in arr:
        image = Image.open(a[0])
        image = image.resize((32, 32))
        image_arr = np.array(image)
        train_img.append(image_arr)
    train_img = np.array(train_img)
    train_arr = np.ndarray((train_img.shape[0], *train_img[0].shape))
    for t in test_arr:
        image = Image.open(t[0])
        image = image.resize((32, 32))
        image_arr = np.array(image)
        test_img.append(image_arr)
    test_img = np.array(test_img)
    test_arr = np.ndarray((test_img.shape[0], *test_img[0].shape))
    print("test[0]␣shape" + str(test_img[0].shape))
    print("train[0]␣shape" + str(train_img[0].shape))
    for i in range(train_img.shape[0]):
        for j in range(train_img[0].shape[0]):
            for k in range(train_img[0].shape[1]):
                for l in range(train_img[0].shape[2]):
                    train_arr[i, j, k, l] = train_img[i][j, k, l]
    for i in range(test_img.shape[0]):
        for j in range(test_img[0].shape[0]):
            for k in range(test_img[0].shape[1]):
                for l in range(test_img[0].shape[2]):
                    test_arr[i, j, k, l] = test_img[i][j, k, l]
    print("train[0]␣type" + str(type(train_img[0])))
```

```
    print("test␣shape" + str(test_img.shape))
    train_label = labels[:103]
    test_label = labels[103:]
    model = models.Sequential()
    model.add(layers.Conv2D(32, (3, 3), activation='relu', input_shape=(32, 32, 3)))
    model.add(layers.MaxPooling2D((2, 2)))
    model.add(layers.Conv2D(64, (3, 3), activation='relu'))
    model.add(layers.MaxPooling2D((2, 2)))
    model.add(layers.Conv2D(64, (3, 3), activation='relu'))
    model.add(layers.Flatten())
    model.add(layers.Dense(64, activation='relu'))
    model.add(layers.Dense(10))
    model.compile(optimizer='adam',
                  loss=tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True),
                  metrics=['accuracy'])

    history = model.fit(train_arr, train_label, epochs=10,
                        validation_data=(test_arr, test_label))
    predictios = model.predict_classes(test_img)
    print("predictions:")
    print(predictios)
    plt.plot(history.history['acc'], label='accuracy')
    plt.plot(history.history['val_acc'], label='val_accuracy')
    plt.xlabel('Epoch')
    plt.ylabel('Accuracy')
    plt.ylim([0.5, 1])
    plt.legend(loc='lower␣right')
    plt.show()

    test_loss, test_acc = model.evaluate(test_arr, test_label, verbose=2)
    print(test_loss)
    print(test_acc)
# Refer to https://www.tensorflow.org/tutorials/images/cnn
```

Appendix B: Code for Logistic Regression Model

```
data = read.csv("data_merged7.csv")
data_pn= data[data$Lab.Status == "Positive␣ID" | data$Lab.Status == "Negative␣ID",
              c("Lab.Status","suitability","time_normalization","Location","Image")]
data_n = data[data$Lab.Status == "Negative␣ID",
              c("Lab.Status","suitability","time_normalization","Location","Image")]
data_p = data[data$Lab.Status == "Positive␣ID",
              c("Lab.Status","suitability","time_normalization","Location","Image")]
data_pn <- data_pn %>%
  mutate(Lab.Status = if_else(Lab.Status == "Positive␣ID", 1, 0))

summary(data_pn$suitability)
summary(data$time_normalization)

```

```
data_pn= data[data$Lab.Status == "Positive␣ID" | data$Lab.Status == "Negative␣ID",
                  c("Lab.Status"," suitability ","time_normalization","Location","Image")]
data_pn <- data_pn %>%
  mutate(Location = as.factor(Location),
         Lab.Status = as.factor(Lab.Status)
         ) %>%
  mutate_if(is.character, as.factor) %>%
  dplyr::select(Lab.Status,Location, everything())

ggplot(data, aes(x=Lab.Status, fill =Lab.Status)) +
  geom_bar() +
  xlab("Lab␣Status") +
  ylab("Count")

ggplot(data, aes(x=suitability )) +
  geom_histogram() +
  scale_ fill _brewer(palette = "Paired") +
  xlab(" Suitability ") +
  ylab("Count")

ggplot(data, aes(x=time_normalization)) +
  geom_histogram() +
  stat_bin(bins = 10) +
  xlab("Time") +
  ylab("Count")

p0 <- ggplot(data_pn, aes(Lab.Status, fill = Location))+
geom_bar(position = "fill")+ylab("Percentage")
p0

set.seed(399973)
data_p$Lab.Status = 1
m = sample(1:3237,17)
data_n_sample = data.frame()
for (i in (0:17)){ data_n_sample = rbind(data_n_sample,data_n[m[i],])}
data_n_sample$Lab.Status = 0
ourdata = rbind(data_n_sample,data_p)

fit_full = glm(Lab.Status~., family = binomial(link = "logit"), data = ourdata,
                  control=list(maxit=1000))
as_flextable( fit_full )
fit_red = glm(Lab.Status~suitability+Location+Image, family = binomial(link = "logit"),
                  data = ourdata, control=list(maxit=1000))
as_flextable( fit_red)
vif_red = as.data.frame(vif( fit_red))
vif_red
anova = anova(fit_red, fit_full, test = "Chisq")
flextable(data = anova)
coef = as.data.frame(exp(coef(fit_red)))
names(coef) = "Coefficient"
coef
```

```
as.data.frame(exp(confint(fit_red)))
 fit.od = glm(Lab.Status~., family = quasibinomial(link = "logit"), data = ourdata,
                control=list(maxit=100))
pchisq(summary(fit.od)$dispersion * fit_red$df.residual, fit_red$df.residual, lower = F)
true = as.factor(ourdata$Lab.Status)
pred = as.numeric(predict(fit_red,type = "response", ourdata)>0.5)
pred = as.factor(pred)
confusion_matrix = confusionMatrix(data = pred, reference = true)
confusion_matrix
```