Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2108410

The control of biological invasion based on deep learning

Summary

The Asian giant hornets were discovered in Washington state in September 2019. Helplines and a website were created to receive reports from the public about Asian giant hornets. Our team was expected to creat mathematical models to interpret the public reports and prioritize them for further investigation.

For **Question 1**, our team established the **Spreading model of pests based on the Maximum entropy**. We find the regression of distribution to environmental covariates, and then predict the spread of the species using the maxium entropy principle. Firstly, 14 samples in the data set were selected, whose Lab status is "positive ID". Next, we generate 10,000 background points using the Ramdonkfold algorithm. After selecting covariates, the objective function is established and the constraint relation is determined according to **maximum entropy principle**. Then we solved the parameters in the regression equation and obtained the probability distribution of the Asian giant hornets. Finally, **AUC test and Jackknife test** were used to test the accuracy of the model and the contribution of environmental covariates.

For **Question 2 and 3**, our team established **Scoring model of the reports.** First, we score each report from three dimensions: Notes, Images, Latitude and Longitude. Then the total score of each report is weighted and calculated. For Question 2, the likelihood of a mistaken classification is calculated according to the score. For Question 3, we rank the scores from highest to lowest, and prioritize the investigation based on the score. In the first place, natural language processing (**NLP**) is used to score the "Notes". Then the image recognition algorithm based on convolutional neural network (**CNN**) was used to identify the sightings and obtain the probability of sighting as real Asian giant hornets. Next, according to the probability distribution of pests in Washington State obtained in Question 1, the probability of Asian giant hornets appearing in the observed positions in the report was calculated.Finally, the comparative weight method in **AHP** is used to give weight to the three indexes. The final score is obtained by multiplying the scores of the three indicators of each report by the weights respectively.

For **Question 4**, With the addition of new reports, first of all, we need to re-select the roots of high frequency. And we need to modify the known distribution of pests in the Spreading model of pests based on the Maximum entropy. And then **solve the new probability distribution** based on the objective function and the constraint relationship. Next, we recalculate the Location (Latitude and Longitude) score in the report. Then we calculate the final score according to the weight. As for the update frequency, our team concluded that the model should be updated every **100** Reports. We found that there was at least an 60-percent probability that the report was true in the dataset of 40 samples. We infer that there will be one sample out of 100 with a probability of more than 60% that it is a true report. So it would be sensible to update the model every 100 reports.

For **Question 5** To judge and test pest eradication in Washington State, we created **the Model of population change of Asian giant hornets**. For the queen, we created an exponential growth model. For workers, we created a linear variation model. Then the model was modified by adding interference factors to the model to take into account the human interference to the population of pests. By drawing the Asian giant hornets population under different measures of the change curve, we found that **destroying their nests** interfered most with the pests. According to our model, when the number of destroyed nests is **2** per month, thepopulation declines to zero after **25 months**.

Keywords:species invasion, biological control, Maximum entropy, CNN, deep learning.

Contents

1 Introduction

1.1 Problem Background

Asian giant hornets, also called sparrow wasps , are a potentially invasive wasp from eastern Asia . The scientific name of Asian giant hornet is Vespa mandarinia Smith, 1852. The Asian giant hornet is a member of the family Vespidae, including yellowjackets, hornets, and paper wasps[14]. And Bees, wasps, and related insects belong to the same order as the Asian giant hornets. Asian giant hornets are in striking colors, with yellow heads, a black thorax, and yellow and black or brown striped abdomens. The Asian giant hornet queens can be more than two inches long, with a wingspan of about three inches. Asian giant hornet workers can grow up to 1.5 inches long[14]. Asian giant hornets are accustomed to building their nests underground. Dead, hollow trunks or roots of trees which are less than 3 to 6 feet above the ground are their typical choice. Native to temperate and tropicaleastern Asia, Asian giant hornets are most commonly encountered in rural areas of Japan.

There are a number of other species that are similar to Asian giant hornets, such as European hornets, cicada killers[14] and so on. European hornets, also known as Vespa crabro, are similar to Asian giant hornets in size, shape, and color. Eastern cicada killers are native wasps that are similar in size to Asian giant hornets.

The image below shows how the Asian giant hornets differ from other species that look alike.

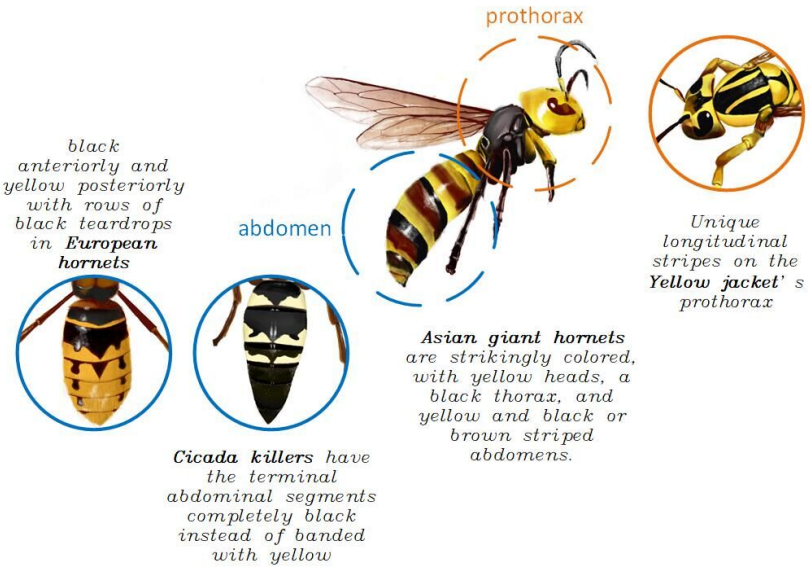


Figure 1: Comparison of Asian giant hornets with other look-alike species

1.2 Problem Review

In September 2019, a colony of Asian giant hornet was discovered and destoryed on Vancouver Island. The news of the event spread rapidly throughout the area. Since that time, several confirmed sightings of the pest have occurred in neighboring Washington State, as well as a multitude of mistaken sightings.

Due to the potential severe impact on local honeybee populations, the presence of Vespa mandarinia can cause a good deal of anxiety. The State of Washington has created helplines

and a website for people to report sightings of these hornets. Based on these reports from the public, the state must decide how to prioritize its limited resources to follow-up with additional investigation. While some reports have been determined to be Vespa mandarinia, many other sightings have turned out to be other types of insects.

The primary questions for this problem are " How can we interpret the data provided by the public reports?" and "What strategies can we use to prioritize these public reports for additional investigation given the limited resources of government agencies?"

Our team are expected to explore and address the following aspects:

- Address and discuss whether or not the spread of this pest over time can be predicted,and with what level of precision.
- Most reported sightings mistake other hornets for the Vespa mandarinia. Use only the data set file[1] provided, and (possibly) the image files provided, to create, analyze, and discuss a model that predicts the likelihood of a mistaken classification.
- Use your model to discuss how your classification analyses leads to prioritizing investigation of the reports most likely to be positive sightings.
- Address how you could update your model given additional new reports over time, and how often the updates should occur.
- Using your model, what would constitute evidence that the pest has been eradicated in Washington State?

2 Problem analysis

Question1

In this section, our team is expected to solve whether the model of pest spread over time can be predicted and the accuracy of the prediction. We now know the latitude and longitude of the "positive ID" in the data set. The goal is to use the data to predict the probability of pest distribution in other parts of Washington state. This is a typical predictive distribution problem in ecology. A class of models in biology predicts the distribution of species by linking environmental variables to their presence or absence.

Maximum entropy (MaxEnt)[13] , genetic algorithm for rule-set production (GARP), and ecological niche factor analysis (ENFA) are tools that predict species suitability using presence-only data. Compared to the other two models The MaxEnt model has better prediction results. This is because the MaxEnt model fits the species probability distribution by the maximum entropy principle under the assumption that there is no environmental restriction.

Question2

In this section, our team is expected to create a model that predicts the likelihood of a mistaken, using only the data set file provided, and (possibly) the image files provided. After data cleaning of the data in the given data set, A few key indicators can be extracted. Based on these indicators, the report is scored and evaluated, and its likelihood of a mistaken classification is judged according to the score.

Question3

In this section, our team is going to discuss how your classification analyses leads to prioritizing investigation of the reports most likely to be positive sightings. According to the established scoring model, the scores are sorted from the highest to the lowest order, so that the survey priority can be known.

Question4

In this section, what our team wants to solve is the problem of model update. The original model is based on existing data sets. If there are newly added Reports, we need to modify the parameters in the model and then update the model.

Question5

According to the reproductive characteristics of Asian giant hornets, we could establish the model of population quantity changing over time. When the population drops to zero and the spatial distribution probability drops to zero Washington's pests were considered extinct.

3 Preparation of the Models

3.1 Assumptions

1. The range of spreading of population is dependent only on the range of nesting sites.

2. Assume that the Asian giant hornets has no natural predators in Washington State.

3. Rich in natural resources that is needed for the pests to survive.

4. The population of Asian giant hornets does not overlap over generations

5. Assume that the reports in the data set are all true.

3.2 Notations

The primary notations used in this paper are listed in Table

Table 1: Basic Information about Three Main Continents (scratched from Wikipedia)

Symbol	Definition
$\mathbf{z}$	the environmental covariates
L	the area we are going to explore the possibility of the absence of the pests
$f(z)$	the probability density of covariates across L
$f_1(z)$	the probability density of covariates across locations within L, where the species is present, and similarly
$Pr(y=1\|z)$	the probability of presence of the species
$Pr(y=1)$	the prevalence of the species
$h(\mathbf{z})$	the vector of featuresand
β	the vector of coefficients

x_j^l	the output of the J channel of convolutional layer I
$\otimes$	Convolution Operation
W_j	a subset of the input feature map
W_{ij}^l	the convolution kernel matrix
b_j^l	the bias of the convoluted feature map
k^l	the weight coefficient of the fully connected network
$\alpha^{'}$	the learning rate
C_1	the score of Notes
C_2	the score of image
C_3	the score of location(latitude and longitude)
ω_i $(1 \leqslant i \leqslant 3)$	the weight of C_i
a_{ij}	the ratio of Ci to Cj s impact on the final score
d	the likelihood of a mistaken classification

4 The Models

4.1 Spreading model of pests based on the Maximum entropy

In this section, we are going to predict the spread of Asian giant hornets, that is, to solve the problem of probability distribution of species. In biology, the distribution problem is essentially a system composition problem. Since maximum entropy principle is an important tool to study system composition, we adopt maximum entropy principle to establish mathematical model. A brief principle of maximum entropy is as follows.

4.1.1 Introduction of the principle of maximum entropy

The principle of maximum entropy[2] states that when learning a probability model, the model with the highest entropy is the best probability distribution model. In information theory, entropy is used to describe the uncertainty of a probability distribution. The principle of maximum entropy embodies the idea that keeping options open ensures maximum uncertainty, it is therefore the safest option.

Let's say the probability distribution of the discrete random variable X is $P(X)$, is

$$H(P) = -\sum_x P(x) \log P(x) \tag{1}$$

Entropy satisfies the following inequality:

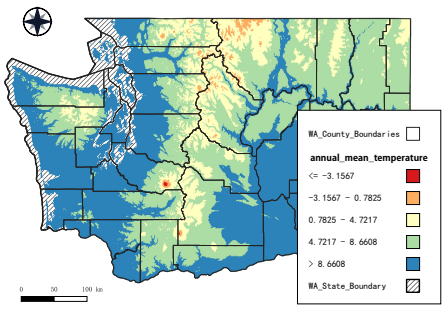
$$0 \leqslant H(P) \leqslant \log |X| \tag{2}$$

In the distribution of the discrete random variable, the equal sign on the right hand side of the above equation holds if and only if the distribution of X is uniform. Which is to say that entropy is maximum when X is randomly distributed uniformly.

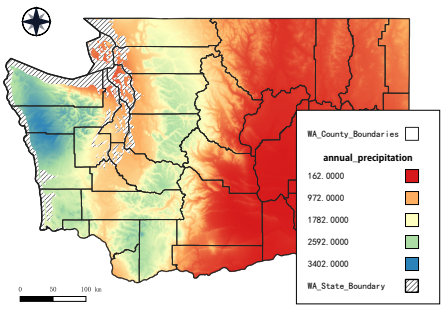
4.1.2 Spreading model of pests based on MaxEnt machine learning algorithm

The "positive IDs" of 14 Asian giant hornets were given in the dataset. These are the established distribution of Asian giant hornets in Washington state. According to the principle of maximum entropy, there is no reason to expect the pests to prefer any environment over other species[3].

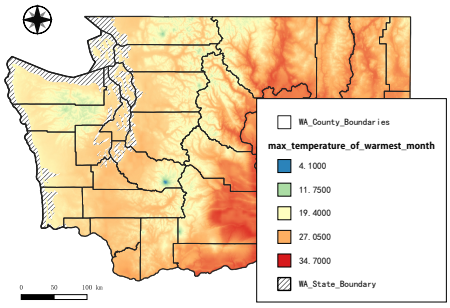
The data of 17 ecological factors[4, 5, 6, 7, 8] and 14 distribution points of Asian giant hornets were imported into MaxEnt model software for processing. According to the results of multiple iterations of the MaxEnt model, 10 factors whose total contribution value is more than 90% are selected as the main ecological factors. Order the environmental covariates(some of them are shown in the following figures) to be $\mathbf{z} = (z_1, z_2, ..., z_{10})$ z_1 donotes Annual Mean Temperature, z_2 denotes Mean Diurnal Range (Mean of monthly (max temp - min temp)), z_3 denotes Max Temperature of Warmest Month, z_4 denotes Min Temperature of Coldest Month, z_5 denotes Temperature Annual Range, z_6 denotes Mean Temperature of Wettest Quarter, z_7 denotes Annual Precipitation, $z_8$denotes Precipitation of Wettest Quarter, z_9 denotes Precipitation of Driest Quarter, z_{10} denotes Forest Cover.



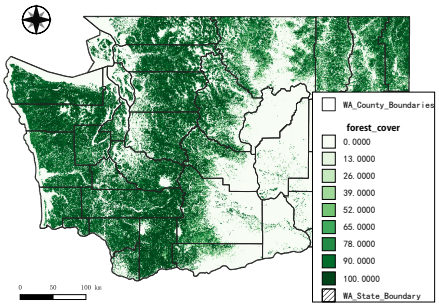
(a) Annual Mean Temperature



(b) Annual Precipitation



(c) Max Temperature of Warmest Month



(d) Forest Cover

Figure 2: Environmental covariate

Let y = 1 denotes presence, y = 0 denotes absence. L denotes the area we are going to explore the possibility of the absence of the pests. Define $f(z)$ to be the probability density of covariates across L, $f_1(z)$ to be the probability density of covariates across locations within L, where the species is present, and similarly.

What we want to estimate is Pr(y=1|z), denoting the probability of presence of the species, depending on the environment. Bayes rule gives:

$$\Pr(\mathrm{y} = 1 \mid \mathbf{z}) = f_1(\mathbf{z}) \Pr(\mathrm{y} = 1)/f(\mathbf{z}) \tag{3}$$

Pr(y = 1) denote the prevalence of the species. Now what we want to know is the ratio: $f_1(z)/f(z)$.

MaxEnt uses the covariate data from the occurrence records and the background sample to estimate the ratio $f_1(\mathbf{z})/f(\mathbf{z})$. By convention, we first generated 10,000 Background Samples in MaxEnt. In MaxEnt, the relative entropy of $f_1(z)$ with respect to $f(z)$ denote thedistance from $f(z)$.

The purpose of the constraint is to make the model reflect information about the environmental covariates with the absence of the pests. One covariate z_1 is Annual Mean Temperature, then constraints ensure that the mean Annual Mean Temperature for the estimate of $f_1(z)$ is close to its mean across the locations with observed presences. In fact, to enable MaxEnt to make predictions for more complex models, it actually fits the model on features that are transfor- mations of the covariates.So the constraint changed from the constraint on the mean of the environmental covariate to the constraint on the eigenmean of the environmental covariate. We will call the vector of features $h(\mathbf{z})$ and the vector of coefficients β.

As explained in Phillips et al. (2006), minimizing relative entropy results in a Gibbs distribution.

$$\begin{aligned} f_1(\mathbf{z}) &= f(\mathbf{z})\mathrm{e}^{\eta(\mathbf{z})} \\ \text{where } \eta(\mathbf{z}) &= \alpha + \beta \cdot h(\mathbf{z}) \end{aligned} \tag{4}$$

and α is a normalizing constant that ensures that $f_1(\mathbf{z})$ integrates (sums) to 1. So now we want to solve for β. It has been theoretically proved (Elith et al., 2011) that the process of solving β is equivalent to minimizing the relative entropy subject to constraints. It has been confirmed that[3] maximizing the penalized log likelihood is equivalent to minimizing the relative entropy subject to constraints.

$$\max_{\alpha,\beta} \frac{1}{m} \sum_{i=1}^{m} \ln\left(f(\mathbf{z}_i)\, e^{\eta(\mathbf{z}_i)}\right) - \sum_{j=1}^{n} \lambda_j \left|\beta_j\right| \tag{5}$$

subject to $\int_L f(\mathbf{z}) e^{\eta(\mathbf{z})} d\mathbf{z} = 1$

where z is the feature vector for occurrence point i of m sites,

and for $j = 1...n$ features.

After rasterizing the environmental covariates and adding 10,000 random background points, the suggested transformations applying to the covariates. We take the linear, product, quadratic, hinge and threshold transformations of covariables.

4.1.3 Conclusion of the model

Maxent uses the Leave One Out algorithm to estimate $f_1(\mathbf{z})/f(\mathbf{z})$ And then according to formula (3) to calculate the distribution probability of the Asian giant hornets.

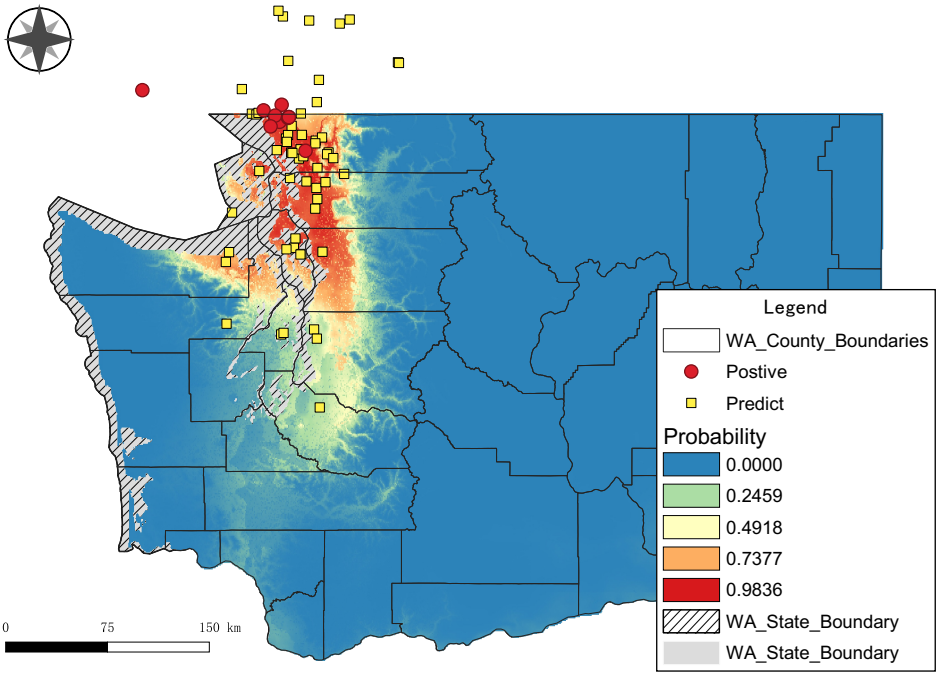


Figure 3: The map of the probability distribution of the Asian giant hornets based on MaxEnt

In the map, different colors represent different distribution probabilities.

4.1.4 Test of the model

Model accuracy test

The next picture is the receiver operating characteristic (ROC) curve for the same data, again averaged over the replicate runs. The area under the ROC curve (AUC) is used as the measurement index of model prediction effect, with a value range of 0 to 1. The larger the value is, the better the model prediction effect is. The average test AUC for the replicate runs is 0.978, and the standard deviation is 0.017.

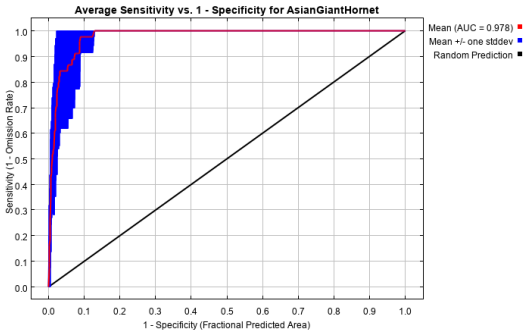


Figure 4: Area under curve for Asian giant hornets

The prediction accuracy of this model for the probability distribution of Asian giant hornets reached an excellent level.

Analysis of variable contributions

The following picture shows the results of the jackknife test of variable importance.

Variable	Percent contribution	Permutation importance
Max_Temperature_of_Warmest_Month	40	0.8
forest_cover	25.7	25.6
Precipitation_of_Wettest_Quarter	24.6	60.2
Precipitation_of_Driest_Quarter	5.1	3.4

Figure 5: estimates of relative contributions of the environmental variables

From the Figure 5, we can know that the relative contribution of Max Temperature of Warmest Month is 40%, which is the highest among the variables.

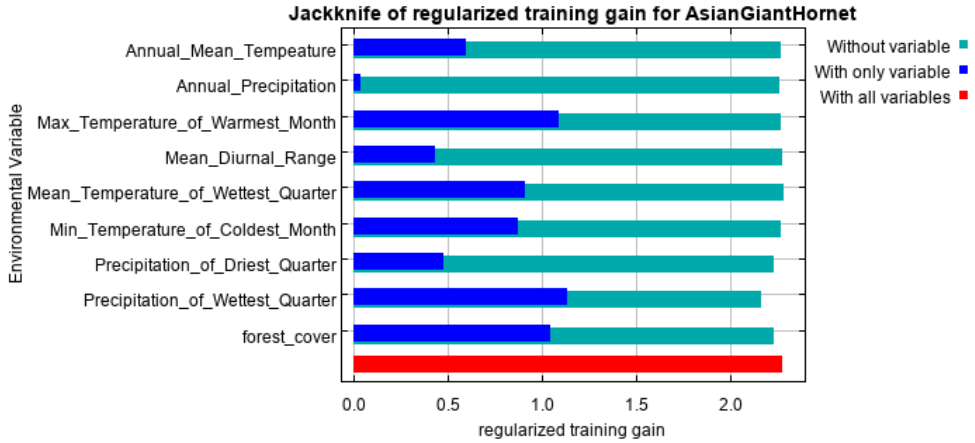


Figure 6: Jackknife of traing gain

The environmental variable with highest gain when used in isolation is Precipitation of Wettest Quarter, which therefore appears to have the most useful information by itself. The environmental variable that decreases the gain the most when it is omitted is Precipitation of Wettest Quarter, which therefore appears to have the most information that isn't present in the other variables.

4.2 Scoring model of the reports

In this section, we are going to create a model to predict the likelihood of a mistaken classification.

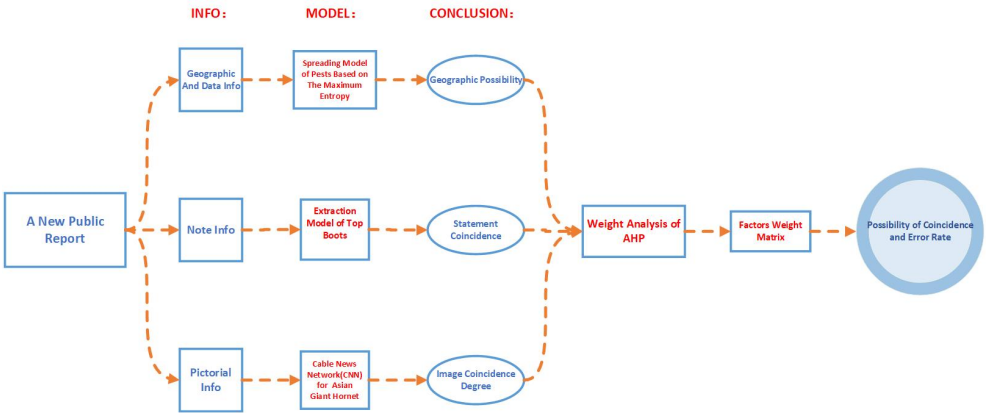


Figure 7: The flow chart of Scroing model

4.2.1 Extraction model of top roots based on NLP

- Screen the reports: select the reports that were detected after September 2019.
- Remove the valid word based on the Python-based NLTK library:remove prepositions, conjunctions, articles, and pronouns.
- Extract the words of high frequency.
- Extract the stem of words of high frequency:remove affixes from verbs, nouns, adjectives, and adverbs to obtain root words.

In all the data sets provided, the data can be divided into 4 categories. Among them, there are more than two thousand pieces of negative data, accounting for about 50% of all the data, and most of the negative data have comments given by the laboratory. Most of the negative sightings have mistaken other species of bees for Asian hornets. Because there are many descriptions and expressions in the negative data that are inconsistent with the characteristics of the Asian Hornet. We decided to use the NLP method to count, filter and analyze the words in the messages and comments of the negative data, and obtain relevant words with a high probability of false witnessing. And the words are classified according to the wasp characteristics described by the words.

$$\text{CLASS} = [c_1, c_2, \ldots, c_i, \ldots, c_a]$$

Negative word classification results:

$$\text{CLASS}_n = [\text{ "color","environment","size","appellation","area", "feeling", "other"}]$$

Table 2: Negative

color	environment	size	appellation	area	feeling	other
red	wall	small	bee	Pennsylvania	slight	teardrop
brown	high	little	yellowjacket	country	mild	drop
white	tall	tiny	paper	island	marginally	solo

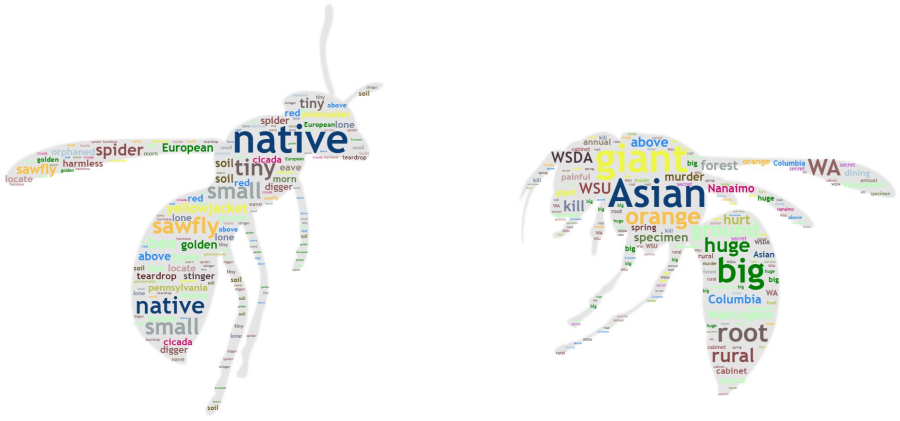
For keywords that can reflect the sightings as positive, considering that the amount of positive data in the data table is small, and the unverified data cannot be used as a statistical basis

for correct vocabulary, we analyzed the text data and selected positive vocabulary of description. Corresponding to the classification of positive words and negative words, a corresponding feature classification vocabulary set is formed.

$$\text{CLASS}_{\,p} = \text{CLASS}_{\,n}$$

Table 3: Positive

color	environment	size	appellation	area	feeling	other
orange	underground	giant	murder	british	painful	social
yellow	forest	big	Asian	columbia	allergic	annual
stripe	root	huge	tiger	Washington	hurt	spring



(a) Negative (b) Positive

Figure 8: negative and positive

We will assign the weight of each feature category according to the importance of the features described by the two types of vocabulary that we have counted, so as to obtain the feature classification weight matrix of the two types of word sets. $S = [s_1, s_2, \ldots, s_i, \ldots, s_a]$

s_i corresponds to the weight of c_i. For a newly obtained message, first use NLP to divide and process the sentence to obtain the vocabulary of nouns, adjectives and other description features. Set the initial value M_0 of the message evaluation index to 0.5. Then compare the filtered vocabulary with the negative and positive word sets. If the matching is successful, the corresponding feature weight of the word is obtained according to the feature classification weight matrix of the word set, the evaluation index is updated and calculated, and the final message score is obtained after iteration. $M_{j+1} = M_j + M_j s_i$

4.2.2 Image recognition based on convolutional neural network(CNN)

Introduction of CNN model

Convolutional neural network is widely used in image recognition in recent years.

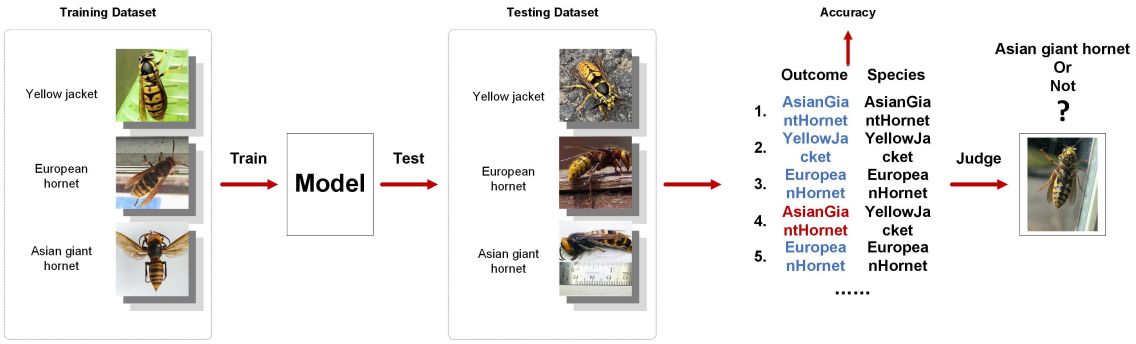


Figure 9: Image recognition process based on CNN model

As the picture shows, there two steps of the process of image recognition by convolutional neural network. First of all, set the parameters according to the existing training. Secondly, calculate the likelyhood of a right classification for a given image according to the parameters.

CNN consists of convolutional layer, excitation layer, pooling layer, full connection layer and so on[9]. To put it simply, the work of the convolution layer is to obtain the feature map of the image through the convolution operation. Images are stored in the computer as pixels. For a computer, an image is just a matrix of 0 or 1 elements. The parameter to be trained by the convolutional neural network is the convolutional kernel in the convolutional layer. Convolving two matrices is just multiplying the corresponding entries together and adding them together. According to the preset step size, select a series of submatrices rehearsed in order in the matrix. The solution process of convolution kernel is as follows:

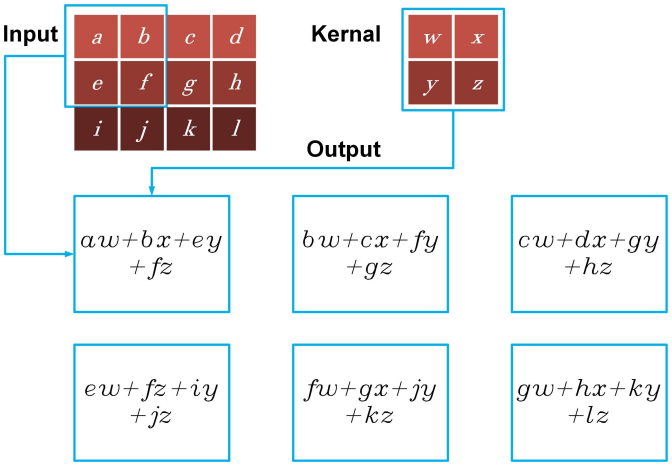


Figure 10: The solution of the convolution kernel

Take the submatrix and convolve it and then take the average. The average value is used to represent the submatrix to get the feature graph of the original pixel matrix. The excitation layer simplifies the operation by nonlinear activation mapping. The biggest effect of the pooling layer is to reduce the amount of data. For the pixel square matrix of fixed size in the feature graph obtained by the convolution layer, the maximum value or average value is used to represent the square matrix, so that the feature graph with less data volume can be obtained. A convolutional neural network can make multiple composites of the convolutional layer, excitation layer, and pooling layer. Finally, the likelihood of the right classification is obtained by applying the full connection network to the feature graph after multiple compositions. Next, we will graphically summarize convolutional neural networks with a diagram.

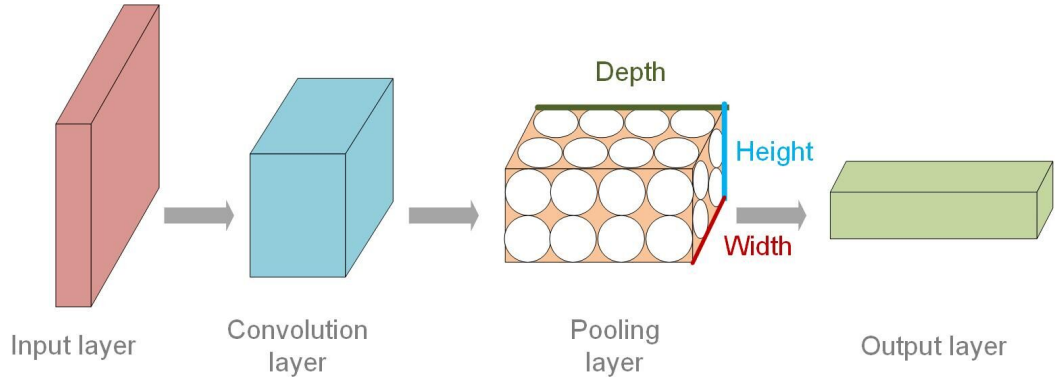


Figure 11: Convolutional neural network model diagram

The parameter to be trained for convolutional neural network is convolutional kernel.

The initial convolution kernel is defined randomly, and then we need to train the convolution kernel. We train the convolution kernel with a tagged training set. Suppose that we're going to use convolutional neural network algorithms to identify cats and dogs, the training set should be a collection of pictures of cats and dogs. Images in the training set were pre-tagged with either a cat or a dog. And then use the randomly generated convolution kernel to calculate the probability that the image in the training set is a cat or a dog. Then the Euclidean radial basis function is used to calculate the deviation between the probability obtained by the convolution kernel and the real probability. According to the error obtained, BP back propagation algorithm is used to correct the convolution kernel.

Convolutional neural network is used for image recognition, that is, the trained convolutional kernel is used to calculate likelihood of right classification of images.

Identification of Asian giant hornets based on CNN model

Firstly, we found the publicly available image recognition training set[10] for the Asian giant hornets. A convolution kernel is trained according to the images in the training set. The trained convolution kernel is used to determine the correct probability of a given image set.

In the convolution layer, the Feature map of the upper layer is convolved by a learnable convolution kernel, and then the output Feature map can be obtained through an Activation function[11].

$$x_j^l = f\left(\sum_{i\in M_j} x_i^{l-1} \otimes W_{ij}^l + b_j^l\right) \tag{6}$$

$(\sum_{i\in M_j} x_i^{l-1} \otimes W_{ij}^l + b_j^l)$,the net activation of the convolutional layer is obtained by convolutional summation and bias of the output Feature map x_i^{l-1} of the previous layer. x_j^l is the output of the J channel of convolutional layer I. f is the activation function. $\otimes$represents Convolution Operation. W_j represents a subset of the input feature map used to compute the u_j^l. W_{ij}^l is the convolution kernel matrix, and b_j^l is the bias of the convoluted feature map.

In the lower sample layer, the feature map of each input is mapped into the output feature map.

$$x_j^l = f\left(\beta_j^l \operatorname{down}\left(x_j^{l-1}\right) + b_j^l\right) \tag{7}$$

$\beta_j^l(\operatorname{down}\left(x_j^{l-1}\right) + b_j^l)$ is obtained from the output feature map x_j^{l-1} of the previous layer after weighted subsampling and bias. It is the net activation of J channel in the lower sampling

layer I. β represents the weight coefficient of the lower sampling layer. $b_j^l$is the offset term. *down* represents the subsampling function. The purpose of using the subampling function is to reduce the amount of data.

In the fully connected layer, the feature images of two-dimensional images are spliced into one-dimensional features as the input of the fully connected network.

$$x^l = f\left(w^l x^{l-1} + b^l\right) \tag{8}$$

$k^l$is the weight coefficient of the fully connected network. $b^l$is the offset term of fully connected layer I

Finally, BP back propagation algorithm was used to optimize the parameters. The parameters to be optimized include: convolution kernel parameter k, network weight β of the lower sampling layer, network weight w of the full connection layer and bias parameter b of each layer. The square of the difference between the expected and actual output at the output end is used as an error function

$$E = \frac{1}{2}\sum_{n=1}^{N} \|t_n - y_n\|^2 \tag{9}$$

$t_n$represents the true value and y_n represents the predicted value.

The gradient:

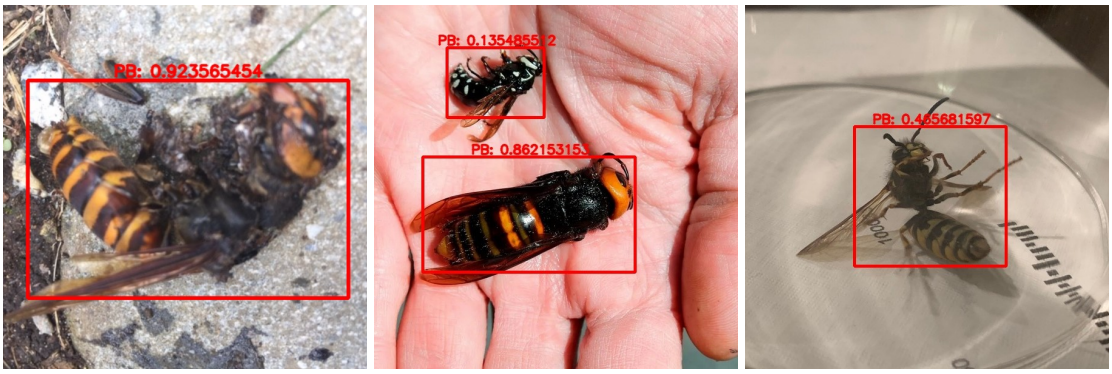
$$\nabla E = \left(\frac{\partial E}{\partial w}, \frac{\partial E}{\partial \beta}, \frac{\partial E}{\partial k}, \frac{\partial E}{\partial b}\right) \tag{10}$$

Gradient descent formula:

$$\Theta^1 = \Theta^0 - \alpha^{'} \nabla E(\Theta) \tag{11}$$

$\alpha^{'}$ is the learning rate. The learning rate affects the accuracy of the algorithm. By adjusting the learning rate in the program, the optimal solution of the parameters can be obtained.

Image recognition results of Asian giant hornets based on CNN

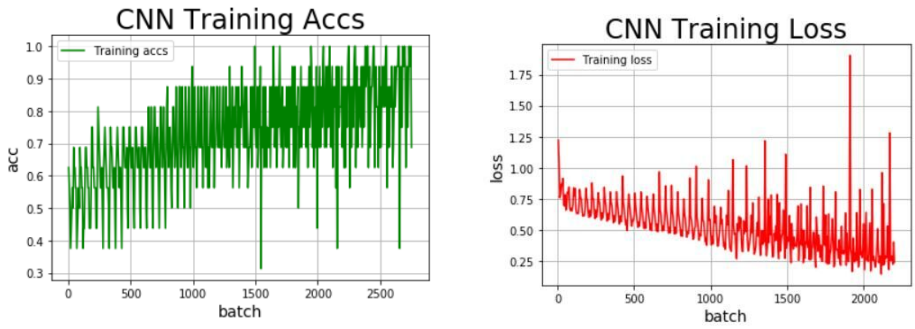


(a) 1 (b) 2 (c) 3

Figure 12: Image recognition

The probability of the creature in the picture is an Asian giant hornet is shown in the picture.

Test of the CNN model



(a) Accuracy of CNN model (b) Loss of CNN model

Figure 13: test of the CNN model

From the images, we can see that the accuracy rate of CNN training model is about 60%, and the error is 20%. It can be considered that the trained model is accurate and can be put into use.

4.2.3 Weight analysis of AHP

Score respectively

In this section, we consider the likelihood of classificaton from three dimensions of Notes, image, longitude and latitude. The contribution of each report was quantified by scoring. Each item has a maximum score of 100, and the score for each of these three items is C_1 for Notes, C_2 for image, and C_3 for location(longitude and latitude). According to the Compare-Weighting method of Analytic hierarchy process, the weight of each indicator $\omega_1, \omega_2, \omega_3$ could be known. The final score C can be calculated using the following formula:

$$C = C_1 \times \omega_1 + C_2 \times \omega_2 + C_3 \times \omega_3 \tag{12}$$

The principle of scoring is as follows.

1. Notes:After processing Lab comments corresponding to Negative ID with NLP natural language, 10 high-frequency word roots were konwn. If the notes in the report contain M words with high-frequency word roots that we screened. C_1 could be known by the following formula.

$$C_1 = 100M \tag{13}$$

2. Image:The probability δ of each image being an Asian giant hornets is known from the CNN model. C_2 could be known by the following formula.

$$C_2 = 100 \times \delta \tag{14}$$

3. Latitude and longitude:According to the model established in Question 1, we can get the probability ζ that the locations in reports with the distribution of Asian giant hornets. C_3 could be known by the following formula.

$$C_3 = 100 \times \zeta \tag{15}$$

Determine the weight

Notes, image and location(latitude and longitude) are equivalent to the criterion layer in the analytic hierarchy process. The weight determination method is as follows:

Now, there are three existing decision factors C_1, C_2, C_3 Take two factors each time and use a_{ij} to represent the ratio of C_i to C_j 's impact on the final score. All the comparison results can be expressed by paired comparison matrices.

$$A = (a_{ij})_{n\times n}, a_{ij} > 0, a_{ji} = 1/a_{ij} \tag{16}$$

The corresponding matrix is:

$$A = \begin{pmatrix} 1 & 5 & 5 \\ 0.2 & 1 & 3 \\ 0.2 & 1/3 & 1 \end{pmatrix}$$

Weighting formula:

$$\omega_i = \frac{1}{n}\sum_{j=1}^{n}\frac{a_{ij}}{\sum_{k=1}^{n} a_{kj}} \tag{17}$$

We could know the ω_i by the formula.

$\omega_1 = 0.11$, $\omega_2 = 0.21$, $\omega_3 = 0.68$.

λ denotes the maximum eigenvector of matrix A, whose value is 3.13, calculated by Matlab.

Finally, we need to check the consistency of the matrix.

Consistency index

$$CI = \frac{\lambda - n}{n-1} = 0.065 \tag{18}$$

For fixed n, Random consistency index is also fixed[12].

$$RI = 0.58 \tag{19}$$

Consistency ratio

$$CR = \frac{CI}{RI} = 0.09 \tag{20}$$

The inconsistency of matrix A is less than 0.1, which is considered to be within the allowable range.

4.2.4 Conslusion of the scoring model

1.Predict the likelihood of a mistaken classification

d donates the likelihood of a mistaken classification.

$$d = 1 - \frac{C}{100} = 1 - \frac{C_1\times\omega_1 + C_2\times\omega_2 + C_3\times\omega_3}{100} \tag{21}$$

2.Prioritize investigation of the reports

According to the scoring model we created in Question 2, each report has a corresponding score $(C = C_1\omega_1 + C_2\omega_2 + C_3\omega_3)$. Prioritize investigation of the reports in the order of score each report gets.

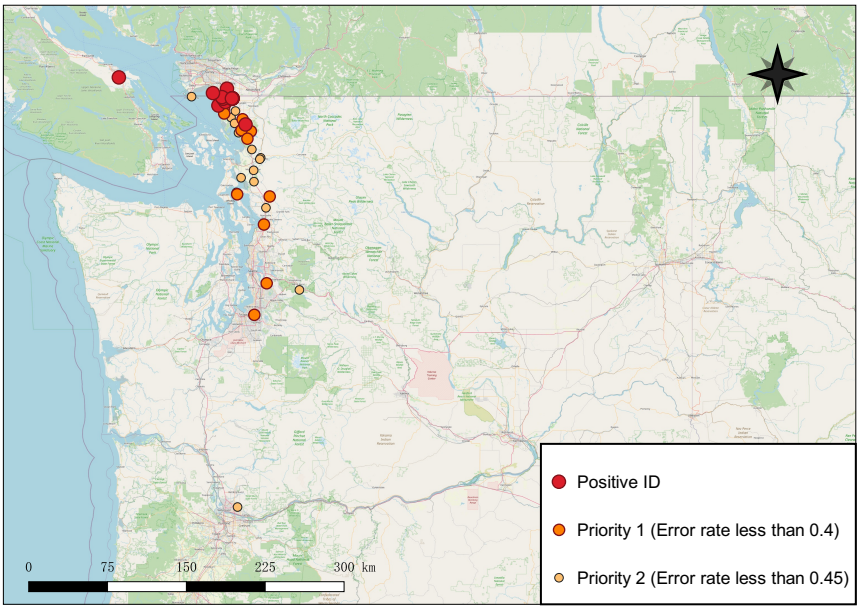


Figure 14: distribution of the reports with prioritization

In the figure, we have marked the positive, the part with error probability less than 0.4 and the part with error probability less than 0.45. Reports with an error probability of less than 0.4 are considered of high value and should be investigated first. Reports with an error probability of less than 0.45 are considered generally valuable and should be investigated as a secondary priority.

4.2.5 Test of scoring model

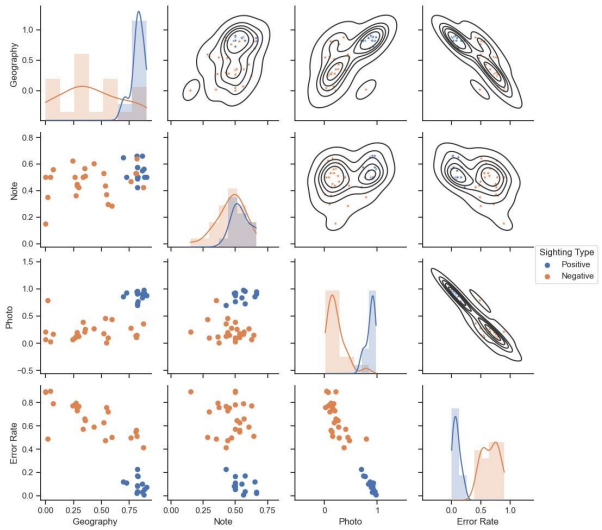


Figure 15: Test of scoring model

Our scoring model was used to test the positive ID and negative ID in the data set. It can be clearly seen from the figure that the two parts are clearly separated, indicating that our model is reasonable.

4.2.6 Update of the scoring model

In this section, we are going to address Question 4. That is how to update our model given additional new reports over time, and how often the updates should occur.

The core formula of our scoring model is $C = C_1\omega_1 + C_2\omega_2 + C_3\omega_3 = (100M)\omega_1 + 100\delta\omega_2 + 100\zeta\omega_3$ Among them, with the new reports, roots of high frequency changes over time. In addition, the probability distribution of Asian giant hornets based on the "Spreading model over time" will also makes changes.

If there are newly added reports, we need to update M to $M^{'}$ in formula $C_1 = 100M$ and update ζ to $\zeta^{'}$ in formula $C_3 = 100 \times \zeta$. NLP algorithm was used to extract the Lab Comments roots of high frequency of Negative ID. And then it is necessary to update the top 10 of roots.

At the same time, the value of $f(\mathbf{z})$ and $f_1(\mathbf{z})$ of the Asian giant hornets propagation model over time were changed to solve the probability distribution again. And then we use the following formula to update the score corresponding to the report.

$$C = (100M^{'})\omega_1 + 100\delta^{'}\omega_2 + 100\zeta^{'}\omega_3 \tag{22}$$

When it comes to the frequency of updates, even though the model we create can be updated timely, the impact of a few newly added reports on the parameters in the model is ignorable. According to the Scoring model, among more than 4000 reports provided in the data set, there are 40 samples with the likelihood of ture classification exceeding 60%.

τ donates the likelihood of true classification of a new report.

$$\tau = \frac{40}{4000} = \frac{1}{100}$$

Estimate the overall property through the sample, we know that the likelihood of ture classification of one sample in about 100 is more than 60%. So it is sensible to update the model for every 100 reports added.

4.3 Eradication judgment of Asian giant hornets

To determine and test pest control in Washington State, we modeled the population changes of Asian bumble bees over time.

Again, in this model we assume

- The Asian has no natural enemies in Washington State
- Rich in natural resources that is needed for the pests to survive.
- The population of Asian giant hornets does not overlap over generations

4.3.1 The growth of the queens population

On the premise of satisfying the model hypothesis, the number of queens of Asian giant hornets can be described by the exponential growth model. Assume that a queen can produce ϕ mating females during reproduction, the initial number of queen bees in Washington is n_0. Take the queen reproduction cycle as a unit of month, the growth model of queen bee population is as follows:

$$N_t = \phi^{\frac{1}{12}} N_{t-1} \tag{23}$$

4.3.2 The growth of the workers population

Considering that the life cycle of workers is one year, the population slowly increases throughout the spring and summer, peaking at about 100 bees per nest in August. Therefore, we construct a periodic linear change function to describe this phenomenon.

$$\varphi(t) = \begin{cases} \frac{100}{8}t & 0 < t < \frac{2}{3}T_0 \\ -\frac{100}{4}t + 300 & \frac{2}{3}T_0 < t < T_0 \end{cases} \tag{24}$$

The period T_0 is 12 months, and the unit of independent variable is month. Considering that a queen can represent a hive, the change model of the total number of workers is as follows:

$$N_w(t) = N_0(t) \cdot \varphi(t) \tag{25}$$

4.3.3 Human intervention in population size

Monthly elimination

This method of elimination is mainly aimed at the workers who often fly outside. Assuming that the number of elimination workers per month is a, then the factor is:

$$n_1 = a\varphi(t) \tag{26}$$

Nests destory

This extermination may not kill all the workers at once, but it can kill the queen and prevent further reproduction of the colony. Assuming that the number of nests destroyed per month is γ, the factor is

$$n_2 = \gamma t \tag{27}$$

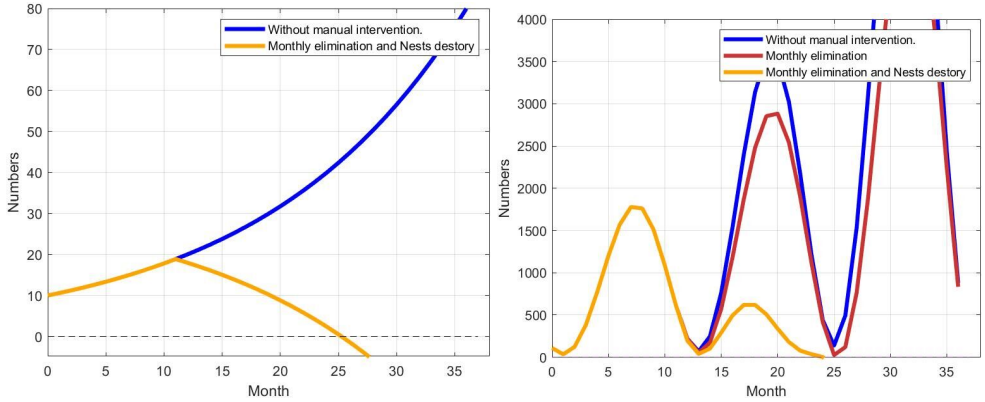
In summary, the model of Asian giant hornets population changes over time can be defined as:

$$N(t) = ((n_0 + \frac{\gamma}{1-\phi^{\frac{1}{12}}})\phi^{\frac{t-1}{12}} - \frac{\phi}{1-n^{\frac{1}{12}}})(1+\varphi(t)) - \gamma t \tag{28}$$

According to the population change model, we can draw three population change lines of Asian giant hornets.

- Without manual intervention.
- Monthly elimination.
- Monthly elimination and Nests destory.

The three curves are shown in the square diagram below.



(a) The change in the number of queens (b) Total population change diagram

Figure 16: Graph of changes in population size,measures were taken from September 2020

It can be seen that the effect of "Nests destory" is extremely significant. Therefore, according to our model, when the number of destroyed nests is 2 per month, the population of the pests will decline after 25 months and eventually fall to zero.

5 Model commentary

5.1 Spreading model of pests based on the Maximum entropy

Strengths

- High accuracy:maximum entropy model of all the models that meet the constraints of the maximum entropy of information model.
- Flexibility of constraint selection:it is convenient to adjust the adaptability of the model to unknown data and the degree of fitting to known data.

Weaknesses

- Large computation:The maximum entropy model has a large amount of computation by iterating the parameters.

5.2 Scoring model of the reports

Strengths

- High sample recognition rate and prediction accuracy of CNN: Compared with the shallow learning model, deep learning constructs the learning model with multiple hidden layers, so as to process the big data and improve the sample recognition rate or prediction accuracy.
- Reduction of subjectivity and uncertainty of AHP:through the method of influencing comparison, subjectivity is reduced to some extent.

Weaknesses

- The CNN model requires manual adjustment of parameters.

Memorandum

To: Washington State Department of Agriculture
From: Team 2108410
Date: February 9th, 2021
Subject: Research results about Asian giant hornets

In short, our team create "Spreading model of pests" and "Scoring model of the reports" to solve the problem.

As for the spread of Asian giant hornets over time, the results are clearly shown in following Figure. Based on the existing distribution of pests, known as the Positive ID and its latitude and longitude, we predicted the probability distribution of Asian giant hornets across Washington State. And plotted the probability distribution of Asian giant hornets in various parts of Washington State on the following Figure. It divides Washington State into a number of blocks based on the probability of pest distribution, clearly marked in different colors in Figure.

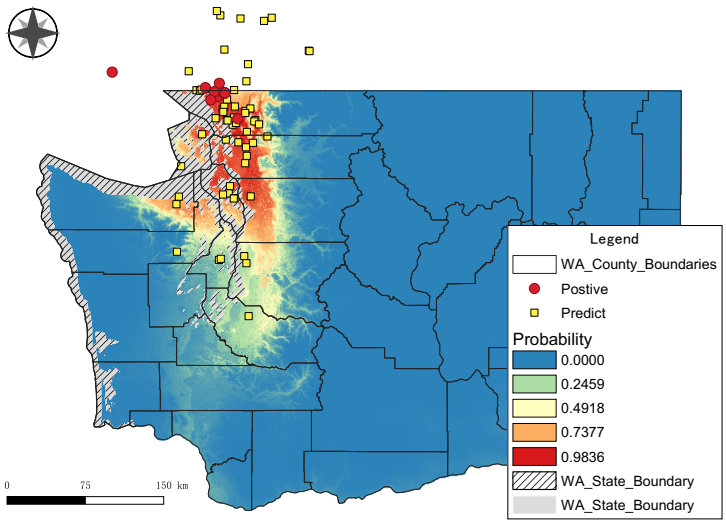


Figure 17: A map of the distribution of the Asian giant hornets in Washington State

As for the analysis of the data provided by the public reports, our team scored each report from three dimensions: Notes, Images, Latitude and Longitude. Our team judge the value of each report according to the score. Rank the scores from highest to lowest, and prioritize the investigation according to the score.

According to the data provided in the data set, except the positive ID and negative ID that have been processed, our team deal with report whose Lab status is unverified or unprocessed. We have found that the value of a report can be judged by its offerings of Notes, Images, Location(latitude and longitude). We scored the report from three aspects: Notes, Images, Latitude and Longitude, and then weighted the scores to get the final score of a report. Notes would be valuable if it contained key information about the Asian giant hornets. For example, yellow heads, a black thorax, and yellow and black or brown striped abdomens, etc on the characteristics of the pests. Its score is determined by the number of top roots it contains. The Images are also the necessary information to judge whether a report is valuable or not. According to the model we trained, the probability of the images being true determines its score. Latitude and longitude are also attached importance to. So far we have a forecast of the probability distribution of the pests in Washington State. We believe that a report is more valuable if the locations

have a greater probability of the pests. The probability of pest distribution determines the score. After we got the scores of the three dimensions, our team gave weight to each score to get its total score. In order to get more objective results, our team learned from the comparative weight method in the analytic hierarchy process(AHP) to get the contribution degree of the three parts to the total score. The weight of the three dimensions are 11%, 21%, 68% The likelihood of a mistaken classification can be determined after the score of each report is calculated. By sorting the scores in order of highest to lowest, the priority of investigation could be konwn.

We also need to consider how we should judge the investigative value of new reports if they are added. After analysis, our team got the result that the model was updated every 100 Reports. Although our model can be updated timely, a few reports have minimal impact on the parameters in the model. At the same time, we found that there was at least an 60% that the Asian giant hornets was true in the dataset of 40 samples. We then infer that there will be one sample out of 100 with a probability of more than 60% that it is a real Asian giant hornets. So it would be sensible to update the model every 100 reports.

Given that only queens are allowed to migrate, even though worker ants outnumber queens, the task of killing hornets should focus on trapping queens and destroying nests. As an annual species, Asian giant hornets build new nests every year. When winter comes, all the hornets in the nest are dead except for the queens. When spring comes, the overwintering queens begin to build nests underground. And these nests grow slowly throughout the spring and summer. It peaked at about 100 workers in August. Then queens begin to produce males and females in September. In October and November, males and females leave the nest to mate. When the males and females leave, the colony is in disarray, and the nest becomes dead again in winter except for the queens.

Given the reproductive characteristics of the colony, the methods used to eliminate the Asian bumblebee should be distinctly seasonal. In the early spring and late fall, the whole population is in the stage of the queen colony, the pre-nesting stage, the breeding stage, and the probability that the hornets is the queen is greater at this time. It is a good method to directly trap queens. While during the summer (July to October), the population is in the cooperative and polyether stages, nests can be located and destroyed. Asian giant hornets are used as food and medicine in many Asian countries. After investigation, our team suggests that when it comes to eliminating the hornets, government should consider capturing nests during the larvae and pupae season to make food for them. During the adult season of the pest, the adult hornets are captured to make medicinal wine.

Based on our population growth model, we found that destruction of nests had the greatest impact on the populations in Asia giant hornets. And if two nests are destroyed every month, the pests will be completely eradicated after 25 months.

In addition, our team has several suggestions for the control of Asian giant hornets.

- Establish a unified coordination and management organization: coordinat manpower and material resources.
- Monitor sequentially: track and monitor sites with a high probability of the pests and destroy the nests as soon as possible.
- Improve the evaluation system: further improve the risk assessment system with respect to reports so that limited resources can be devoted to useful investigations.

References

[1] Washington State Department of Agriculture. 2020 Asian Giant Hornet Public Dashboard. https://agr.wa.gov/departments/insects-pests-and-weeds/insects/hornets/data

[2] And, Aleksandar Radosavljevic , and R. P. Anderson . "Making better Maxent models of species distributions: complexity, overfitting and evaluation." *Journal of Biogeography* (2014).

[3] Elith, Jane, et al. "A statistical explanation of MaxEnt for ecologists."*Diversity and distributions*17.1 (2011): 43-57.

[4] "WA State Boundary."*Data-Wadnr.Opendata.Arcgis.Com* 2021, https://data-wadnr.opendata.arcgis.com/datasets/wa-state-boundary

[5] "WA County Boundaries".*Data-Wadnr.Opendata.Arcgis.Com*, 2021, https://data-wadnr.opendata.arcgis.com/datasets/wa-county-boundaries.

[6] 2021, https://biogeo.ucdavis.edu/data.Accessed 7 Feb 2021. "Global Forest Change | Google Crisis Map".*Google Crisis Map*, 2021, http://earthenginepartners.appspot.com/science-2013-global-forest.

[7] "High-Resolution Global Soil Moisture Map."*Jpl.Nasa.Gov*, 2021, https://www.jpl.nasa.gov/images/high-resolution-global-soil-moisture-map.

[8] "Global Forest Change | Google Crisis Map".*Google Crisis Map*, 2021, http://earthenginepartners.appspot.com/science-2013-global-forest.

[9] CHANG Liang, DENG Xiao-Ming, ZHOU Ming-Quan, WU Zhong-Ke, YUAN Ye, YANG Shuo, WANG Hong-An . Convolutional Neural Networks in Image Understanding. Acta Automatica Sinica, 2016, 42(9): 1300-1312. doi: 10.16383/j.aas.2016.c150800

[10] "Bee Or Wasp?"*Kaggle.Com,* 2021, https://www.kaggle.com/jerzydziewierz/bee-vs-wasp.

[11] Fu L S, et al." Convolutional neural network based image recognition for multi-cluster kiwifruit in field." *Transactions of the Chinese Society for Agricultural Engineering 34*.02(2018):205-211. doi:.

[12] Foroughi, M. and Struckmeier, J., 2003. The mathematical model. Hamburg: Fachbereich Mathematik der Univ. Hamburg.

[13] Kimathi, Emily, et al. "Prediction of breeding regions for the desert locust Schistocerca gregaria in East Africa."*Scientific Reports*ă10.1 (2020): 1-10.

[14] "Asian Giant Hornets". Penn State Extension, 2021, https://extension.psu.edu/asian-giant-hornets.

Source of the figures used in the article

- Figure 1 was drawn by ourselves.
- Figure 2, 3, 4, 5, 6 were drawn by ourselves with MaxEnt tool.
- Figure 7, 9, 10 ,11 were drawn by ourselves with Visio.
- Figure 8a, 8b was drawn by ourselves.
- Figure 12 was from the image files provided.
- Figure 13, 15 were drawn by ourselves with python.
- Figure 14 was drawn by ourselves with qgis 3.16.3
- Figure 16 was drawn by ourselves with matlab.

Appendix B: Program Codes

CNN.py

```
class CNNnet(fluid.dygraph.Layer):
    def __init__(self):
        super(CNNnet, self).__init__()
        self.convpool01 = ConvPool(3, 256, 3, 2, 2,2, act="relu")
        self.convpool02 = ConvPool(256, 512, 3, 2, 2,3, act="relu")
        self.pool_5_shape = 512 * 7* 7
        self.fc01 = fluid.dygraph.Linear(self.pool_5_shape,4096,
            act="relu")
        self.fc02 = fluid.dygraph.Linear(4096,train_parameters
        ['class_dim'],act="softmax")
    def forward(self, inputs, label=None):
        out = self.convpool01(inputs)
        out = self.convpool02(out)
        out = fluid.layers.reshape(out, shape=[-1, 512*7*7])
        out = self.fc01(out)
        out = self.fc02(out)
        if label is not None:
            acc = fluid.layers.accuracy(input=out, label=label)
            return out, acc
        else:
            return out
```

enmevaluate.R

```
bg <- xyFromCell(dens.ras2, sample(which
(!is.na(values(subset(env, 1)))),
 10000, prob=values(dens.ras2)[!is.na(values(subset(env, 1)))]))

enmeval_results <- ENMevaluate(occ, env, method=
"randomkfold", kfolds = 9,
algorithm='maxent.jar', bg.coords = bg)
```