Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2114434

Beat the Vespa mandariniz: Interpretation and Exploration of Sighting Reports

Summary

In September 2019, a group of Vespa mandarinia was discovered on Vancouver Island in British Columbia, Canada, causing heated discussion and panic. Since then, a large number of sightings have also taken place in neighboring Washington State, but only a few of them have been confirmed as Vespa mandarinia sightings. The reports provided by the public contain detection date, notes, pictures, geographic information and global ID. With limited resources, it is vital for Washington State Department of Agriculture to effectively analyze and evaluate these report data, which is helpful to judge the status of the report and take corresponding preventive measures.

To address this problem, we make statistics, drop the missing and delete invalid data. According to the adhesion between different indicators, we get secondary labels to explore more indicators that may contribute to judging the results of the report. The indicators are expressed in equations. We construct a most likely positive scoring model to give a score based on the information provided by citizens. We increase the richness of the final result—re-score all the positive, negative and unverified reports to control them in an interval. The closer the score to the right end of the interval, the greater the possibility of the report to be positive, which means the tendency of more Vespa mandarinia in the future. Therefore, even unverified reports, they have a score that reflects the trend.

On the processing of the data in the report, we use the K-means text clustering method based on TF-IDF algorithm to classify the lab comments of reports. We obtain some of the most confusing Vespa mandarinia insects and divided the images of negative reports into some categories. We continue to build a CNN-based image classification model based on the divided images as training sets. As a result, the unverified image set is divided and lead to a relatively exact result. For the construction of indicators, we take five first-level indicators—detection time, geographic information, data integrity, figure and notes. The detection time is evaluated by the number of reports received and the number of positive reports received per week. The notes is analyzed by special word and validity index based on semantic analysis result. Ultimately, we build a PCA-based Most likely positive scoring model with statistical analysis of each indicator. Moreover, based on ARIMA model, we make a double fitting of the scores and the weekly average number of reports to predict the occurrence of in Washington State. When results are stable at 0 for a certain period, we can conclude that it was eradicated. In addition, for the biological properties of the Vespa mandarinia, we establish a Logistic model to explore its variation law of number over time, aiming to determine the update frequency of the report and model.

We find some insightful conclusions based on our results: (1) the number of positive reports is proportional to the number of reports received, both of which fluctuate with the quarter; (2) Vespa mandarinia's latitude and longitude vary in a small range; (3) positive reports are mostly occurd during Vespa mandarinia's reproduction; (4) several kinds of species are most likely to be confused with Vespa mandarinia by the semantic analysis of lab comments; (5) the high-frequency words and words categories are obtained using the semantic analysis of the notes; (6) the importance of the report integrity to the outcome is obtained by PCA. It is also beneficial to improve the neatness of the figure data format, the clarity and accessibility of the images.

Based on the above analysis, we provide some strategies and suggestions on report processing, collection and public education to Washington State Department of Agriculture, which are fully elaborated in the following memorandum.

Keywords: semantic analysis, TF-IDF, K-means, PCA, ARIMA, Logistic, CNN

Memorandum

To: Washington State Department of Agriculture

From: Team # 2114434

Date: 8 February 2021

Subject: Memo to Washington Department of Agriculture

Regarding the sightings of Vespa mandarinia in Washington State, we have done a series of studies based on the data and information you provided. We are of great honor to have this opportunity to present our research results to you. Considering this incident is related to casualties, we sincerely hope that you can consider our advice, our memo mainly contains three modules: 1. How can the collection and feedback of memos be more effective. 2. How to conduct a reasonable result analysis of the submitted witness report. 3. Suggested guidance for pest control.

1. Collect and confirm information:

In the most likely positive scoring model we constructed, we find that the indicator of data integrity occupies a larger weight by using PCA, and it has a strong guiding role in the judgment of the results. We have counted all the information you provided. After deleting the invalid information, only 60 of the 2327 unverified samples contain valid image information. Therefore, the website should set the image upload option in the column of witnessing report submission as mandatory and attach There are warnings to remind the public not to upload other types of pictures and files in other formats, because we found that identifying multiple formats takes more time than identifying a single format.

In addition, it can be found that the detection date and the submission date are out of sync. Excluding the detection date before 2020, there are still dates that are out of sync between the two. We think it is necessary to make the two completely synchronized, which is more effective for determining the results. There will be no hysteresis error. In order to avoid Vespa mandarinia to the greatest extent possible infestation, the review of laboratory personnel needs to be synchronized with the submission date as much as possible.

For the notes column in the witness report, after our semantic analysis of notes, the main keywords we extract involve size, color, where, and species, these four categories. These 4 categories are very helpful to the expert's judgment. You can write these prompt words in the prompt column of notes to remind the public to write a message in the above range.

2. Analysis of the results of the report:

We found that the number of weekly report submissions is directly proportional to the number of weekly positive samples. The two change with seasonal volatility. Positive samples mainly appear in the second and third quarters, and concentrate in the 4 months of 7, 8, 9, 10. The numbers of all samples and positive samples are both high, and experts need to focus on these months and quarters. However, comparing 2019 with 2020, it is found that the value of the (number of positive samples/total number of samples in the week) is significantly lower when more positive samples appear, which indicates that the proportion of target samples is significantly reduced. Only a single method of expert review will increase the error rate, and it is best to introduce a computer system for assistance.

The CNN model that we built with a three-layer convolution kernel pooling layer is trained and classified under the expanded data set, and its accuracy can reach 85%. Therefore, the website can set up an automatic judgment function based on this model, and continuously put the pictures submitted by the public into the training set for learning, thereby improving the accuracy of the model.The model gives the possibility of judging a sample as positive. In our model, when the probability is higher than 70%, and it will be sent to the expert for a second judgment. It increases the accuracy of judgment and reduces manpower input.

When the submitted report meets the recommendations listed in 1, and becomes a standardized report, we recommend using the Most Likely Positive Scoring Model proposed in this essay to score. When the comprehensive score meets the score interval of the positive sample, a second manual judgment is performed. Therefore, you can create a website based on this evaluation model to facilitate data management.

We can use the modified evaluation model to detect the dynamics of Vespa mandarinia.When the curve moves closer to the positive sample range, you can take some prevention and control measures in advance, or use the ARIMA model in essay to make real-time predictions. Based on this result, you can take some actions in advance to prevent the occurrence of pests.

3. Measures to prevent Vespa mandarinia in advance:

After semantic analysis and clustering of expert feedback, we find that people tend to confuse Vespa mandarinia with the other four types of insects: golden digger wasp, horntail, cicada, and bee. I suggest you organize some training and education activities to teach people living in areas with a high incidence of pests can correctly distinguish these types of insects to reduce the number of false sightings, and they can also take precautions in advance. Regarding the high-prone areas of pests, we estimate the possible circular areas in the first small model above. You can also implement some measures in advance based on the biological characteristics of Vespa mandarinia and the month of reporting.

We are very grateful for this opportunity to give you some advice on the Vespa mandarinia report incident, and we believe our suggestions can help you solve some problems. If you want to know more, please feel free to contact us.

Best,

Team # 2114434

Contents

1 Assumptions and Notations

1.1 Assumptions

In order to simplify our modelling, we make the following assumptions:

Assumption 1. The data provided is true and reliable. We can only use the provided data files to solve the problem, and our analysis is only valid when the data is true and reliable.

Assumption 2. Since 2010, the ecosystem in neighbouring Washington State has been kept stable, namely, the number and proportion of organisms have always been relatively stable.

Assumption 3. The COVID-19 has not affected the spread of Vespa mandarinia, that is, the COVID-19 has no impact on Vespa mandarinia itself, and its impact on human behavior does not exert influence on Vespa mandarinia.

Assumption 4. Vespa mandarinia is independent individuals, and its spread is not affected by each other.

Assumption 5. No other related random variable changes affect the time series analysis.

Assumption 6. people will choose to upload information when they see a suspected species, and provide information like notes and images as much as possible.

1.2 Notations

In this work, we use the nomenclature in Table 1 in the model construction. Other none-frequent-used symbols will be introduced once they are used.

Table 1 Notations used in this literature

Symbol	Definition
TF	Term Frequency
IDF	Inverse document frequency
$\mathrm{Lat}[j]$	Latitude of the j-th data
$\mathrm{Lon}[j]$	Longitude of the j-th data
Notes_if_valid	Whether it is valid information in notes
notes_if_special	Whether there are special words in notes
Geo	Score of geographic information
Data_complete	The integrity of the data
Figure_result	Score of image information
week_total_amount	The total number of reports received in a week
week_positive_amount	The total number of positive reports received in a week

2 Data Processing and Analysis

As shown in the figure 1 below, we have processed the data in the table files first. As for figures, we make the data cleaning before model modelling later.

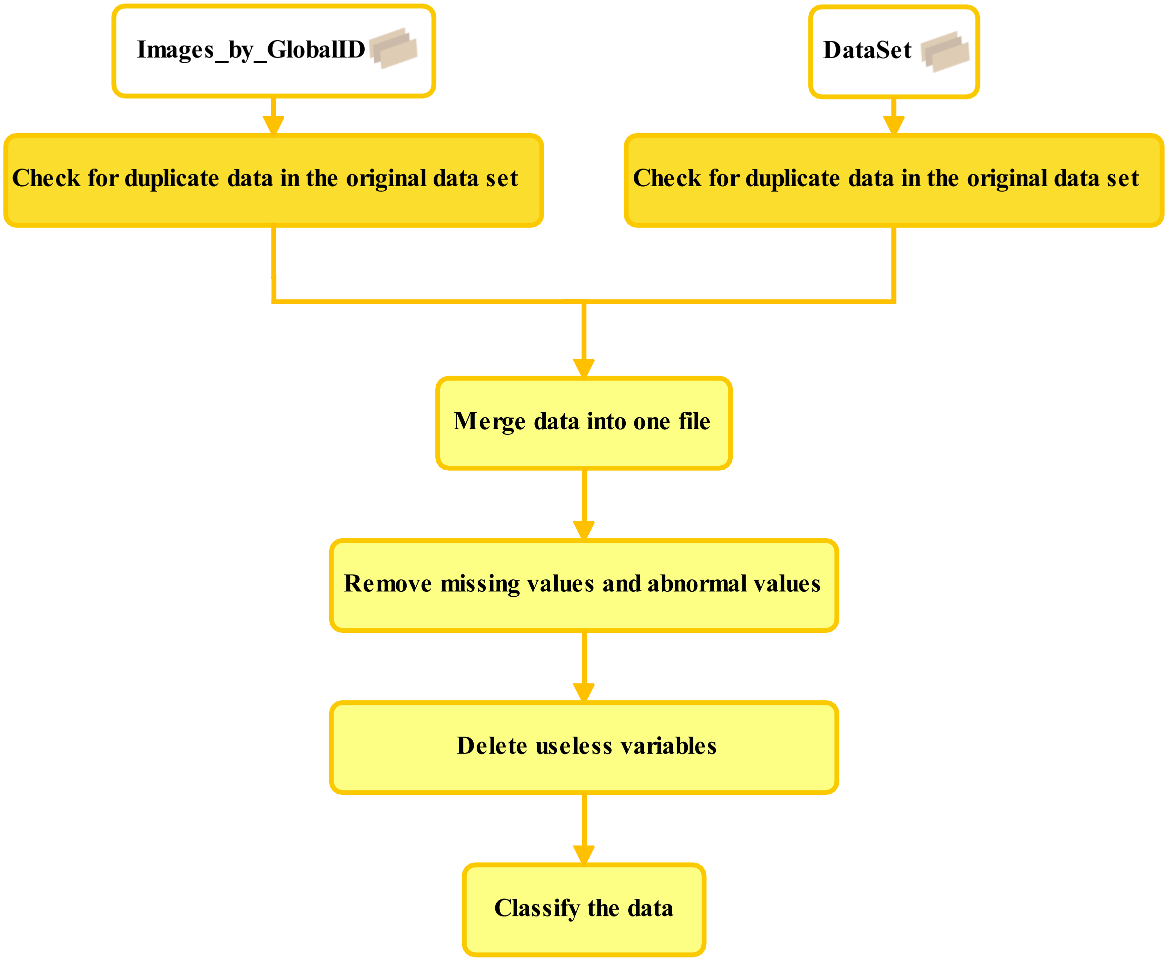

Images_by_GlobalID
DataSet
Check for duplicate data in the original data set
Check for duplicate data in the original data set
Merge data into one file
Remove missing values and abnormal values
Delete useless variables
Classify the data

Figure 1 Data processing flowchart

Firstly, we check for duplicate data in the original data files, but we do not find any.

Then, we merge the data in the two files by the unique global ID of different reports. There are 4440 data in the table file *DataSet* and 3305 data in the table file *Images_By_GlobalID*. After combination, there are 4440 data in the merged file.

After that, we remove missing values and abnormal values. Missing values refer to the null data, which only exist in detective date, notes and lab comments. We have removed three missing data in detective date, since we consider figure legal when it has right detection date. Also, the Asian giant hornet was discovered in September 2019 and there was a relatively long time between adjacent reports for some data, so we wipe out the data before 2010 and malformed data, which is 13 in total.

Moreover, we removed the useless variable—submission date. The submission date is often much later than the detection date. In order to obtain a more accurate time of Asian giant hornet emergence and predict its spread, the detection date is meaningful than the submission date.

Meanwhile, according to lab status, we have divided the data into four categories—Positive ID (14 data), Negative ID (2068 data), Unverified (2327 data), and Unprocessed (15 data).

Finally, 16 data are deleted from the table file, 4424 data and 10 variables are reserved.

3 Model Construction

3.1 Space-time Spread Prediction Model

3.1.1 Problem Analysis

In order to predict the spread of Vespa mandarinia over time, we establish a space-time spread model adopting geographic information and detection date to obtain time nodes and regional scope that Vespa mandarinia may occur. The geographic information is analyzed with the accuracy of kilometers, and the detection date is analyzed with the accuracy of quarter and month.

3.1.2 Model Establishment and Solving

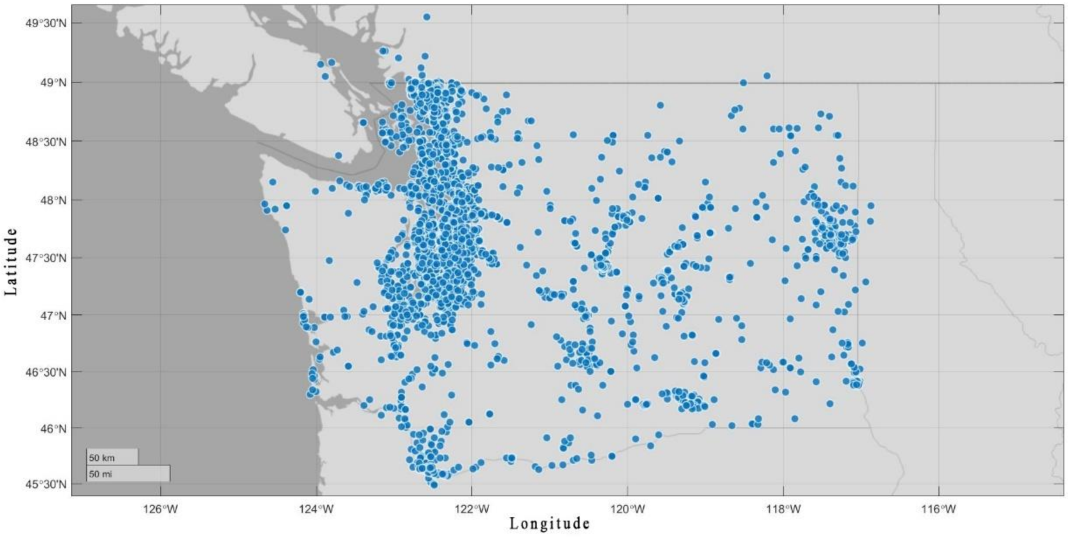


Figure 2 Geographic location of all reports

As for space prediction, we notice that all the geographic range of all reports is relatively wide as figure 2 shows. Regarding geographic information, we extract the latitude and longitude of 14 positive reports from the total data. Then, based on time, we calculated the distance dis_i between j-th positive report and its adjacent positive reports using the equation (1) as follows:

$$\mathrm{dis}_i = \mathrm{geodesic}\big((\mathrm{Lat}[j], \mathrm{Lon}[j]), (\mathrm{Lat}[j+1], \mathrm{Lon}[j+1])\big). \mathrm{km}, i = 1 \dots 14 \qquad (1)$$

where $(\mathrm{Lat}[j], \mathrm{Lon}[j])$ represents latitude and longitude of the j-th data respectively.

Fitting by the *sum of sine* curve, we obtain that $\mathrm{SSE} = 2.878$, $\mathrm{RMSE} = 1.697$ and $\mathrm{R-square} = 0.9976$, which indicate that the fitting effect is great. Based on the result, we predict that the possible ranges of the two future positive reports will be 8.1484 and 3.4021. Therefore, the possible range is $\bar{r} = 5.7753$ by the average value.

Then, we measure the *possible center point* of the region where Vespa mandarinia may occur, which is calculated by the center point of latitude and longitude of 14 positive reports. We consider the earth as a sphere and convert geographic information into polar coordinates firstly. We perform a one-to-to mapping from sphere to the plane and obtain the possible center point $\bar{o} = (-122.7282587308041°, 48.98327185343943°)$. The space prediction is a circle with $\bar{o}$ as the center and $\bar{r}$ as the radius.

As for time prediction, we drop the data for the detection date with few reports firstly. From 2019 to 2020, we count the total number of positive reports and all reports for each quarter and each month, respectively. The probability of positive reports is regarded as the proportion of the number of positive reports to the total number of reports in a quarter or month. We use the *polynomial* and *smoothing spline* fitting to make correlation fitting for quarter and month, respectively. For monthly analysis, we obtain $\mathrm{SSE} = 0.2178$, $\mathrm{RMSE} = 0.1559$ and $\mathrm{R-square} = 0.8992$, which shows its great fitting effect.

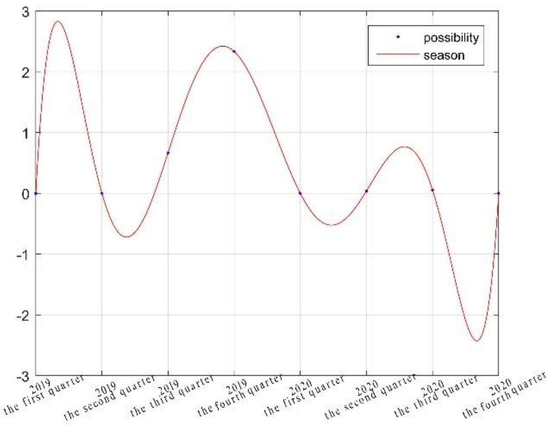


Figure 3 Fitting of the probability of positive reports each quarter and month

According to figure 3, we find that the probability of positive reports fluctuates seasonally—the frequency of occurrence is higher in the second and third quarters, and lower in the first and fourth quarters. Meanwhile, excluding the uneven distribution of sample data, we observe that there are more positive reports from June to October, compared with other months in the same year. It can be confirmed that the quarters and months mentioned above are the result of time prediction.

3.2 CNN-based Image Classification Model

3.2.1 Problem Analysis

With regard to the prediction of the possibility of a mistaken classification, we reprocess the provided data set files and image files. Then, we extract the seven indicators—detection date, notes, Lab Status, Lab comments, Latitude, Longitude, and image information. We focus on each indicator of negative reports, in order to explore the correlation between indicators and establish the second-level object of each indicator. Consequently, we are able to list the possible categories of people's mistaken judgments and arrive at a function that can confirm the report is positive.

3.2.2 Model Establishment and Solving

Firstly, we extract all the non-null lab comments of negative reports, whose number is 2053 in total. We hope to extract keywords and labels of each type by analyzing the responses of experts. Then, we want to find the optimal K value using *K-means text clustering* as shown in figure 4.

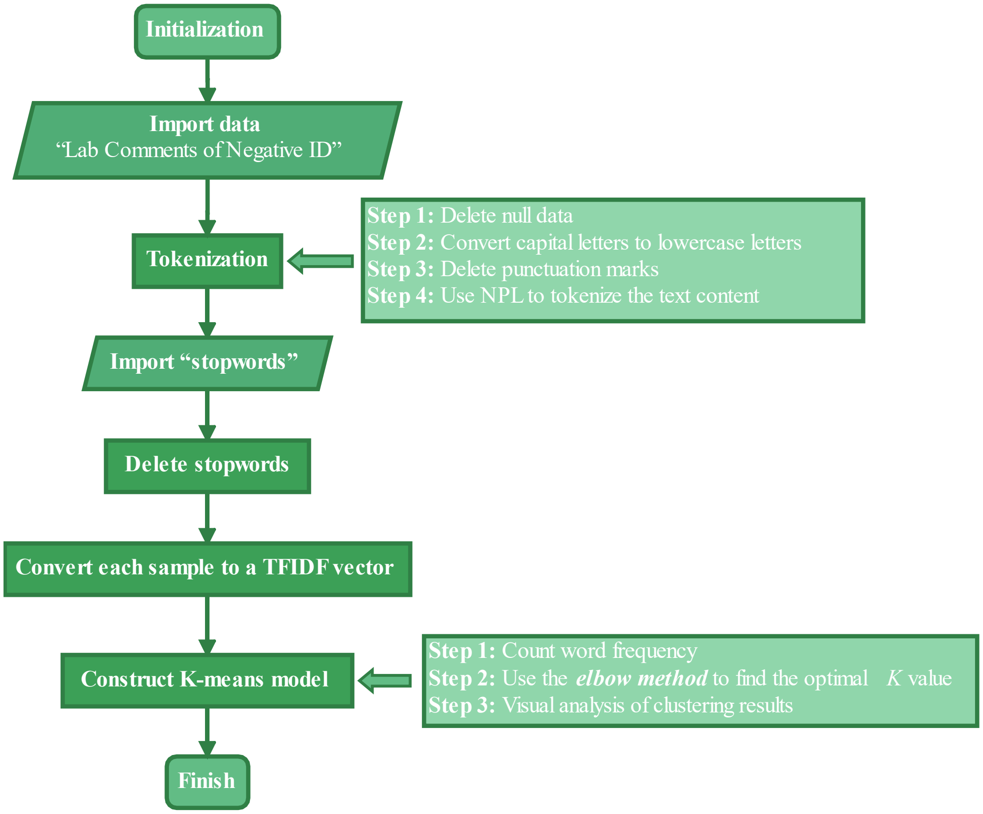

Initialization
Import data
"Lab Comments of Negative ID"
Tokenization
Step 1: Delete null data
Step 2: Convert capital letters to lowercase letters
Step 3: Delete punctuation marks
Step 4: Use NPL to tokenize the text content
Import "stopwords"
Delete stopwords
Convert each sample to a TFIDF vector
Construct K-means model
Step 1: Count word frequency
Step 2: Use the elbow method to find the optimal K value
Step 3: Visual analysis of clustering results
Finish

Figure 4 K-means algorithm flowchart

The K-means algorithm takes the output parameter k and divides a set of n objects into k clusters, which enables the similarity in the same cluster is the highest, and lower between different clusters. While analyzing text data, because the text is unstructured data, it is necessary to preprocess the document before clustering—convert the text data into numerical data. The basic steps of preprocessing are text word tokenization, removal of stopwords, text feature selection and text quantification. Also, we adopt the most classic *VSM model* for text quantization—we use vector $\langle tf-idf_1, rf-idf_2, \ldots, tf-idf_n\rangle$ to represent document A. The equation to calculate term frequency(tf)–inverse document frequency(idf) is

$$tf-idf(t,A)=\frac{\mathrm{lb}(tf(t,A)+0.1)\times \mathrm{lb}(\mathrm{N}/n_i)}{\sqrt{\sum_{i=1}^{n}[\mathrm{lb}(tf(t,A)+0.1)\times \mathrm{lb}(\mathrm{N}/n_i)]^2}} \tag{2}$$

where n_i is the number of documents containing the word t_i, N is the total number of documents ($N=1$ in this literature) and $tf(t,d)$ is the number of times that the word t_i appears in document A.

Then, we want to calculate the text similarity. Assume that the coordinates of the vector **a** and **b** are (x_1, y_1) and (x_2, y_2), respectively. We use the *cosine similarity* algorithm as the text similarity, and the equation is

$$\cos\theta=\frac{x_1x_2+y_1y_2}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}} \tag{3}$$

Therefore, we obtain the distance from the reports to k centers and select the nearest center as the classification point. We Keep iterating until the offset value of the cluster center meets the clustering condition. We get the optimal K value is 9 by the elbow figure shown in Figure 4.

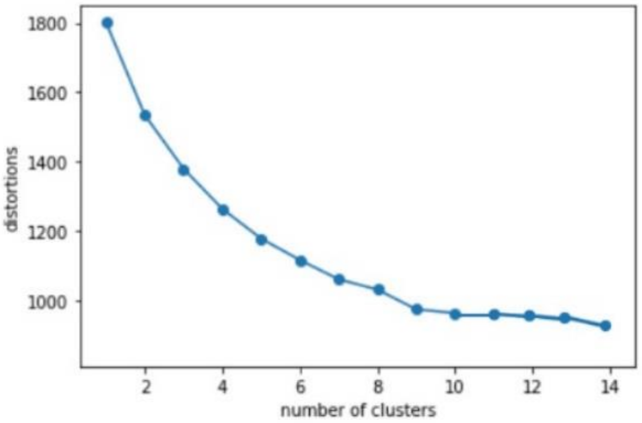
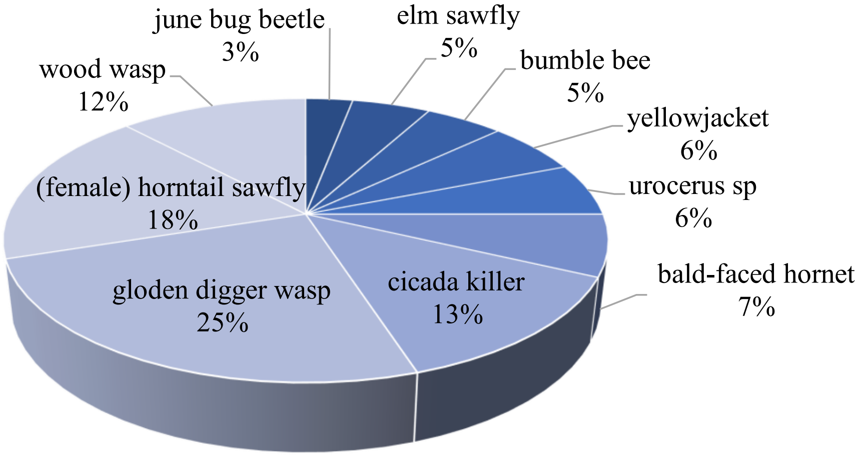


Figure 5 Elbow figure to find the optimal K value in lab comments of negative reports

3.2.3 Analysis and Exploration of Results

As shown in figure 5 and 6 below, we present the hot words and their occurrence frequency in the form of a pie chart and a word cloud chart—the larger the shape, the higher the frequency. It is obvious that in the pie chart, gloden digger wasp, (female) horntail sawfly, and cicada killer account for 25%, 18%, and 13% respectively. Therefore, we conclude that they are most easily confused with Vespa mandarinia and other possible confusing species include bee, etc.



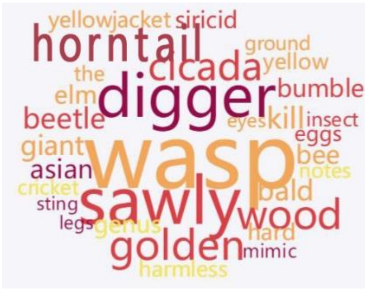


Figure 6 The percentage of look-like species in negative ID

Figure 7 Word cloud of lab comments in negative ID

Then, we try to perform image recognition on 60 unverified figures. According to the results of semantic analysis, we divide the images of negative reports into four categories—glode digger wasp, horntail, cicada killer and bee. We treat the above four types of images and the target image (Vespa mandarinia) as the training set. As figure 8 represents, we can establish a image classification model based on CNN by continuously learning with the convolutional neural network.

Before making the training set, we firstly preprocess the images. We exclud data with abnormal formats in the original data set, and perform screenshots for videos and unzip the compressed package. Finally, we convert all images into jpg format. We notice that, excluding the overly bright, espeically dark, and low-pixel images, only 14 of the 688 selected images are positive reports. It imply that the sample is extremely unbalanced. We enlarge the small sample by 10 times by performing operations such as flipping, cutting, translation, filling, etc.

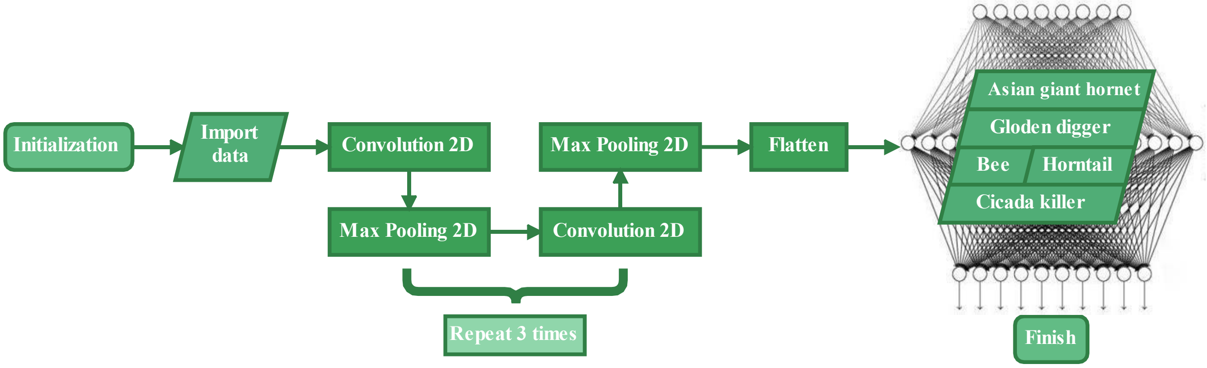


Figure 8 CNN flowchart

CNN consists of three neural network layers—convolutional layer, pooling layer, and fully-connected layer. The convolutional layer is calculated by sliding windows one by one on the upper input layer with k convolution kernels. Each parameter in the convolution kernel extracts the features of each level of the input signal, and the pooling layer reduces the number of useless features, which strengthen the invariance of image features. We adopt three superimposed convolutional layers and pooling layers to improve the accuracy of image recognition.

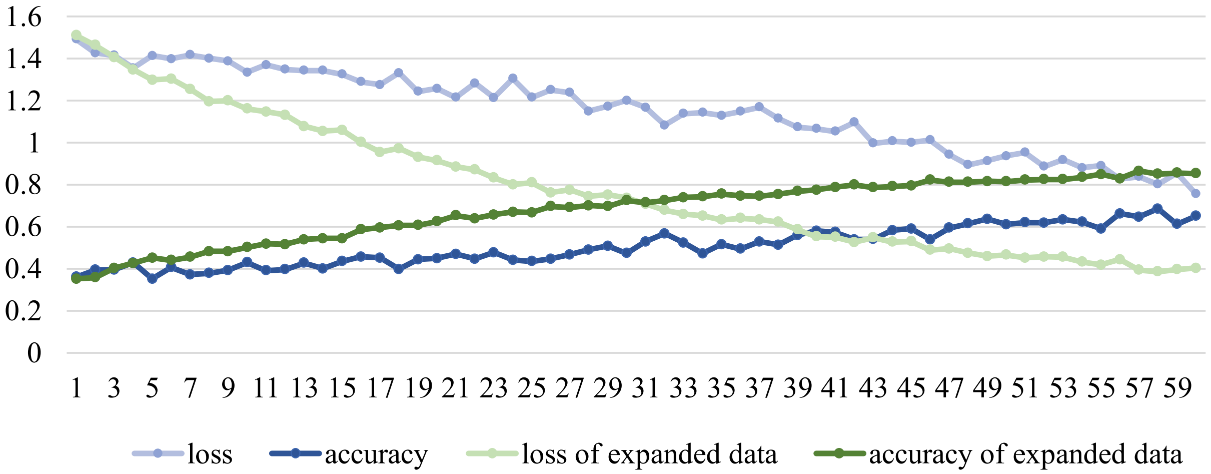


Figure 9 Percentage of loss and accuracy in CNN

In figure 9, blue polyline is the loss and accuracy of the original unbalanced sample after machine learning, and green polyline is the loss and accuracy of the enlarged large sample. Obviously, it is vital to expand the sample, which contributes to improving the accuracy and reducing the loss of image recognition. Finally, the accuracy of image classification reaches 85.28%, which represents that the model learning effect is good.

3.2.4 Evaluation of Results

After constructing the image classification model, we input unverified images (60 available) as the test set. The judgment results of the model are that Vespa mandarinia is 6, cicada killer is 5, horntail is 18, gloden digger wasp is 14 and bee is 14. It is impossible to accurately assess the proportion of each type of images, because the number of test sets is too small. However, we find that horntail and gloden digger wasp account for the largest proportions, which indicates that they are most likely to be confused with the target. Also, this result is consistent with the result obtained by text analysis before.

As a result, we can roughly divide people's mistaken judgments into five categories from the most easily confused to the least likely to be confused—gloden digger wasp, horntail, cicada, bee, other species.

We consider that people's confusion has a high degree of similarity in biological characteristics among these species. In addition, it is also related to the richness and type of species in neighboring Washington State. We find that gloden digger wasp is very similar to Vespa mandarinia. The head of gloden digger wasp is black and yellow, while Vespa mandarinia has a black-yellow head. Also, the body of Vespa mandarinia is black and yellow while gloden digger wasp usually occupies a large part of orange-red. As for horntail, it is mostly pure black, and its length is similar to that of Vespa mandarinia. Moreober, the size of cicada killer is very similar to Vespa mandarinia, but the end of cicada killer's abdomen is black while ASH is yellow. Also, cicada killer's head does not have any yellow. However, because the color types of cicada killer and Vespa mandarinia are exactly the same—black and yellow, so it is easy to confuse. Regarding bee, its size is smaller than Vespa mandarinia, and the yellow of bees is much deeper than that of Vespa mandarinia. Also, the bees are mostly gray and black. This can explain people's mistaken judgment results.

To sum up, people have a 30% chance of considering gloden digger wasp as Vespa mandarinia, a 20% chance of regarding a horntail as Vespa mandarinia, a 15% chance of thinking of a cicada killer as Vespa mandarinia, and a 10% chance of taking a bee as Vespa mandarinia, and there is a 25% chance of recognizing other species as Vespa mandarinia.

3.3 Most Likely Positive Scoring Model

3.3.1 Problem Analysis

After obtaining a report, we can only judge the possibility of this report as positive based on the items listed in the report and historical data. We consider dividing the items in more detail and use principal component analysis to build the most likely positive scoring model. We determine the status of the report according to different levels obtained by the final score.

3.3.2 Model Establishment

As for detection date, we divide it into two indicators—the total number of reports received in a week (week_total_amount) and the number of positive reports received in a week (week_positive_amount). From the detection date of 10th February 2010, we use a week (seven days) as a step to count the number of reports and the number of positive reports for each week and report.

As for notes, we divide it into two indicators—whether it is valid information in notes (notes_if_valid) and whether there are special words in notes (notes_if_special). We find that there are many messy and useless words in notes so we choose some meaning ful words for valid information—inch(es)/large/huge/big/giant/yellow/black/nest/bee/wasp/hornet), which are very helpful for determining a positive report. The effective information is divided into four categories—digital features x_1, color words x_2, nest-related words x_3, and species words x_4, each with a weight of 0.25. The equation to calculate the score of the notes is

$$\text{Notes\_if\_valid} = 0.25 \times (x_1 + x_2 + x_3 + x_4) \tag{4}$$

$$x_i = \begin{cases} 1, & \text{if } x_i \text{ appears} \\ 0, & \text{elsewhere} \end{cases}$$

Regarding special vocabulary, through semantic analysis, we screen some high-frequency words that are opposite to the positive reports, which exert a negative effect on positive reports—hornet(s)t_1/bee(s)t_2/flyingt_3/oranget_4/wasp(s)t_5.

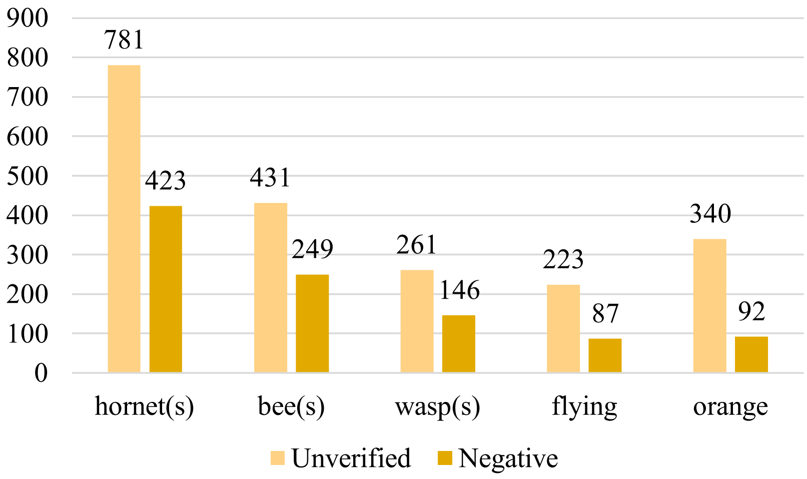


Figure 10 Frequency of special words with antisense

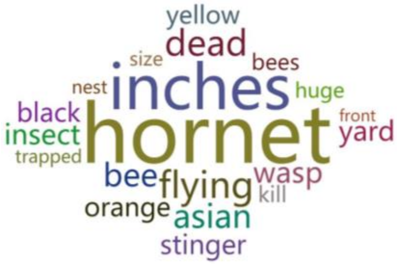


Figure 11 Word cloud of notes in negative reports

Through semantic analysis, the words in figure 10 are those extracted in unverified and negative reports. We select words with a higher number in both two types of reports. We believe that these two kinds of reports are opposite to positive reports, and there is no difference in the judgment of the final result of report status. Therefore, this can be used as an indicator to assist judgment. The equation is

$$\text{Notes\_if\_special} = 1 - 0.2 \times (t_1 + t_2 + t_3 + t_4 + t_5) \tag{5}$$

$$t_i = \begin{cases} 1, & \text{if } t_i \text{ appears} \\ 0, & \text{elsewhere} \end{cases}$$

For geographic information (Geo), we establish the following equation in order to determine whether there is a positive report within 30km of the report:

$$\begin{aligned}\text{Geo} \\ &= 0.6 \times \frac{1}{\ln(\mathrm{e} + \Delta\mathrm{dis}[1])} + 0.25 \times \frac{1}{\ln(\mathrm{e} + \Delta\mathrm{dis}[2])} + 0.1 \times \frac{1}{\ln(\mathrm{e} + \Delta\mathrm{dis}[3])}) \\ &+ 0.05 \times \frac{1}{\ln(\mathrm{e} + \Delta\mathrm{dis}[4])}\end{aligned} \tag{6}$$

where $\Delta\text{dis}[k], k = 1,2,3,4$ is the array of distance difference. Taking i-th report as the current report, it traverses from the nearest detection date to the farthest from i-th report to j-th report. The traversal will end if the distance difference between the dates of i-th report and j-th report exceeds 14 days. During the traversal, if j-th report is found as a positive sample, then $\Delta\text{dis}[k]$ is the distance difference between i-th report and j-th report. If $\Delta\text{dis}[\text{k}] \leq 30$, we store it and discard otherwise. To calculate more convenient, we only store four distance values that meet the conditions. We consider 30 km as the longest flight mileage of Vespa mandarinia in 14 days.

The weight of the above formula is set as 0.6, 0.25, 0.1 and 0.05 according to the detection date from nearest to farthest. We consider the cumulative positive feedback effect, so the weight is successively decreased.

Data integrity is a first-level indicator we add. We believe that it is critical for the identification of reports status. Incomplete data includes missing picture information or notes. Through *fuzzy evaluation*, because the importance of pictures is far greater than notes, we assign weights of 0.9 and 0.1 to pictures and notes, respectively. The equation of Data integrity is

$$\text{Data\_complete} = 0.1 \times \text{z}_1 + 0.9 \times z_2 \tag{7}$$

$$z_1 = \begin{cases} 1, & \text{if it has notes} \\ 0, & \text{elsewhere} \end{cases}$$

$$z_2 = \begin{cases} 1, & \text{if it has figures} \\ 0, & \text{elsewhere} \end{cases}$$

For the image information (Figure_result), we regard it as a human eye judgment. We assume that the result of machine judgment refers to human eye judgment. We firstly recognize all unverified images with the *image classification model*, where the wight of the recognition as Vespa mandarinia is 0.85 (it is the accuracy of the image classification model). Then, we assign the wight of 1 to the figure result of positive reports, and the wight of 0 to the figure result of negative reports.

In conclusion, as shown in figure 11, we divide the assessments that are most likely to be positive reports into two levels, of which five categories that directly affect the assessment results—data integrity, detection date, geographic information, notes, and figure. As for the detection date, we consider both the total number of positive reports received in a week (week_positve_amount) and the total number of reports received in a week (week_total_amount). As for notes, we consider whether it contains valid information (notes_if_valid) and whether there are special words with antisense (notes_if_special).

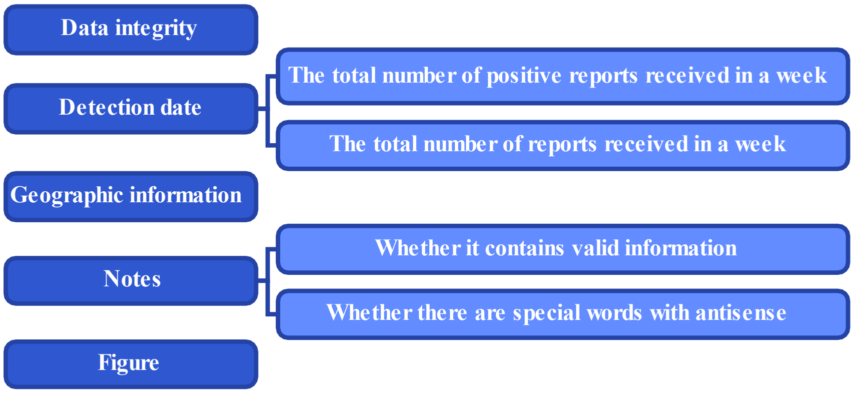

Data integrity
Detection date
The total number of positive reports received in a week
The total number of reports received in a week
Geographic information
Notes
Whether it contains valid information
Whether there are special words with antisense
Figure

Figure 11 Indices of principal component analysis Index

3.3.3 Principal Component Analysis

Based on data analysis, in order to accurately analyze the report results, we adopt *principal component analysis* to extract principal components first.

Given the data composition matrix of n reports and p indicators:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

In order to make the evaluation indicators comparable, we use equation (8) to standardize the data:

$$y_{ij} = \frac{x_{ij} - \overline{x_j}}{s_j}, i = 1 \cdots n, j = 1 \cdots p \tag{8}$$

2here

$$\overline{x_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \tag{9}$$

is the sample mean of the j-th index,

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_{ij} - \overline{x_j}\right)^2 \tag{10}$$

is the average variance of the j-th index.

Therefore, we obtain standardized data matrix:

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{pmatrix}$$

The eigenvalues of the correlation coefficient matrix are obtained through the matrix $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Moreover, the contribution rate of each principal component is

$$\omega_i = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{11}$$

The principle is to select the number of principal components to make it more than 70%~85%, and select r principal components for further research.

3.3.4 Analysis and Evaluation of Results

The final result of principal component analysis:

Component score coefficient matrix

	component		
	1	2	3
Zscore(valid_points)x_1	-.012	-.103	-.733
Zscore(complete_points)x_2	.829	.378	-.143
Zscore(special_words)x_3	-.014	-.012	.534
Zscore(week_amount)x_4	.197	-.212	.034
Zscore(week_positive_amount)x_5	.103	.555	-.004
Zscore(Figure_result)x_6	-.045	.648	.15
Zscore(Geo)x_7	.092	.384	-.023

Table 2 Component score coefficient matrix

We take the three principal components, and their elements are as follows:

$$Q_1 = -0.012x_1 + 0.829x_2 - 0.014x_3 + 0.197x_4 + 0.103x_5 - 0.034x_6 + 0.092x_7 \tag{12}$$

$$Q_2 = -0.103x_1 + 0.378x_2 - 0.012x_3 - 0.212x_4 + 0.555x_5 + 0.684x_6 + 0.384x_7 \tag{13}$$

$$Q_3 = -0.733x_1 - 0.143x_2 + 0.534x_3 + 0.034x_4 - 0.004x_5 + 0.15x_6 - 0.023x_7 \tag{14}$$

Then, we make qualitative analysis of each principal component.

In the first principal component, data integrity and the number of reports received in a week are more weighted. In the second principal component, image recognition, geographic information, and the number of positive reports received in a week are more weighted. In the third principal component, valid information and special words have greater weight.

The contribution rates of these three principal components are normalized, and the processed data is as follows:

Principal component ratio

Q_1	Q_2	Q_3
0.49874	0.34926	0.29566

Table 3 Principal component ratio

The equation of scoring system obtained by quantitative analysis is

$$z = 0.49874Q_1 + 0.34926Q_2 + 0.29566Q_3 \tag{15}$$

We make the score according to the scoring system obtained by principal component analysis, and finally standardize the data to z^* to get the final score.

Finally, we perform *fuzzy analysis* based on the obtained data, and classify the standardized data.

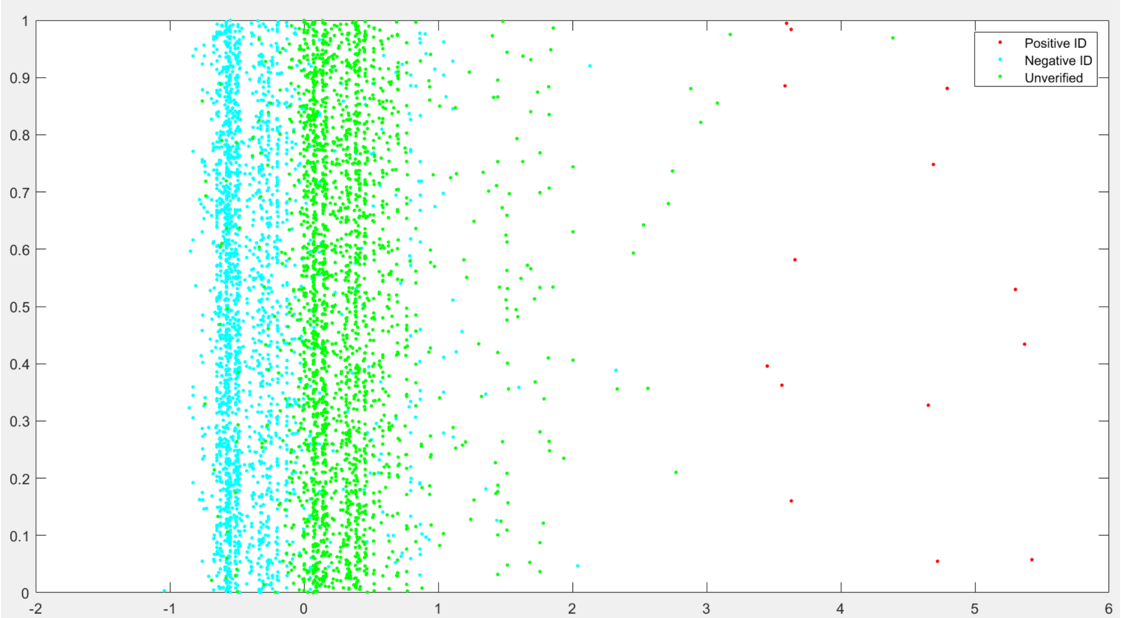


Figure 13 Standardized data in principal component analysis

According to the original data obtained by the principal component score, there are 1897 data in the 2068 negative reports with a score less than 0. In the 2327 unverified reports, only 218 data are less than 0, and 2006 data are concentrated in $(0,1)$. Moreover, the result values of 14 positive samples are all between $(3.5, 5.5)$.

This result is very close to our expected results. We hope to use the *most likely positive scoring model* to estimate the true results of the report with higher accuracy. The scoring model we built successfully separates negative, positive, and unverified reports. Also, as the opposite of positive reports, the numerical distance difference of negative reports and unverified reports is very consistent as well.

3.4 Logistic Update Model

3.4.1 Model Establishment and Solving

Given more new reports, considering Vespa mandarinia are only seen outside the nest when they are hibernating or in the spring before workers have emerged. Under normal conditions, growth rate and mortality rate have a greater contribution rate to population growth. Therefore, we consider the effects of geographical environment, population competition and Vespa mandarinia breeding period on the model over time, Based on logistic model, We make a series of transformations on the blocking factor R, The following differential equations are obtained：

$$\begin{cases} \dot{x} = (r + \alpha) \cdot x, \qquad x(0) = x_0 \\ r = P_i \times E_i \\ P_i = \exp\left(-\dfrac{\Delta_{ij}^2}{2\sigma_a^2}\right) \\ E_i = \gamma \sum_j \exp\left(-\dfrac{\Delta_{ij}^2}{2\sigma_b^2}\right) \times Terrain_j \end{cases} \tag{16}$$

where α is the reproductive rate that changes dynamically with the quarter and month, P_i is the strength of density-dependence experienced by nest, Δ_{ij} is the distance between two nests, γ is a normalizing parameter, σ_a is a measure of local interaction distances and σ_a is the mean foraging distance.

The environmental suitability at a given site is based on a local (Gaussian) average over all nearby locations of the habitat type, as captured by the parameter *Terrain*. [2] Therefore, the environmental suitability is the weighted average across all habitats that are likely to be visited.

Habitat Type	*Terrain* Value
Artificial surfaces	1 (by default)
Agricultural areas	0.87
Forest & Semi-natural areas	0.14
Wetlands	0.02
Other areas	0

Table 4 Habitat suitability as captured by the parameter *Terrain*, these preferences come from recording of nest locations from French national data [1][3]

3.4.2 Model Analysis and Evaluation

We consider that a new queen has a range estimated at 30 km for establishing her nest. Firstly, when the newly reported Vespa mandarinia is more than 30 km away from the last positive report, we should drop it and not screen it. Based on the notes and photos of the new report, we use the previous image classification model to make judgments. By understanding the geographical environment of the discovery site and whether it is a breeding period at the time of discovery, it is easy to obtain Δ_{ij} from the collated data and pass the new report. We predict the possibility of the positive report by solving the differential equations and combing the image classification model.

In addition, we can introduce the migration ability index to optimize the model. Among them, the migration ability index is based on the density and reproduction ability of Vespa mandarinia. A quantitative indicator, Wthat is, when the density is high and it is worth the breeding season, Vespa mandarinia can reach a range greater than 30 km. Through the data given, we can see that the Vespa mandarinia base is relatively small, so it will not reach this situation soon, which can help governments to inform local residents to take preventive measures in time.

Based on space-time spread model, we get the results of seasonal changes in Vespa mandarinia. As for update concerns, we add multiple interference items such as geographic environment, population competition, and reproduction rate. Therefore, we should shorten the cycle on the basis of updating the model on a quarterly cycle. Using principal component analysis, the contribution rate of geographic environment, population competition and reproduction rate to Vespa mandarinia q\are calculated. Therefore, the quarterly period is weighted to obtain the frequency at which the model should be updated.

3.5 ARIMA Model

3.5.1 Problem Analysis

Regarding determining whether pests have been eradicated in Washington State, we consider ARIMA fitting to the number of reports received each week and the weekly mean value of the most likely positive scoring model we have constructed before. If the number of reports submitted every week in the future becomes 0, our weekly average value is less than or equal to 0. (Note: the distribution remains the same after the data is normalized, and the positive sample score average exceeds 2). When both are reduced to the lowest, it can be determined that Washington State has eliminated this pest. We take the weekly average value to calculate, because we consider that the prediction should be staged forecasting on a weekly basis, not for discrete dates.

3.5.2 Model Establishment and Sloving

After observing the data, we selected the year with the largest cumulative number of reports received to forecast. We set 14 July 2018 as the first day of the initial week, with 7 days as the period. Then, we count the total number of reports received and the average scores obtained by the most likely positive scoring model every day within the week.

Taking into account the species reproduction cycle, we believe that these two data are affected by seasons and other factors, which have certain trend items and seasonal effects. Therefore, we want to establish an ARIMA (p, d, q) model to predict the future change trend of the number and score of weekly received reports. Take the weekly score as an example. Firstly, we preprocess the data and make a sequence diagram of the original sequence. We notice that the original data has an upward trend and is not stable. We draw the curve after the first-order difference and the second-order difference through difference processing. We find that the first difference is a stationary time series. Therefore, we build an ARIMA model with $d = 1$.

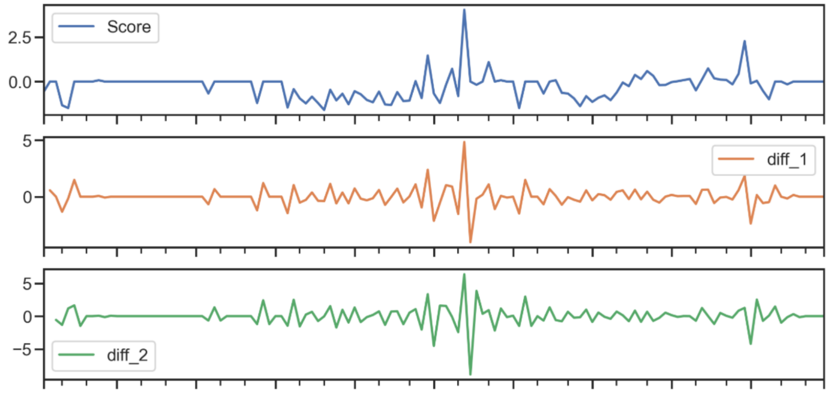


Figure 14 Original sequence, 1st order and 2nd order difference result

We continue to determine the order of the model. p and q are the number of autoregressive terms and the number of moving average terms in the ARIMA model. To select the optimal parameter combination, we adopt the method of judging the truncation, tailing, and overlap of the autocorrelation (ACF) and partial autocorrelation (PACF) coefficients of the stationary series after d times of differences. Then, we combine with the establishment of different parameter models for comparison to determine the best model structure. From figure 15, we get that it is best when $p = 1$, $q = 1$.

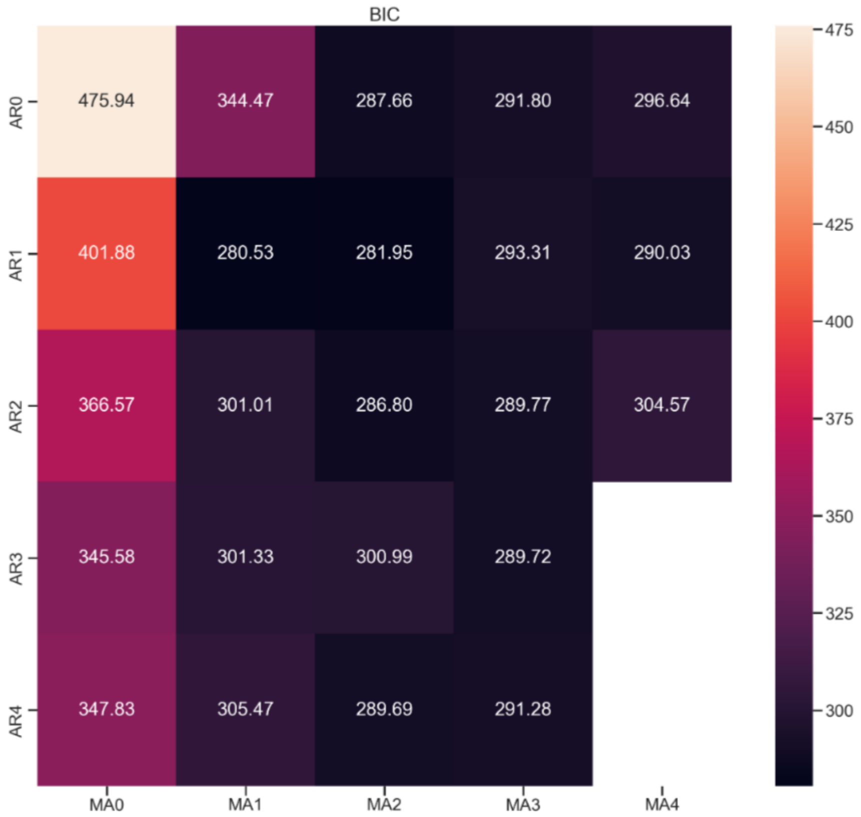


Figure 15 Heat map

We use the ARIMA (1, 1, 1) model d with the best goodness of fit to fit and predict the data. The results are shown in the figure 16 and 17 below.

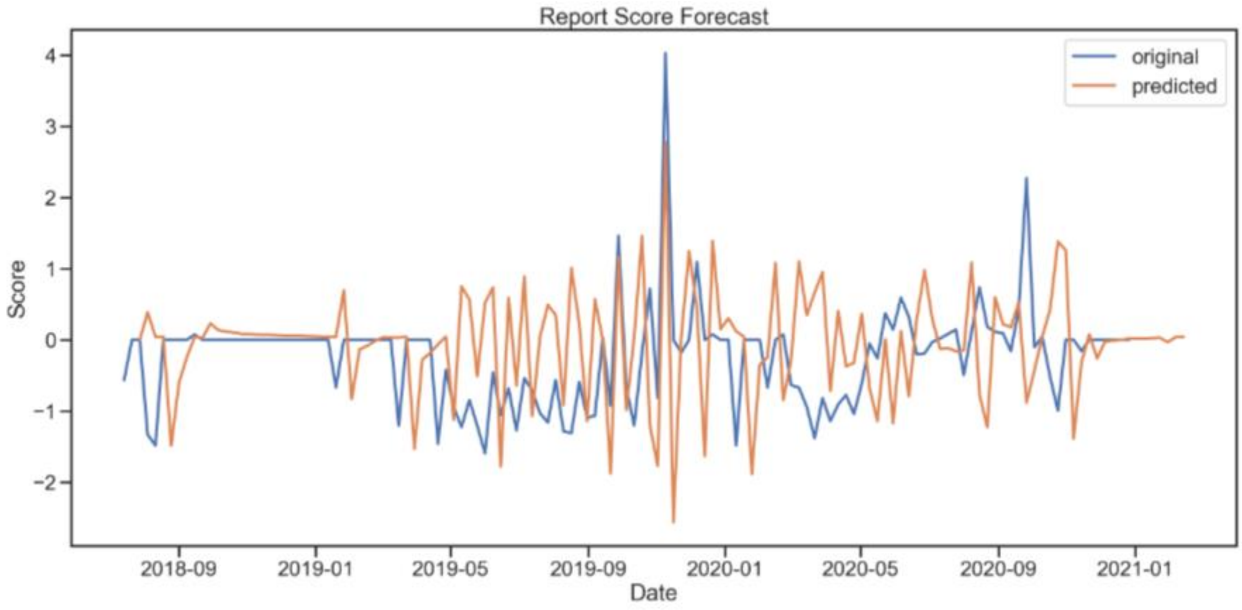


Figure 16 Forecast of report score

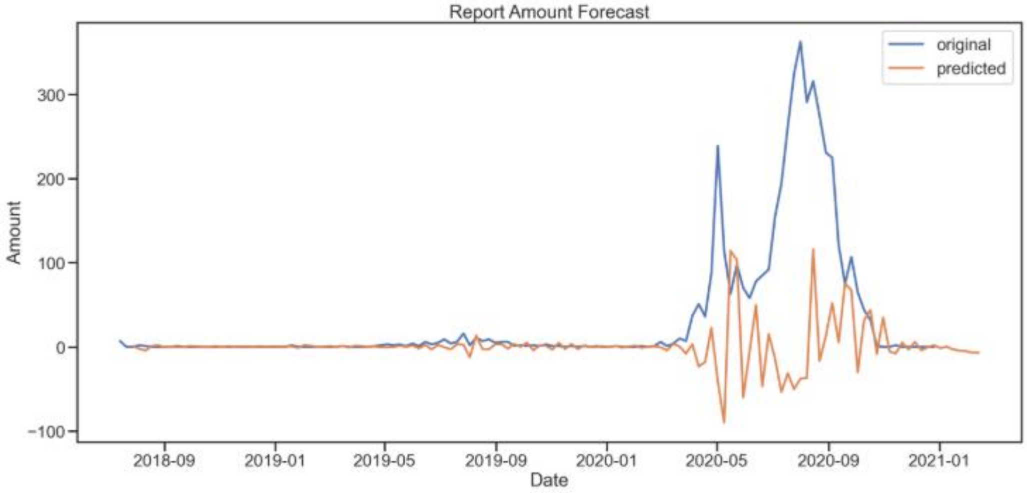


Figure 17 Forecast of report amount

We obtain the average score of the weekly report forecast for the next 4 weeks, which is very close to 0. At the same time, the average number of weekly report is also extremely close to 0. The results can indicate that both will tend to be 0 in the next week to two weeks. Also, it is in line with the analysis results, indicating that the pests have been eradicated in Washington State.

4 Model Evaluation

4.1 Strengths

1. In the image preprocessing process, we screen out the bright, dark, and low-pixel graphics for the first time. After that, we use a three-layer convolution kernel pooling layer to build a CNN model. To balance the small sample problem, we perform operations such as rotation, translation, filling, and segmentation on all selected images. As a result, the sample set is expanded to 10 times the original, and the accuracy of the training result of the CNN classification model can reach more than 80%, which is great.

2. Our most likely positive scoring model shows strong accuracy and robustness. It can be easily implemented to other data with our source code. Also, our model is applicable to any invasive species with obvious biological characteristics.

3. We have made good use of the data provided. Within the scope of the data required by the problem, we construct indicators to the greatest extent, deeply researched the relationship, the constraints, adhesion and their contribution to the outcome of the dependent variable of data.

4.2 Weaknesses

1. The logistic update model we have established is weak. We simply assume that the blocking factor of the logistic model is the product of geographic environment and population competition, but the actual relationship between the two is not necessarily a simple product relationship. Moreover, it may involve more complicated operations. In addition, blocking factors may also include climate indicators, human factors, etc., which have not considered yet in our model.

2. The error in the prediction results of the ARIMA model is a little large due to the poor stability of the data used. The data set provided to us in this problem has a low degree of standardization and a small amount. It may be better if we are offered with more data.

5 Conclusion

To track the Vespa mandariniz's invasion, we proposed a series of novel models to address the sub-issues by selecting the most informative notes and identifying image's descriptors. The proposed models achieve a high accuracy.

1. We establish a space-time spread model adopting geographic information and detection date to obtain time nodes and regional scope that Vespa mandarinia may occur. We find that the probability of positive reports fluctuates seasonally. Meanwhile, excluding the uneven distribution of sample data, we observe that there are more positive reports from June to October.
2. With regard to the prediction of the possibility of a mistaken classification, we can roughly divide people's mistaken judgments into five categories from the most easily confused to the least likely to be confused—gloden digger wasp, horntail, cicada, bee, other species.
3. We establish a most likely positive scoring model to predict the possibility of this report as positive based on the items listed in the report and historical data. We divide the assessments that are most likely to be positive reports into two levels, of which five categories that directly affect the assessment results—data integrity, detection date, geographic information, notes, and figure. Also, we make the score according to the scoring system obtained by principal component analysis. Finally, we perform fuzzy analysis based on the obtained data, and classify the standardized data.
4. On the basis of logistic update model and space-time spread model, we obtain the results of seasonal and monthly variation, accompanied by multiple interference terms such as geographical environment, population competition and reproduction rate. Therefore, the period should be shortened on the basis of the original quarterly update model. We use principal component analysis to calculate the contribution rate of geographical environment, population competition and reproduction rate of Vespa mandarinia, and then weight quarterly cycle to get the update frequency of the model.
5. By our ARIMA model, the average score of the weekly report forecast for the next 4 weeks is very close to 0. Meanwhile, the average number of weekly reports is also close to 0. The results can indicate that both will tend to be 0 in the next week to two weeks. Also, it is consistent with the analysis results, indicating that the pests have been eradicated in Washington State.

Reference

[1] Archer, M. Taxonomy, distribution and nesting biology of the Vespa bicolor group (Hym., Vespinae). *Entomologist's Monthly Magazine* 130, 149-158 (1994).

[2] Keeling, M.J., Franklin, D.N., Datta, S. et al. Predicting the spread of the Asian hornet (Vespa velutina) following its incursion into Great Britain. *Sci Rep* 7, 6240 (2017). https://doi.org/10.1038/s41598-017-06212-0.

[3] Villemant, C. et al. Predicting the invasion risk by the alien bee-hawking Yellow-legged hornet Vespa velutina nigrithorax across Europe and other continents with niche models. *Biological Conservation* 144, 2142-2150 (2011).

Appendices

Appendix A Code of distance calculation

```
import pandas as pd
import numpy as np
from geopy.distance import geodesic
```

```
import time
import datetime
data = pd.read_csv('geodata.csv')
temp_df = Date.str.split('-',expand=True)
temp_df.columns = ["year","month","day"]
data = pd.concat([data,temp_df],axis=1)
data['res'] = 0
data['res_1'] = 0
data['res_2'] = 0
data['res_3'] = 0
data['res_4'] = 0
num = 0
for i in range(170,4424):
    num = 0
    k = i-1
    for j in range(1,400):
        if int(data['year'][k]) == int(data['year'][i]):
            if int(data['month'][i])==int(data['month'][k]):
                if (int(data['day'][i])-int(data['day'][k]))>13 or (int(data['day'][k])-int(data['day'][i]))>13:
                     break
            else:
                if int(data['month'][i]) - int(data['month'][k])>1 or int(data['month'][i]) - int(data['month'][k])>1:
                    break
                else:
                    tmp_date = int(data['day'][i]) + 30 - int(data['day'][k])
                    if tmp_date>13:
                        break
        else:
            if int(data['year'][k])!=12 and int(data['year'][i])!=1:
                break
            else:
                tmp_date_1 = int(data['day'][i]) + 30 - int(data['day'][k])
                if tmp_date_1>13:
                    break
        if data['ifpositive'][k] == 1:
            distance = geodesic((data['Latitude'][k],data['Longitude'][k]),
(data['Latitude'][i],data['Longitude'][i])).km
            if distance<30:
                if num == 0:
                     data['res'][i] = distance
                     num = num+1
                elif num == 1:
                    data['res_1'][i] = distance
                    num = num+1
                elif num == 2:
                    data['res_2'][i] = distance
                    num = num+1
                elif num == 3:
                    data['res_3'][i] = distance
                    num = num+1
                elif num == 4:
                    data['res_4'][i] = distance
                    num = num+1
        k = k-1
```

Appendix B Code of K-means

```
import pandas as pd
import numpy as np
neg_comments = pd.read_table('comments.txt',names = ['neg_comments'],encoding = 'utf-8')
neg_comments = neg_comments.dropna()
neg_comments['neg_comments'] = neg_comments.neg_comments.str.lower()
stopwords = pd.read_csv("stopwords.txt",sep = "\t",index_col = False,quoting = 3,names =
['stopwords'],encoding = 'utf-8')      ####Import stopwords
content = neg_comments
content = content.values.tolist()
from nltk import word_tokenize      ####Split words
content_temp = []
for line in content:
    current = line
    current = ".join(current)
    current = word_tokenize(current)
    content_temp.append(current)
def cutwords(contents,stopwords):      ####Remove stopwords
    contents_clean = []
    allwords = []
    for line in contents:
        temp_clean = []
        for word in line:
            if word in stopwords:
                continue
            temp_clean.append(word)
            allwords.append(str(word))
        contents_clean.append(temp_clean)
    return contents_clean,allwords
stopwords = stopwords.stopwords.values.tolist()
contents = df_content.content_temp.values.tolist()
contents_clean,allwords = cutwords(contents,stopwords)
df_content = pd.DataFrame({'contents_clean':contents_clean})
notes_words = []
for line_index in range(len(X)):
    try:
        words.append(' '.join(X[line_index]))
    except:
        print (line_index)
from sklearn.feature_extraction.text import TfidfVectorizer      ####Convert to TFIDF vector
vectorizer = TfidfVectorizer(max_features = 688,lowercase = False)
vectorizer.fit(notes_words)
from sklearn.cluster import KMeans      ####K-means clustering
KM_model = KMeans(n_clusters=8)
KM_model.fit(vectorizer.transform(words))
label_pred = KM_model.labels_
```