Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2123737

No Where to Hide: Find Vespa Mandarinia with Data

Summary

Owing to Vespa mandarinia's catastrophic impact on European honeybees and its appearance similarity to other wasps, to precisely detect the presence of Vespa mandarinia can efficiently cut down on the economic and agricultural loss and anxiety of the locals. One main source of the sightings of these hornets is the public reports while the quality and precision of these reports can't always be ensured, therefore, how to correctly interpret the data from the reports and prioritize the limited sources to make suitable strategies accordingly has become a concern for the government and public.

To distinguish whether the spread of the pest can be predicted or not, we utilize time series analysis method under the assumption that the change of the pest's location is a wide-sense stationary random process. Considering the weak temporality of latitude and longitude, we then use autoregressive model to predict the location of the pest with an absolute error of latitude less than 0.05 and of longitude less than 0.2.

When it comes to predicting the likelihood of a mistaken classification, we propose Twice Discriminant Model. We first use logistic regression model to locate the major area of the pest's appearance location and achieve a precision of 92.86% on the provided data set. In order to discriminate those negative samples that are mixed up with the positive ones, we train the data in the LeNet convolutional neural network and balance the difference between the number of positive and negative samples with oversampling and undersampling methods. Values of precision and F1-Score are used to measure the effectiveness of the model's prediction and statistical significances of parameters are guaranteed by the hypothesis testing and interval estimation.

A ranking model is designed to evaluate the priority of the reports processing based on a weighted distance measurement of the two feature indexes in the Twice Discriminant Model so that investigation can be devoted to those reports that are most likely to be positive sightings. We also discuss the model's adaptation to new reports data and the suitable frequency of providing new reports.

According to our assumptions and models, we assess the conditions of the pest's control in Washington State and write a memorandum for the Washington State Department of Agriculture, including our model results and some targeted recommendation.

Key words: Time Series Analysis, Logistic Regression, Bayesian Probability, LeNet CNN, Hypothesis Testing, Weighted Distance Measure

Contents

1. Introduction

1.1 Restatement of the Problem

The invasion of alien species will not only destroy the local ecological agriculture development, but also affect the survival of other species, and even directly attack human beings, interfere with normal life, and cause economic losses. Using statistical modeling methods to analyze historical data, we can help staff to more accurately locate the scope of pest activities and eliminate them in the early stages of their development. To efficiently deploy limited resources, we need to use the data in historical reports to complete the following tasks:

- Determine the spread discipline of the pest over time
- Establish a classification model of the report to enable it to predict the possibility of misclassification
- Establish a prioritization model for processing reports to achieve a reasonable allocation of limited resources
- Propose the update method and update frequency of the model
- Assess the current pest control situation in Washington State to determine whether the pests have been eliminated

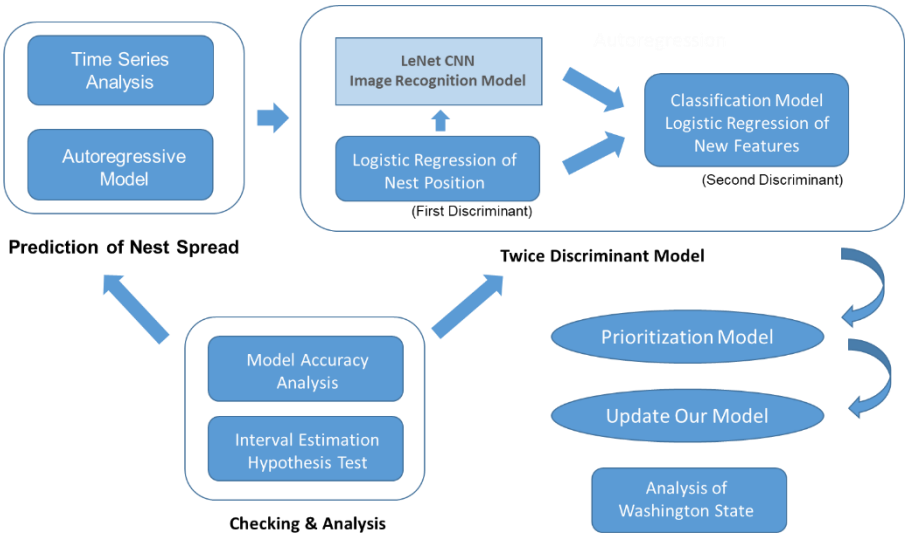
1.2 Our Work

We mainly carried out statistical modeling for the geographic information and picture information in the report and proposed a series of models for realizing the prediction of nests' propagation, classification of reports, and ranking processing priority of reports.

In Section 4, we assume that the position change of the nests is a generalized stationary random process. We use time series analysis and autoregressive models to model the propagation characteristics of the nests and test the statistical significance of the parameters in the model.

In Section 5, based on the logistic regression model and the image recognition model, we propose a secondary discriminant model for report classification, which comprehensively considers the location information and picture information in the report. In the image recognition model, we used the classic LeNet convolutional neural network for training, and combined the oversampling and undersampling methods to deal with the imbalance of positive and negative samples. The model uses a combination of accuracy and F1-Score to measure the quality of the model. We have performed hypothesis testing and interval estimation of the statistical significance of parameters for all statistical models used, as well as the significance testing of the model itself. Based on Bayesian probability, we also give a model for the possibility of misclassification in the model. Moreover, the calculation of the prior probability in the model is done by using the law of large numbers.

In Section 6, we use two feature indicators in the twice discriminant model, based on the weighted distance measurement formula, to establish a report processing priority ranking model. The Section 7 introduces the update method of the model, and Section 8 evaluates the pest control situation in Washington State. Finally, we discussed the sources of modeling errors and the ways to improve the models in Section 9. The modeling process is shown below.



2. Model Assumptions

1. Assume that the random noise in the autoregressive model obeys a normal distribution with a mean value of 0 and a constant variance.
2. In the logistic regression model, it is assumed that the target value of the data obeys the Bernoulli distribution (only values 0 or 1). Therefore, in the model training phase, we do not consider the situation, in which the value of "Lab Status" to be "Unprocessed" and " Unverified".
3. It is assumed that the actual classification and the results of the model discrimination reported are independent of each other.
4. Suppose that the reported true classification is a sequence of independent and identically distributed random variables, and its mathematical expectation exists.

3. Notations

Symbol	Description
$\{X_t\},\{Y_t\}$	Random process of nests' latitude and longitude changes
$R_X(v),R_Y(v)$	Autocorrelation function of the random process $\{X_t\},\{Y_t\}$
α_1,α_2,c	Undetermined parameters of the autoregressive model AR(2)
S_i	The true classification status of the report, 1 means Positive ID
Lat_i	Latitude of the i-th report
Lon_i	Longitude of the i-th report
$p(Lat_i,Lon_i)$	Output probability of logistic regression model
Y_i	Classification results output by the logistic regression model

(Continued)

p_0	The threshold value of the logistic regression model
N_i	Probability output by image recognition model
NS_i	Classification results output by the image recognition model
p_1	The threshold value of the image recognition model
$p'(p_i, N_i)$	Probability output by the twice discriminant model
C_i	Classification results output by the secondary discriminant model
p_2	The threshold value of the twice discriminant model
P^{error}	Probability of misclassification
d_i	Prioritization model
w_1, w_2	The weight coefficient of the priority ranking model d_i

4. The Spread of Vespa Mandarinia Nest over Time

We take out the report data with the field "Lab Status" as positive ID from the file (14 in total). As shown in Fig 1, the label of data point indicates the time sequence of the pest's appearance.

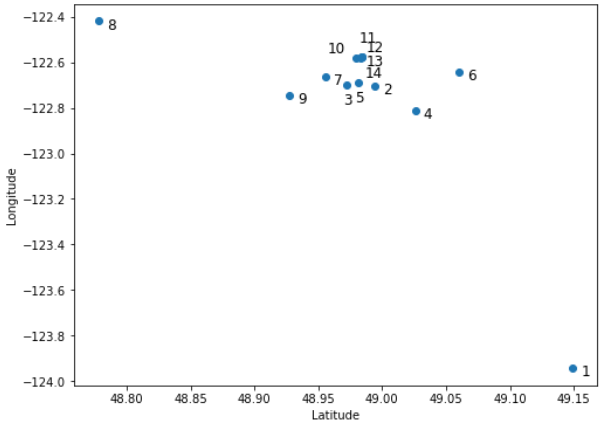


Fig 1: The Distribution of Honeycomb over Time

From the rough analysis of Fig 1, we can see that the geographical distribution of the pest is relatively concentrated, and there is no obvious time distribution trend, which is relatively random. Then we will use the method of time series analysis to discuss the spread of the pest in detail, and examine whether the distribution of the pest over time is predictable.

4.1 Time Series Analysis

We use $\{X_t\}, \{Y_t\}$ to represent the latitude and longitude data in the reports respectively, where t represents the sequence of events. Suppose that the change of the pest's position with time is a generalized stationary discrete stochastic process [1], so the mean and variance of the random process $\{X_t\}, \{Y_t\}$ are constant:

$$\begin{aligned} \mu_X &= E(\{X_t\}) \approx 48.9827463 \\ \mu_Y &= E(\{Y_t\}) \approx -122.7285601 \\ \sigma_X &= D(\{X_t\}) \approx 0.0793194 \\ \sigma_Y &= D(\{X_t\}) \approx 0.3624098 \end{aligned} \tag{4.1}$$

For processes $\{X_t\}, \{Y_t\}$, the autocorrelation functions $R_X(v), R_Y(v)$ are defined as:

$$\begin{aligned} R_X(v) &= \frac{E\left[(X_t - \mu_X)(X_{t+v} - \mu_X)\right]}{\sigma_X^2} \\ R_Y(v) &= \frac{E\left[(Y_t - \mu_Y)(Y_{t+v} - \mu_Y)\right]}{\sigma_Y^2} \end{aligned} \tag{4.2}$$

where v represents the time order. Since the processes $\{X_t\}, \{Y_t\}$ are discrete processes, we take v as integer.

Since only 14 reports have been successfully identified, we take the time order v as $1,2,\ldots,7$. Fig. 2(a) shows the variation trend of the drawing processes $\{X_t\}, \{Y_t\}$, and Fig. 2(b) is the autocorrelation function of the two processes.

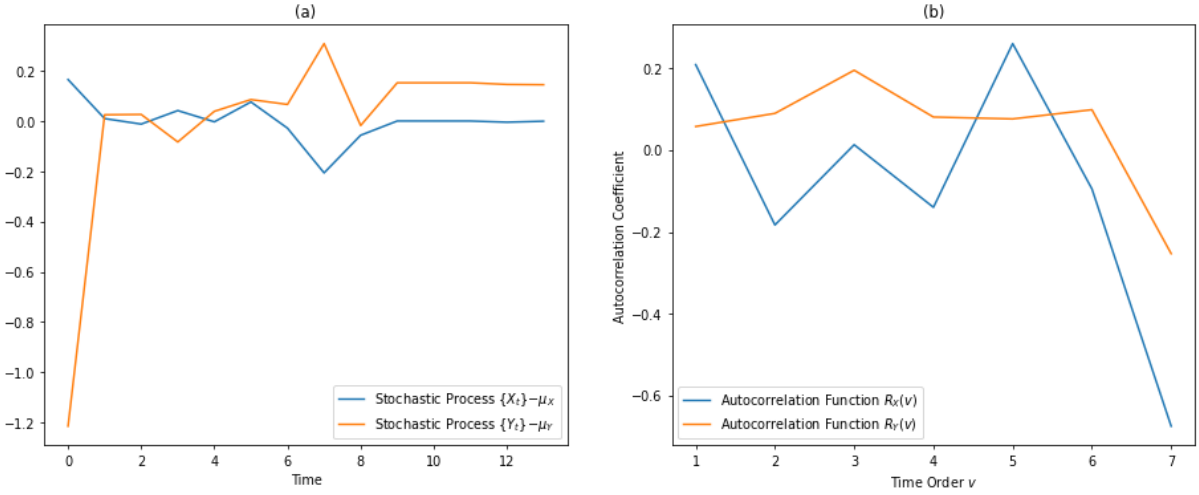


Fig 2: Process $\{X_t\}, \{Y_t\}$ changing trend (a), Autocorrelation function (b)

The value of the autocorrelation coefficient reflects the time sequential property of the random process to some extent. It describes the linear correlation between the random sequence before and after a period of time interval, so as to reflect the causality of data changes with the time. The more the absolute value of the autocorrelation coefficient tends to 1, the stronger the correlation is. Conversely, the more the value of the autocorrelation coefficient tends to 0, the weaker the correlation is. As it can be seen from Fig 2(b), when the time order v takes different values, the autocorrelation coefficient values of processes $\{X_t\}$ and $\{Y_t\}$ are both small, and both the processes show weak time sequential properties.

4.2 Predicting the Spread of Nest

From the previous section, we know that the two processes $\{X_t\}$ and $\{Y_t\}$ about the pest's location distribution have weak time sequential property, and this property is similar to that of discrete white noise. Since the change trend of the two processes is gentle and stable, and there is no obvious trend, we consider to use autoregressive model to realize the prediction of the pest's position.

For the process $\{X_t\}$, we consider the low order autoregressive model AR(2) [2]:

$$(1-\alpha_1 L-\alpha_2 L^2)X_t = c+\varepsilon_t \tag{4.3}$$

Where L is the forward time operator (i.e. $L(X_t) = X_{t-1}$), α_1, α_2 and c are the undetermined parameters of AR model. $\{\varepsilon_t\}$ is an independently identically distribution Gaussian random noise. For the existing 14 coordinates data, we divide them into the first 10 and the last 4 parts. The former part of data is used as the training set to solve the AR model's parameters, and the second part of data is used as the verification set in order to examine the accuracy of the prediction model.

Before fitting the model, two processes $\{X_t\}$ and $\{Y_t\}$ need to be zero-averaged first. For the parameters obtained from the solving, we further investigated whether they have statistical significance or not. We choose the significance level $\alpha = 0.05$ and only when the p value of the model parameter is less than a, we accept the model; otherwise, we can consider to reduce the order of the AR model, or to choose other regression models [2], [3].

We get parameters of AR model of processes $\{X_t\}, \{Y_t\}$ as shown in Table 1:

Table 1: Results of AR Regression Model Parameters

Stochastic Process	Parameter	Value	Standard Error	p-value
$\{X_t\}$	c	3.8234×10^{-4}	0.0311	0.0992
	α_1	0.5122	0.5732	0.0371
	a_2	-0.4466	0.3517	0.0204
$\{Y_t\}$	c	-0.0294	0.1729	0.0865
	α_1	0.8821	0.2265	9.847×10^{-5}
	a_2	-0.4520	0.1539	0.0033

As can be seen from the results in Table 1, AR model performs well. Except that the p value of the constant term c of the two models is slightly higher, other parameters can pass the significance level of 0.05 test. We get propagation prediction model of the pest over time:

$$\begin{aligned} X_t &= \mu_X + 0.5122(X_{t-1} - \mu_X) - 0.4466(X_{t-2} - \mu_X) + 3.8234\times10^{-4} + \varepsilon_t \\ Y_t &= \mu_Y + 0.8821(Y_{t-1} - \mu_X) - 0.4520(Y_{t-2} - \mu_X) - 0.0294 + \varepsilon_t \end{aligned} \tag{4.4}$$

Apply the model above to predict the propagation of pest locations over time, and the results are compared with four data in the verification set, as shown in Fig 3:

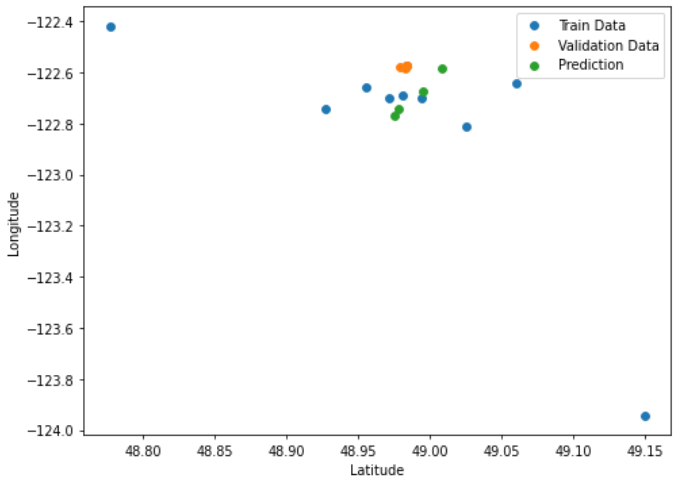


Fig 3: Diagram of AR Model Prediction Results

To sum up, it is relatively difficult to accurately predict the propagation of the pest due to the lack of training data and no obvious characteristics of the propagation position of pest over time. After analyzing Fig 3 we can find that AR model can better predict the propagation position of the pest and the absolute error of latitude prediction is less than 0.05, the absolute error of longitude prediction is less than 0.2.

5. Report Classification Model

From the analysis and discussion in the Section 4, we can see that the location distribution of the pest has obvious characteristics of centralized distribution. However, only using the latitude and longitude data in the report is not enough to make a correct judgment. As shown in Fig 4 (b), in the cluster area of the hive, there are many reports judged as negative ID.

In order to solve the problem of mixed positive and negative cases in the Fig above, we also need to use the image information contained in the report to assist in discrimination and improve the dimension of classification model variables, so as to separate the positive and negative cases apart. Based on the ideas above, we propose a Twice Discriminant Model based on logistic regression and image recognition.

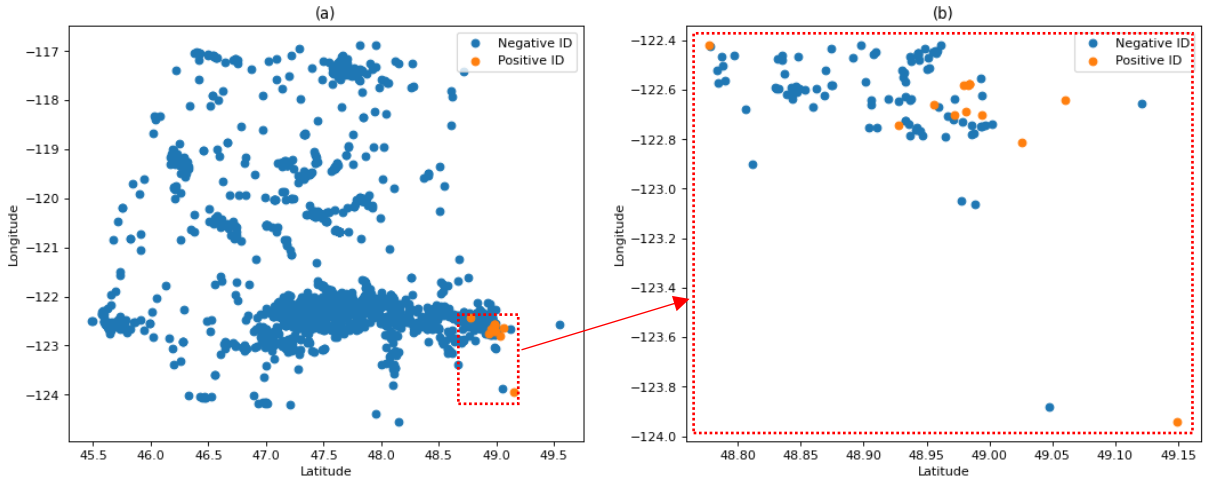


Fig 4: (a) Represents the reports' distribution in Washington State;
(b) represents the reports' distribution of a pest concentration area

5.1 Twice Discriminant Model

Since the pest are concentrated in a certain area of the map, we assume that the special region should be identified from the map based on the logistic regression model by using coordinate information without considering the image information in the report. In order to avoid the mixed positive and negative cases in Fig. 4(b) affecting the division of decision boundaries and improve the classification effect of the model, all negative cases near the area where the pest exists are deleted during the training of the logistic regression model (i.e., the first discriminant model).

In the second step of the Twice Discriminant Model, we focus on the classification of mixed regions in Fig. 4(b). The most straightforward solution is to add image information to the model through image recognition, making the previously inseparable problem easier to deal with after adding additional spatial dimensions.

5.1.1 Logistic Regression Model (First Discriminant)

For Lab Status from the i th report S_i ($S_i = 1$ means the value of this field is "Positive ID", $S_i = 0$ means the value of "Negative ID", We do not consider here the case where the field values are "Unverified" and "Unprocessed") and the corresponding latitude and longitude Lat_i,Lon_i, we define:

$$\begin{aligned} p\left(Lat_i, Lon_i\right) &= P(S_i = 1) \\ 1 - p\left(Lat_i, Lon_i\right) &= P(S_i = 0) \triangleq q\left(Lat_i, Lon_i\right) \end{aligned} \tag{5.1}$$

In Equation (5.1), $q(Lat_i, Lon_i)$ is the possibility of the corresponding report being included in the negative class by the model.

Consider the following logistic regression model [3]:

$$\log\frac{p(Lat_i,Lon_i)}{1-p(Lat_i,Lon_i)}=\beta_0+\beta_1 Lat_i+\beta_2 Lon_i+\beta_3 Lat_i Lon_i \tag{5.2}$$

The addition of quadratic term can enrich the kinds of decision boundary [3]. Before training the model, it is necessary to eliminate all negative cases in the mixed regions, and then standardize the latitude and longitude data to ensure the learning efficiency of model parameters:

$$Lat_i=\frac{Lat_i-\min\limits_i(Lat_i)}{\max\limits_i(Lat_i)-\min\limits_i(Lat_i)},\quad Lon_i=\frac{Lon_i-\min\limits_i(Lon_i)}{\max\limits_i(Lon_i)-\min\limits_i(Lon_i)} \tag{5.3}$$

The cost function that the logistic regression model needs to minimize is the following maximum likelihood function [3]:

$$\zeta(\beta_0,\beta_1,\beta_2,\beta_3)=\prod_{i:S_i=1}p(Lat_i,Lon_i)\prod_{i':S_{i'}=0}\left(1-p(Lat_{i'},Lon_{i'})\right) \tag{5.4}$$

For the output of the model $p(Lat_i,Lon_i)$, we can set a threshold value p_0 and get the classifier Y_i :

$$Y_i=\begin{cases}1, & p(Lat_i,Lon_i)\geq p_0\\ 0, & p(Lat_i,Lon_i)<p_0\end{cases} \tag{5.6}$$

Since there are very few positive class data, we use $Precision$ to measure the wellness of fitness of the model:

$$Precision=\frac{TP}{TP+FP}=\frac{\sum_i\left(Y_i=1|S_i=1\right)}{\sum_i\left(Y_i=1|S_i=1\right)+\sum_i\left(Y_i=1|S_i=0\right)} \tag{5.7}$$

When the model trains the parameters, the original data is divided into training set and verification set according to the ratio of 7:3. Due to the serious imbalance between the sample size of positive and negative data (there are only 14 positive cases and 2069 negative cases), the model is trained by 5-fold cross-validation. Table 2 shows the optimal model parameters and their statistical characteristics.

Table 2: Results of Logistic Regression Model

	Coefficient	Std Err	z-statistics	p-value	[0.025	0.975]
β_0	−71.3378	16.325	−4.370	< 0.0001	−103.333	−39.342
β_1	97.3164	22.178	4.388	< 0.0001	53.849	140.784
β_2	75.7076	25.234	3.000	0.003	26.249	125.166
β_3	−142.4303	41.905	−3.399	0.001	−224.563	−60.298

In Table 2, Coefficient represents the value of regression coefficient, and Std Err is the standard deviation of regression coefficient. Due to the sufficient sample size, we directly calculated the z statistic of each parameter and the corresponding p value by using the law of large numbers. The last two columns in the table represent the interval estimates of the parameters with a confidence of 97.5% [3].

Next, we examine the significance level of the model's parameters and the significance level of the model itself. For the parameter β_i $(i = 0,1,2,3)$, we consider the following hypothesis testing problem:

$$H_0 : \beta_i = 0 \quad \text{v.s} \quad H_a : \beta_i \neq 0 \tag{5.8}$$

It can be seen from Table 2 that the p values of the four parameters that is to be estimated, $\beta_0, \beta_1, \beta_2$ and β_3, are all very small. Under the condition of selecting significance level $\alpha = 0.05$, the hypothesis test is passed and H_0 is rejected, so the parameters $\beta_0, \beta_1, \beta_2$ and β_3 are considered to be significant [3].

For the significance of the model, we consider the following hypothesis testing problem:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{v.s} \quad H_a : \beta_i \neq 0 \quad \text{for some} \quad i \tag{5.9}$$

The χ^2 test is used for the model. The calculated p value of the model is $p = 9.285 \times 10^{-27} \ll \alpha = 0.05$. Therefore, we reject H_0 and consider the regression model to be significant [3]. Use the regression model to predict and the results are shown in Fig 5:

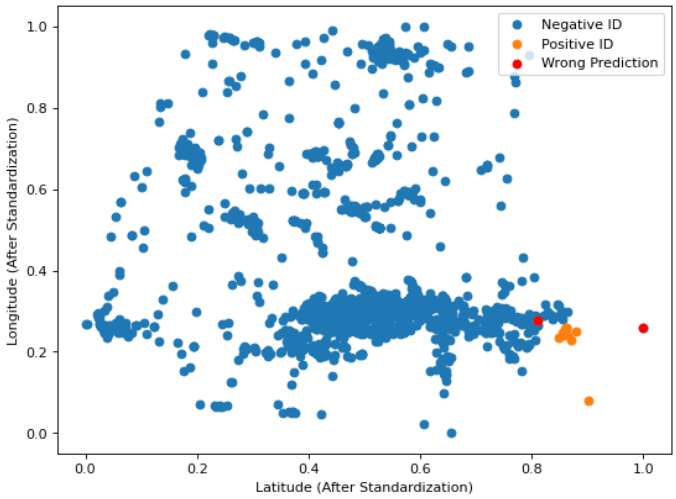


Fig 5: Result of Logical Regression Model Classification

As can be seen from Fig 5, the logistic regression model used for first discrimination can basically correctly identify the concentrated area of the pest. Among the 14 positive cases, one is wrongly classified as negative, and among all the negative cases, one is wrongly classified as positive. The accuracy rate of the calculation model using formula (5.7) is about 92.86%.

5.1.2 Image Recognition of Vespa Mandarinia

After using the logistic regression model to separate the pest concentration areas from the map, we need to deal with the confounding areas in Fig 4(b). With the help of picture information, we consider adding the variable of similarity between the pictures in the report and the Asian Giant Hornet pictures to help the discrimination in order to complete the classification task of the report. About the construction of image recognition model, we will explain in the following parts.

A. Data Preprocessing

The images submitted in the report were uniformly converted to 32×32-pixel RGB three-channel data. For reports lacking images, we filled them with pictures from the report closest to them in geographical location to ensure the integrity of training data.

B. Over-sampling and under-sampling

Due to the large difference in the proportion of positive and negative cases in the training data, we consider using the combination of over-sampling and under-sampling to process the training data. Oversampling strategy was used for the positive cases data (14 cases in total). We consider to rotate the images clockwise by 90°,180° and 270° respectively and flip them horizontally and vertically in order to increase the richness of the positive class training set through image transformation.

As for the negative data (a total of 3098 cases), we use the under-sampling strategy. We randomly divide the negative data into several parts, and ensure that the amount of each negative class data is equal to the number of positive class data after oversampling. The negative and positive class data are trained in multiple rounds. This allows the model to better distinguish the difference between the two classes.

C. Network structure and training methods

For the image recognition task, we choose the convolutional neural network (CNN) because of its better performance and we use the LeNet structure. LeNet Convolutional Neural Network is the first work of Convolutional Neural Network proposed by LeCun in 1998. LeNet reduces the parameters of the network by sharing the convolutional kernel. When counting the layers of the CNN, only the convolutional computing layer and the fully connected computing layer are generally counted, and other operations can be considered as the dependencies of the convolutional computing layer [4]. LeNet has five layers of network, the first two layers convolved, the last three layers with continuous full connection, the network structure is shown in Fig 6. The network takes 3 channels of RGB data as input, and finally outputs the network to two neurons. Softmax activation function is used to change the output value of the network into a probability distribution. The model was trained by the ADAM optimizer, and the learning rate was 0.001. The cost function used in model training is the cross-entropy loss function:

$$Cost = -\frac{1}{n}\sum_{i=1}^{n}\left[S_i \ln N_i + (1-S_i)\ln(1-N_i)\right] \tag{5.10}$$

where S_i is the true label of data and N_i is the ouput of the CNN.

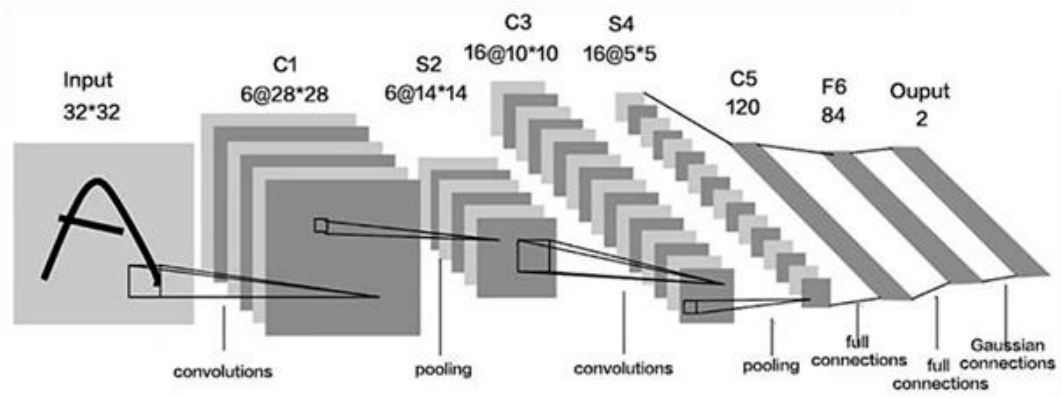


Fig 6: Schematic Diagram of LeNet network

D. Verification set and cross-validation

For each training, we divided the dataset into a training set and a validation at a ratio of 7:3. In addition, due to the small amount of data, in order to prevent the accidental division of training set and validation set from affecting the training results of the model, we used 5-fold cross-validation to evaluate the discriminant ability of the model.

E. Model prediction and evaluation

Since the output N_i of the network is a probability distribution similar to the logistic regression model, we choose a threshold p_1 and define the prediction result of the network as:

$$NS_i = \begin{cases} 1, & N_i \geq p_1 \\ 0, & N_i < p_1 \end{cases} \tag{5.11}$$

Considering the small number of positive cases, we use the following two indexes of accuracy acc and $F1-Score$ to measure the performance of the network [3]:

$$acc = 1 - \frac{1}{n}\sum_{i=1}^{n}\left|S_i - NS_i\right| \tag{5.12}$$

$$F1-Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5.13}$$

$Precision$ and $Recall$ represent the precision rate and recall rate, respectively:

$$\begin{aligned} Precision &= \frac{TP}{TP+FP} = \frac{\sum_i \left(NS_i = 1 \middle| S_i = 1\right)}{\sum_i \left(NS_i = 1 \middle| S_i = 1\right) + \sum_i \left(NS_i = 1 \middle| S_i = 0\right)} \\ Recall &= \frac{TP}{TP+FN} = \frac{\sum_i \left(NS_i = 1 \middle| S_i = 1\right)}{\sum_i \left(NS_i = 1 \middle| S_i = 1\right) + \sum_i \left(NS_i = 0 \middle| S_i = 1\right)} \end{aligned} \tag{5.14}$$

The trained model is used to identify the confounding area in Fig 4(b), and the classification result is shown in Fig 7:

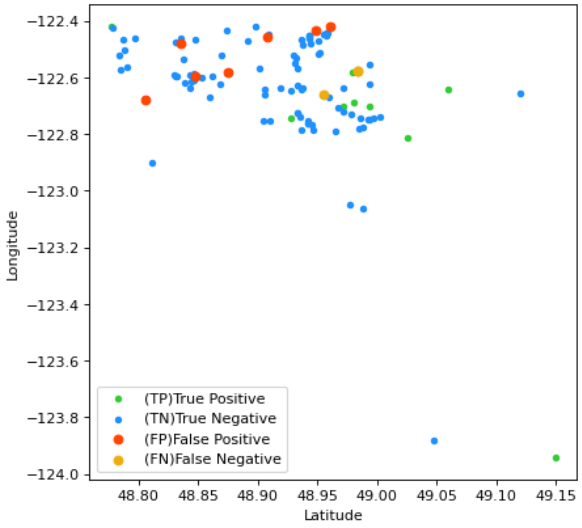


Fig 7: Discrimination results of image recognition model in positive and negative mixed areas

The image recognition model in Fig 4(b) has a recognition accuracy rate of $acc \approx 91.07\%$, a model accuracy rate of 61.11%, a recall rate of 78.57%, and a F1-Score value of 0.6875. It's safe to say the image recognition model performs well.

5.1.3 Compound Discriminant Model (Second Discriminant)

In the image recognition model, LeNet help complete the classification and recognition tasks of mixed positive and negative areas, but we can't just use image recognition to complete the report classification. For example, we apply the image recognition model to the entire data set and classify the submitted reports. The results are as shown in Fig 8:

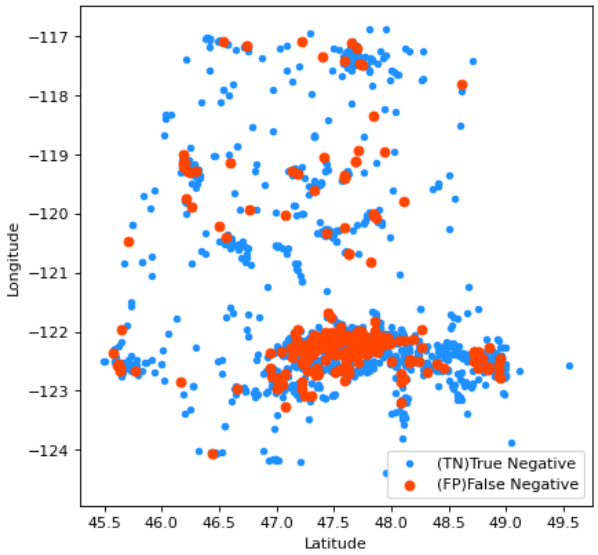


Fig 8: The use of pattern recognition to identify reports across the state of Washington

It can be seen from Fig 8 that there are a lot of misclassifications in the image recognition model. This is because among the pictures uploaded, many are similar to the pest (other kinds of bee species), and the image recognition model intelligently separates the image whose features are far from the pest.

In order to make up for the deficiencies of the logistic regression model in section 5.1.1 and the image recognition model in section 5.1.2, we combine the two to obtain a Twice Discriminant Model. The core idea of this model is to use the logistic regression model to determine whether the geographic location is in the pest cluster area or not, and then use the image recognition information for the second-time discrimination.

In the logistic regression model, we get $p(Lat_i, Lon_i)$ to measure the probability that the sighting report is located in pest concentration area (that is, the report is classified as a positive sample). In the image recognition model, we get N_i to measure the similarity between the uploaded image information and the pest. The value of N_i represents probability, and the value is between [0,1], but most of the N_i is distributed around the threshold p_1. For this reason, we consider standardizing the probability N_i:

$$N_i = \frac{N_i - \min_i(N_i)}{\max_i(N_i) - \min_i(N_i)} \tag{5.15}$$

Now, draw with $p(Lat_i, Lon_i)$ and N_i as the horizontal and vertical coordinates, and the distribution of the two features is shown in Fig 9. It can be seen from Fig 9 that the two types of positive and negative data are not mixed under the two characteristics and can be separated. Therefore, the feature data is sent to the new logistic regression classifier:

Now, draw with $p(Lat_i, Lon_i)$ and N_i as the horizontal and vertical coordinates, and the distribution of the two features is shown in Fig 9. It can be seen from Fig 9 that the two types of positive and negative data are not mixed under the two characteristics and can be separated. Therefore, the feature data is sent to the new logistic regression classifier [3]:

$$\log \frac{p'(p(Lat_i, Lon_i), N_i)}{1 - p'(p(Lat_i, Lon_i), N_i)} = \theta_0 + \theta_1 p(Lat_i, Lon_i) + \theta_2 N_i + \theta_3 p(Lat_i, Lon_i) N_i \tag{5.16}$$

In formula (5.16), $p'(p(Lat_i, Lon_i), N_i)$ represents the logistic regression probability used in the twice discriminant model, where $\theta_0, \theta_1, \theta_2, \theta_3$ are the parameters to be learned. The training of the model is the same as that of equation (5.2), and the cost function also uses the maximum likelihood function of (5.4). For the output $p'(p(Lat_i, Lon_i), N_i)$, set the threshold p_2 to obtain the classifier C_i:

$$C_i = \begin{cases} 1, & p'\left(p(Lat_i, Lon_i), N_i\right) \geq p_2 \\ 0, & p'\left(p(Lat_i, Lon_i), N_i\right) < p_2 \end{cases} \tag{5.17}$$

Similarly, since there are very few positive data, we use F1-Score to measure the quality of the discriminant model:

$$F1 - Score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{5.18}$$

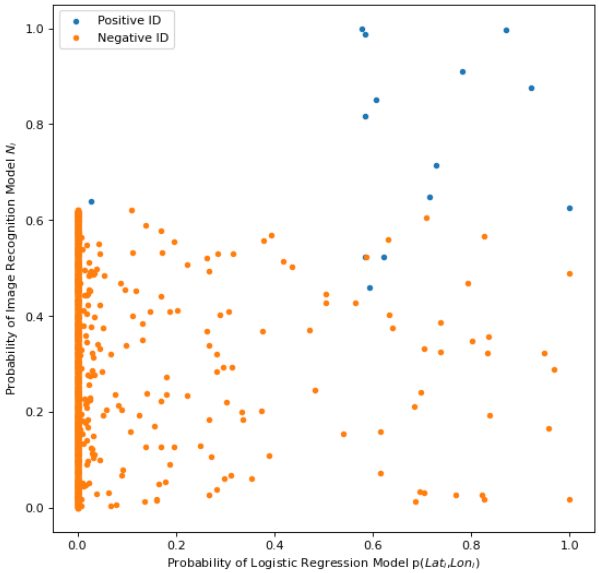


Fig 9: Distribution of positive and negative samples under new features

The model training method is the same as that in Section 5.1.1. Table 3 is the best model parameters and their statistical characteristics that are solved. They will be used to analyze the significance of the model and judge whether the model is reasonable.

Table 3: The solution result of the twice discriminant model

	Coefficient	Std Err	z-statistics	p-value	[0.025	0.975]
θ_0	−18.4128	9.110	−2.021	0.043	−36.268	−0.558
θ_1	9.1803	12.504	0.734	0.463	−15.327	33.688
θ_2	22.2703	15.765	1.413	0.158	−46.196	40.222
θ_3	−2.9868	22.046	−0.135	0.892	−224.563	−60.298

Considering the test with a significance level of α=0.05, a simple analysis of Table 3 finds that, except for θ_0, the p-values of all other model parameters are greater than 0.05. Therefore, for the following hypothesis testing problem, we do not have enough evidence to reject the null hypothesis:

$$H_0 : \theta_i = 0 \quad \text{v.s} \quad H_a : \theta_i \neq 0 \quad \text{for} \quad i = 1,2,3 \tag{5.19}$$

Among the parameters $\theta_1, \theta_2, \theta_3$, the largest p value is θ_3. Therefore, we consider removing the quadratic term from the model (5.16) and modifying the regression model of twice discrimination to [3]:

$$\log \frac{p'(p(Lat_i, Lon_i), N_i)}{1 - p'(p(Lat_i, Lon_i), N_i)} = \theta_0 + \theta_1 p(Lat_i, Lon_i) + \theta_2 N_i \tag{5.20}$$

Using the results in Table 4 to analyze the statistical significance of the model, we can see that the p-values of the three parameters $\theta_0, \theta_1,$ and θ_2 of the current model are all much less than 0.05, so we have enough statistical evidence to reject the null hypothesis that the parameter is 0. The last two columns of Table 4 are the interval estimates of the three parameters with a confidence of 97.5%.

For the significance test of the model, consider the following hypothesis test problem:

$$H_0 : \theta_1 = \theta_2 = 0 \quad \text{v.s} \quad H_a : \theta_i \neq 0 \quad \text{for some} \quad i \tag{5.21}$$

Using the χ^2 test on the model, the p-value of the model is$p = 1.414 \times 10^{-29} \ll \alpha = 0.05$, so H_0 is rejected. We think the twice discriminant model is significant. Using the above model for prediction, the result is shown in Fig 10.

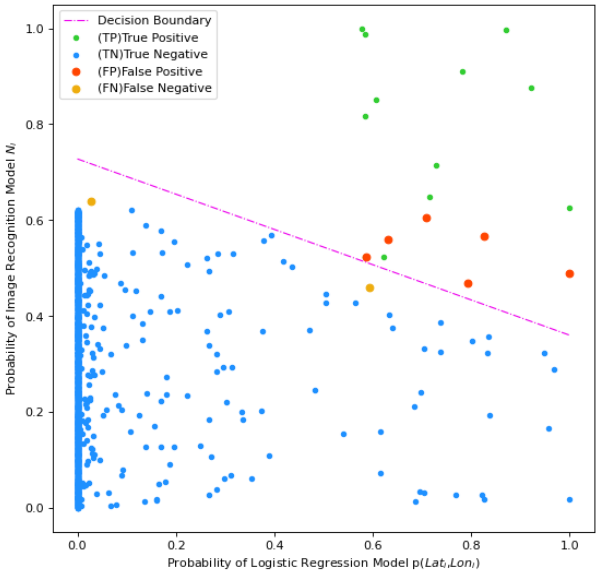


Fig 10: Classification results of the twice discriminant model

From Fig 10, the accuracy rate of the secondary discriminant model classification is 66.67%, the recall rate is 85.71%, and the F1-Score value of the model is 0.75. The model can better complete the classification task of the report.

5.2 Evaluation of the model's accuracy

In this section we discuss the accuracy of the twice discriminant model. There are two possibilities for the secondary discriminant model to report classification errors: judge the positive class as a negative class (false negative) or judge the negative class as a positive class (false positive). Based on the twice discriminant model, we use Bayesian estimation to establish a model about the possibility of classification errors.

Assume that the discriminant model is independent of the reported classification results and the true state of the report (Positive ID or Negative ID). Then the probability P_i^{error} of the article i report being misclassified can be calculated by the full probability formula:

$$P_i^{error} = P\left(C_i = 1 \middle| S_i = 0\right) + P\left(C_i = 0 \middle| S_i = 1\right) \tag{5.22}$$

In formula (5.22), $S_i$is the true state of the report, and $C_i$is the classification result of the secondary discriminant classifier. Because it is assumed that the values of C_i and S_i are independent of each other, the above two conditional probabilities can be calculated by the following formula:

$$P\left(C_i = 1 \middle| S_i = 0\right) = \frac{P\left(S_i = 0, C_i = 1\right)}{P\left(S_i = 0\right)}, \quad P\left(C_i = 0 \middle| S_i = 1\right) = \frac{P\left(S_i = 1, C_i = 0\right)}{P\left(S_i = 1\right)} \tag{5.23}$$

For a single indicator i, it is difficult to calculate the above probabilities and there are too many uncertain factors. We directly use the existing data to estimate these probabilities. For $P(S_i = 0, C_i = 1)$ and $P(S_i = 1, C_i = 0)$, we have:

$$P\left(S_i = 0, C_i = 1\right) = \frac{FP}{TN + FP}, \quad P\left(S_i = 1, C_i = 0\right) = \frac{FN}{TP + FN} \tag{5.24}$$

In summary, the model for predicting the probability of misclassification is:

$$P_i^{error} = \frac{FP}{\mu\left(TN + FP\right)} + \frac{FN}{\left(1 - \mu\right)\left(TP + FN\right)} \tag{5.25}$$

Where μ represents the probability$P(S_i = 0)$, which is a hyperparameter determined by the environment in the model. For the value of μ, according to the law of large numbers, we directly use the frequency of events in all data to represent the probability, and get:

$$\mu = \frac{n_{pos}}{n_{pos} + n_{neg}} \approx 0.6721\%, \quad 1 - \mu = \frac{n_{neg}}{n_{pos} + n_{neg}} \approx 99.3279\% \tag{5.26}$$

From the central limit theorem, we can get the distribution of μ:

$$\mu \sim N(\mu_0, \frac{\mu_0(1 - \mu_0)}{n}) \tag{5.27}$$

in which $\mu_0 = 0.006721$, and the existing sample size $n = 2083$.

From the formula (5.27), the 95% confidence interval of μ is $\mu \in [0.006644,,0.006798]$. The results of the twice discriminant model is used to calculate FP, FN, TN, TP. Based on the existing data, the model error prediction is obtained. The possibilities are:

$$P^{error} = \frac{6}{0.006721 \cdot \left(2063 + 6\right)} + \frac{2}{0.9932791 \cdot \left(12 + 2\right)} \approx 57.53\% \tag{5.29}$$

Among them, the false positive rate is $P(C_i = 1|S_i = 0) = 43.15\%$, while the false negative rate is $P(C_i = 0|S_i = 1) = 14.38\%$, which represents a low level.

6. Prioritizing Investigation of the Reports

In order to distinguish the report with a higher potential to become a true sample in the report, we have given the probability of combining geographical location and image information in the output of the secondary discriminant model to classify the report into a positive category. Therefore, we can arrange the report processing order according to the discriminant model output $p'(p(Lat_i, Lon_i), N_i)$ from high to offset, but because logistic regression almost transforms the positive probability to 1, and the negative probability to 0, so the direct use of probability for sorting will lose a lot of information about the feature quantities $p(Lat_i, Lon_i)$ and N_i.

As can be seen from Fig 10, above the decision boundary, near the upper right corner of the image (that is, $p(Lat_i, Lon_i)$, the direction where N_i tends to 1), the classifier divides the report into positive categories; and below the decision boundary , Near the lower left corner of the image (that is, the direction where $p(Lat_i, Lon_i), N_i$ tends to 0), the classifier divides the report into negative categories. For the feature coordinate reported in the i-th article, we measure the distance d_i from the coordinate (1,1):

$$d_i = w_1\left(p(Lat_i, Lon_i) - 1\right)^2 + w_2\left(N_i - 1\right)^2 \tag{6.1}$$

In formula (6.1), the weight coefficients w_1, w_2 represent the weight of the distance d_i with respect to the two features. The linear coefficient of the decision boundary of the secondary discriminant model carries the weight of the two features. We can directly make the weights w_1, w_2 equal to θ_1, θ_2, and we get:

$$d_i = \theta_1\left(p(Lat_i, Lon_i) - 1\right)^2 + \theta_2\left(N_i - 1\right)^2 \tag{6.2}$$

If d equals to 0.5, 2, 4, 8 respectively, the contour map of the model (6.2) is shown in Fig 11.

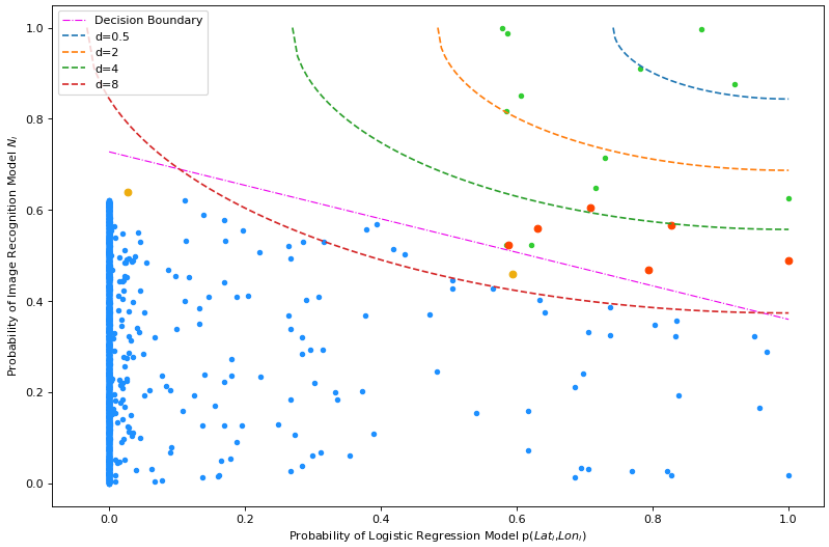


Fig 11: Contour map of report prioritization model

Table 4: Top 20 priority data

Order	d_i	Lab Status	Order	d_i	Lab Status
1	0.125163	Positive ID	11	3.825246	Negative ID
2	0.357567	Positive ID	12	4.072676	Negative ID
3	0.524948	Positive ID	13	4.978837	Negative ID
4	1.293426	Positive ID	14	5.337458	Negative ID
5	1.338132	Positive ID	15	5.727756	Positive ID
6	1.617445	Positive ID	16	5.907368	Negative ID
7	1.969349	Positive ID	17	5.933695	Positive ID
8	2.221676	Positive ID	18	6.102558	Negative ID
9	2.873424	Positive ID	19	6.537266	Negative ID
10	3.144514	Positive ID	20	6.901218	Negative ID

7. Update Model with More Data

The performance of the three models proposed in Section 5 (logistic regression first discriminant model, image recognition model and logistic regression twice discriminant model) are all dependent on the data set. Adding additional new data can help to increase the probability that the discriminant model correctly completes the report classification. On the one hand, we consider the online-offline operation mode of the model in practical tasks. While the model makes predictive analysis on the existing data, the updated data is used in the background for model training and adapting to obtain a predictive model that is more in line with the existing data. At the same time, the model accuracy threshold is set to compare the pros and cons of the new and old models. When the value exceeds the threshold, the new model is considered to have better performance on the existing data, and the model can be updated at this time.

Below we discuss methods used for updating the three models in detail.

- Logistic Regression Model (first discrimination)

This model is mainly used to identify concentrated areas of honeycombs. Negative data are randomly scattered throughout Washington State without obvious characteristics. The classification of the model mainly depends on the distribution characteristics of positive data, and the number of positive data is much smaller than that of negative one. Additional new positive data can significantly affect the decision-making of the model and improve the classifying effect of the model.

We choose to adjust the ratio of positive and negative samples in the input model data to find the most suitable positive sample concentration for the model's prediction and get the best prediction accuracy. Through tests, we have concluded that when the total number of positive classes in the new data set reaches 10, the accuracy of the model prediction reaches a high level at this time, and the new data can be added to the model to retrain.

When the distribution of the positive data does not change too much, the parameters of the model can be updated directly; if the distribution of the positive data is significantly different over time, for example, when the longitude and latitude data in the report changes significantly (from the Northwestern United States to the Eastern United States), it ought to consider making adaptive adjustments to the logistic regression model, such as adding higher-order terms about latitude and longitude to increase the richness of the model. Corresponding regular term should be added into the loss function to alleviate the over-fitting when additional data is put into the model. What's more important is that the significance test of the model parameters is essential for the model itself.

- Image Recognition Model

The amount of data and image quality are key factors that affect the accuracy of image recognition. Existing models can only distinguish useless pictures (such as irrelevant images or other noisy images) that are far from the pest, while for other bees' images, the model's distinguishing ability is relatively limited.

Adopting the same test method in the first logistic regression discriminant model and adjusting the proportion of positive and negative samples in the input, we conclude that when the total number of positive images in the new data set reaches 5, and the proportion of positive and negative samples is greater than or equal to 0.68%, it's safe to add new image data to the model to retrain.

The method of combining over-sampling and under-sampling is still used to expand the positive data set and solve the problem of imbalance between positive and negative classes. In addition, in order to improve the prediction accuracy of the model, a standard pest images can be manually collected as positive input data for training to obtain more reliable results.

- Logistic Regression Model (second discrimination)

The training data used by the secondary discriminant model is related to the results of the above two models. Therefore, we believe that whenever any of the above models is updated, this model should be retrained. Again, it is necessary to perform hypothesis testing on the significance of the updated model parameters and the significance of the model. This is the key to ensure the correctness and reliability of the model.

- Notes Classification Model

We have tried to find a reliable classification model on the current data set by studying the word frequency and sentiment analysis of notes attached to the report in the data set. However, due to the large deviation of the actual number of valid positive and negative samples (much more than the previous three models), and the effective data in the word frequency and emotion analysis does not show obvious trend, jumbled quality of data forces us to give up this kind of information processing. If the text content in the new report can add information with obvious tendency (for example, the degree of confidence in sighting), we can use relevant methods to obtain the sample features contained in the report comments to help achieve more effective classification.

8. Vespa Mandarinia in Washington State

Finally, we use the secondary discriminant model to examine whether there is evidence that Washington State has eliminated this pest. Some of the reports in the data set are under "Unprocessed" and "Unverified" status. Use the model to predict on these data and the classification results are shown in Fig 12:

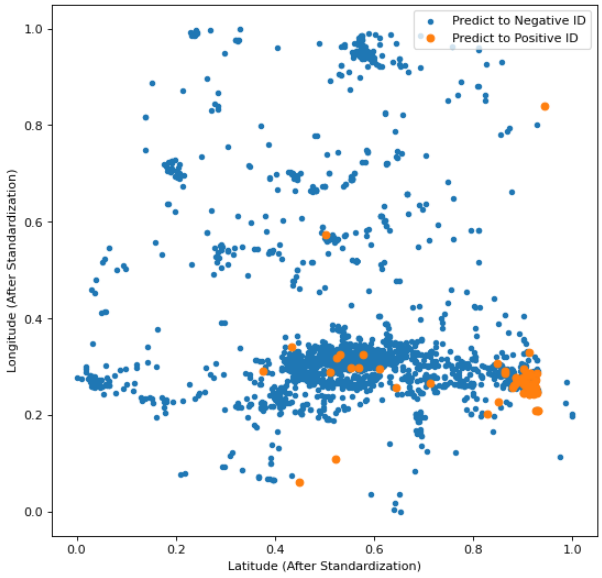


Fig 12: Classification and prediction of Unprocessed and Unverified reports by quadratic discriminant model

From Fig 12, we can find that most of the positive predictions appear in the concentrated area of the pest, and a few positive predictions are randomly scattered in other areas of Washington State. Combined with the analysis of Fig 10 (decision boundary of quadratic discriminant model), the output characteristic $p(Lat_i, Lon_i)$ of the logistic regression model tends to zero in the place that is far away from the pest concentration area, and only when the place is very close to the pest concentration area will the value of characteristic $p(Lat_i, Lon_i)$ tends to 1.

On the side that tends to zero, the decision boundary of $p(Lat_i, Lon_i)$ requires a high value of N_i so that the model will judge the report as positive. Therefore, the positive class far away from the pest concentrated area in Fig 12 is classified as positive because of their images' similarity to that of the pest. In the concentrated area, the model does not require high similarity of image information, and the model mainly relies on the geographic location information uploaded in the report to make judgments.

From the prediction of the number of positive species, the nature of the concentrated distribution of positive cases, and the high false positive rate (43.15%) of the twice discriminant model, we believe that this pest has not been widespread in Washington state, which has been controlled and eliminated.

9. Model Evaluation

9.1 Error Analysis

- The model mainly investigates the geographical location information and image information of the report. As for the time of the report sending, the text information in the report is not processed too much, and the loss of useful information will increase the uncertainty of the model.
- The positive and negative samples of the training data used in the image recognition model are not balanced, and the image quality is low, so the model cannot correctly distinguish the pictures of the pest and other bees, which is one of the reasons for the high false positive rate of the Twice Discriminant Model.
- When we use the Twice Discriminant Model for report classification, we do not consider the distribution characteristics of pest over time. The expansion and movement of pest will greatly change the decision boundary of the discriminant model.

9.2 Improvement

- Analyze the words frequency and meaning of notes submitted in the report to mine useful information. After quantifying it as indicators, it can be added into the training features of the Twice Discriminant Model. Higher dimension of feature can make the separation of positive and negative classes more considerable and reliable.
- Take the influence of the evolution of pest propagation with time on the model into consideration. We can use the transmission model to treat the positive data, the positive data predicted by the autoregressive model and the negative data as confirmed patients, virus carriers and susceptible population respectively. The study of the spreading characteristics of the pest is helpful to control the spread of this pest.
- Collect more positive data images and provide them to the image recognition model for learning and training to improve the possibility of correctly identifying the pest.
- When new reports are classified (especially positive data), update and retrain the model in time, so that the model can obtain the latest information about the development of pests in Washington state.

10. Reference

[1] Ross, Sheldon M . "Introduction to Probability Models, Ninth Edition." Academic Press, Inc. 2006.
[2] Quinn, E. J. Hannanb. G. . "The Determination of the Order of an Autoregression." Journal of the Royal Statistical Society 41.2(1979):190-195.
[3] James, Gareth , et al. An Introduction to Statistical Learning. Springer New York, 2013.
[4] Lecun, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE (1998).

11.Memorandum

From: MCM Team#2123737

To: Washington State Department of Agriculture

Date: February 8, 2021

Dear Washington State Department of Agriculture,

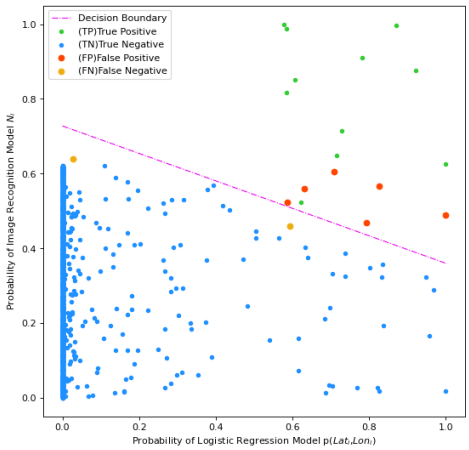
To deal with numerous public reports data about Vespa mandarinia sightings and optimize the strategies to prioritize the reports for additional investigation, making full use of the limited resources, we analyzed the reports data provided by four models we designed. These models have been trained and tested and are fully able to solve the main questions raised, besides, they perform a good ability of adaptation for new reports data.

The analysis and results of the models are as follows:

First of all, based on the geographic information in the reports, we use Time Series Analysis Method to Fig out the distribution of latitudes and longitudes and their numerical characteristics. After discovering the weak temporality of the coordinates, we utilize Autoregression Model to make prediction on the next possible location of the pest and get a very promising results with an absolute error less than $(0.05,0.2)$.

Most of our work is devoted to creating a model that is used to analyze the reports to predict the likelihood of a mistaken public classification. Considering that the major features of the reports are buried in the data of coordinates and images, we provide a Twice Discriminant Model on the foundation of Logistic Regression and Image Recognition methods. When putting some reports into the model, they will first be addressed with the Logistic Regression Model to locate the most possible area of the positive reports because of their spatial distribution characteristic.

Then to separate the negative reports that mixed in the positive ones in the first model, we take the images attached to the reports as a secondary discrimination. We choose the LeNet Convolutional Neural Network to extract features of the images and get a measure N_i of the similarity between the image from the report and the image of a typical Vespa mandarinia. Tests have been run to prove that lack of either of the two features of the reports (information contained in coordinates and images) will make the prediction results unconvincing. Therefore, we still use Logistic Regression method to combine the two discrimination models mentioned above to make sure the integrity of the reports' information and the classification result is showed below. What's worth to be mentioned is that we've carried out hypothesis testing and interval estimation to guarantee the effectiveness of the model's prediction and statistical significances of the parameters.



To optimize the investigation of the reports and reduce resources that should not be wasted, we propose a ranking model based on the results of former problems. To be specific, a kind of measurement is set in which we offer the two features of the report's different weights according to the Twice Discriminant Model's decision boundary. By doing so, when new reports are added into the model, we will be able to give a list of report Ids (as is shown in the table below)which are most probably to be the positive reports, thus leading to targeted and efficient strategies.

Order	d_i	Lab Status
1	0.125163	Positive ID
2	0.357567	Positive ID
3	0.524948	Positive ID
4	1.293426	Positive ID
5	1.338132	Positive ID

The values d_i in the table is able to measure the specific report's distance from the positive class. (how much importance it should be exerted to the report). According to the positive cases in Washington State that are predicted by the model and its false positive rate, it's safe to come to the conclusion that the pest has been controlled and eradicated in Washington State.

In the end, we offer you some suggestions on how to update the models with additional reports data. (1) Due to the dependence of the model's accuracy on the number of positive reports, the performance of our models will be better when given reports with more obvious positive features. For example, update the model with reports with more accurate and explicit images of Vespa mandarinia. (2) What's more, when the coordinates distribution of positive reports doesn't have great change, you only need to update the model's parameters and when the coordinates distribution have some dramatic changes like geographic migration, you can add higher order items about latitude and longitude to realize the adaptation. (3) If provided with more accurate and distinguished notes information, you can use the model we mentioned in Section 7 to make full use of this kind of text data.

Thank you for your precious time on reading this. We truly wish that the results of our models and suggestions above will be of help for you.