Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2106561

Control pests:Biological diffusion simulation and Positive report screening

Summary

In recent years, due to the invasion of Vespa mandarinia in Washington State, the ecological environment of the United States is facing great challenge. WSDA has now received 14 confirmed sighting reports and more than 4000 suspected reports. In order to explore the impact of this creature on the United States, we modeled and analyzed the law of the reproduction and expansion of Vespa mandarinia.

We first define the reproduction rules in the model based on the inherent biological attributes of Vespa mandarinia. It is assumed that the frequency of occurrence of bee colony at various points in its range of activities obeys a two-dimensional normal distribution, and the concept of migration vector is introduced. Use the ARMA time series to predict the value of the migration vector later. On these foundations, an MRV(Migration and reproduction model of Vespa mandarinia) model was established to simulate its expansion process. Because Vespa mandariniad has less historical data, we use other bee species with a larger number to fit the parameters and verify the rationality of our model. After using the MRV model for simulation, we found that without interference from external forces, from 2020 to 2021, the number of bee colonies will increase three times, and a large number of the United States will be troubled by them.

We use the data sets and image files of 4400 reports received by WSDA in the past to predict the credibility of related reports. Three dimensions of information were used to evaluate the report, and a VTP(Event screening model based on vision, text and place) model was established.

In order to reduce the workload of the staff of the WSDA, and because Vespa mandariniad has too little data and the recognition accuracy is too low. Therefore, in the use of the model, we will discard events where the vision_scores part of other classes is greater than 0.5.Then, we tested and evaluated the accuracy of the model. For the MRV model, we use the ADF(Augmented Dickey-Fuller test) test to prove that the sequence is stable.

For model two, we mainly evaluate the image recognition module. For the training effect of the neural network, we use the PR curve and mAP to evaluate. The prediction effect of the model is generally excellent for most classes, but it is not ideal for some classes with less data.Through our fitting of other insect data, we found that the MRV model can only predict the populations of insects reported for 4 consecutive months. At the same time, the prediction of migration vectors requires the participation of historical data. And then, we set the conditions for judging the extinction of insects.

Finally, through our simulation data, a reasonable pest elimination strategy was given to the Washington Department of Agriculture. We expect that Vespa mandarinia will be effectively restrained in Washington State.

Keywords: Two-dimensional Gaussian distribution;Faster r-cnn;Vespa mandarinia ;Species migration vector;Information screening

Contents

1 Introduction

1.1 Background

Vespa mandarinia is one of the largest wasps in the world. It is also known as the Chinese hornet, Taiwan hornet, etc. It is classified into the animal kingdom, arthropod phylum, insect class, Hymenoptera, slender waist Suborders, Vespas, and there are many subspecies of the golden ring wasp, such as the matchmaker wasp in Yunnan and the Japanese hornet in Japan. It is native to temperate and tropical East Asia and is mainly distributed in mid-altitude mountainous areas. In September 2019, an Asian giant wasp nest was discovered and destroyed on Vancouver Island, British Columbia. In December, the Washington State Department of Agriculture confirmed that a dead specimen was found in Washington[1]. This was the first record of this species in the United States. Because Vespa mandarinia is an invasive species, as a predator of European honeybees, Vespa mandarinia greatly affects the living conditions of honeybee populations within 5 miles. After entering the slaughter stage, a group of 20-30 bumblebees can kill 5000-25000 bees within an hour. For example, this wasp queen had been discovered in France in 2004. Just eight years later, its traces have been almost all over France.In 2012, it directly attacked a 54-year-old man and caused his death. A fear of invasive species arouse. A fear of invasive species. The Washington state government has set up a help line and website for people to report the traces of these bumblebees. Of the thousands of reports received, most of them were false alarms, but there were also some real situations. In order to efficiently track the reproduction and development of this species in the United States, the ICM team needs to monitor the reproduction status of Vespa mandarinia. Predictive modeling, and analyze various error classifications from numerous reports.

1.2 Restatement of the problem

In order to prevent a series of losses caused by the invasion of biological species, the Washington state government needs us to obtain information about the presence of Vespa mandarinia in the United States from a large number of unverified suspected reports and follow up on these reports. The specific problems that we needs to solve are as follows:

1. Analyze the growth habits of organisms, establish the breeding model of Vespa mandarinia, visualize their results, and predict their spreading areas and directions.
2. The model is established according to the pictures, videos and words to deal with the eyewitness events and get the possibility of eyewitness event classification. Determine a feasible method to screen the events and get the most likely positive witness reports, so as to reduce the workload of staff.
3. Verify the established model and get the update frequency of the model.
4. Sensitivity analysis of the model.
5. According to the model, when the conditions are met, the pest is proved to be extinct.
6. Determine a plan to eliminate pests.

1.3 Our work

In this article, we established an MRV model based on the habits and activities of organisms. In the model, we consider that the biological activity presents a two-dimensional normal

distribution law under ideal circumstances. At the same time, in order to fit the reality: organisms will migrate due to factors such as temperature, food, altitude, etc., which introduces the concept of biological migration vectors. Use historical data to calculate the migration vector of each period of time, and use the time series ARMA model to predict the migration vector of the next stage.Combining the predicted biological migration vector with the activity law of the two-dimensional normal distribution, the biological transition probability density matrix is obtained, which is used to evolve the change over time, the direction and number of biological reproduction, and help the Washington State Department of Agriculture to predict Vespa The spread of mandarinia.

We extract the text of Lab Comments in 2021MCMProblemC_DataSet.xlsx to obtain 12 different insect categories and classify them. In order to be able to screen a large number of submitted sightings, priority is given to researching incidents that are most likely to be Positive, reducing the useless workload of Washington State Department of Agriculture staff. We integrated the uploaded images, the notes in the incident, and the latitude and longitude of the sightings to establish a VTP model.

First, use the image data in the 2021MCM_ProblemC_Files file to build a Faster r-cnn[4] network for image recognition. Compare the image provided by the event to be processed to obtain vision_scores. Then we searched for the different characteristics of 12 kinds of insects, built a list of text characteristics of 12 kinds of insects, compared them with the Notes of the event to be processed, and got text_scores. Then according to the locations where insects were found in the past, calculate the center of insect gathering, get the place_scores compared with the latitude and longitude of the most frequent activity of each insect in the past. Finally, use vision_scores, text_scores, and place_scores to calculate the VTP score.

In addition, we also showed how to use the VTP model to handle a large number of incidents and reduce the workload of the staff of the Washington Department of Agriculture.

Regarding the migration vector in the MRV model, we give the results of using the migration vector prediction and the reproduction results under real conditions, verifying that the migration vector has a good predictive effect, can predict the future development direction of insects, and test that ARMA can Used in the prediction of migration vectors.

For the VTP model, we give a series of indicators such as AP value, mAP value, False Positive, True Positive, etc. of the image detection results of various types of insects to detect the robustness of the Faster r-cnn network.

Sensitivity analysis of mortality and mutation rate in MRV model was performed. It is found that the mutation rate has no obvious relationship with the number of reproduction, while the mortality rate has a serious impact on the number of reproduction. If you want to prevent and treat Vespa mandarinia, you need to increase its mortality, not interfere with its mutation rate.

According to the MRV model, we give the conditions of how to judge the extinction of pests, and give control suggestions to help local people control pests.

2 Basic assumption

1. Assuming that Vespa mandarinia will only appear in groups, that is, once a new hive appears somewhere, there must be queen bees and worker bees.

2. Assuming that in the process of population expansion, the probability of Vespa mandarinia

appearing in its range of activities obeys a two-dimensional normal distribution.

3. After the position of the hive is determined, the position of a single colony will not change.

4. Vespa mandarinia has expanded to the most suitable environment for its survival.

3 Data preprocessing

In this chapter, we performs label extraction and data elimination operations on the 4440 reports received in the DataSet file to facilitate the application of data in image recognition and models.

1. **Label Extraction** A total of 4440 reports received are listed in the DataSet file. Only a few of them belong to the confirmed Vespa mandarinia haunt report. Through the distinction of the laboratory, all reports are divided into four types: Positive ID, Negative ID, Unverified, Unprocessed. Unverified and Unprocessed types are reports that cannot be classified or have not yet been classified. In the report confirmed as Negative ID, the Lab Comments column indicates the category. By extracting text keywords from the laboratory opinions, the image files in the report can be labeled with their respective types, and the core content in the images can be marked.

2. **Data elimination** Because the sample size of some types is too small or the description is vague, it is difficult to play a positive role in subsequent research. We eliminated the report types and types of reports that were not accurately judged in the laboratory's opinions and the sample size was too small (less than 10).

4 Model I: Migration and reproduction model of Vespa mandarinia

4.1 Biological activity law

1. **Gregariousness[3]**
Vespa mandarinia is a social organism, which has a beehive and a queen bee. The queen bee is responsible for breeding, and the worker bee is responsible for foraging.

2. **Scope of activities**
During the foraging process, Vespa mandarinia is usually only in the area 1-2 km away from the nest and no more than 8 km, so its activity radius is set at 8 km. And the distance between its location and its nest satisfies the normal distribution.

3. **Environmental preference**
Vespa mandarinia is mostly used to living in the mountain area of middle altitude, but the wasp that just appeared in the American continent does not live in the best living condition directly. Therefore, we assume that the population will migrate with time, and gradually move to the best living environment at the same time of population expansion.

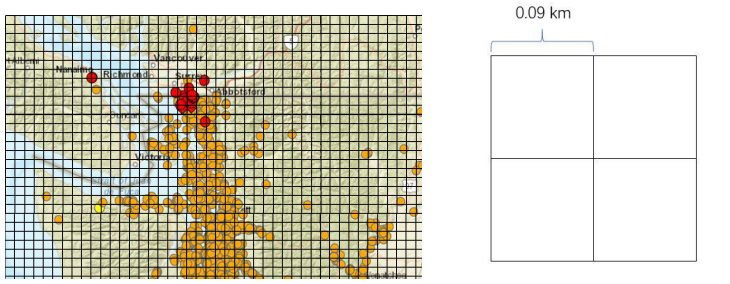
4. **Life cycle**
The life of Vespa mandarinia has a certain periodicity. The peak time of colony population is August, and the time of Queen production is usually September. At the end of December, most of the colonies died, and only queens survived and wintered.

4.2 Establishment of model

We mapped the whole possible space of Vespa mandarinia into a network space map, and regarded the wasp in a cell as a unit without considering the number of individuals in the colony.

4.2.1 Distribution of swarm activity range

In order to accurately simulate the expansion process of Vespa mandarinia, first, its possible living space is meshed with longitude and latitude as the division standard, where the side length of a single cell is 0.01 longitude and 0.01 latitude respectively. By investigating the true distance corresponding to the Longitude [122,124], [Latitude 48,50] area, to simplify the calculation, we assume that the distance spanned by 1 longitude and 1 dimension is the same, and the corresponding relationship is that 1 degree is approximately equal to 90 kilometers,, each cell is a square with a side length of 0.9 kilometers.



(a) Partial area grid

(b) Actual distance of single grid

Figure 1: Schematic diagram of grid generation

At the same time, due to the gregariousness of Vespa mandarinia, we ignore the difference between queen bee and worker bee in population expansion, so there are only two cases of single cell state, the presence of wasps or the absence of wasps.In the model, it is expressed as follows($L_i$represents the binary state of the ith point in the grid):

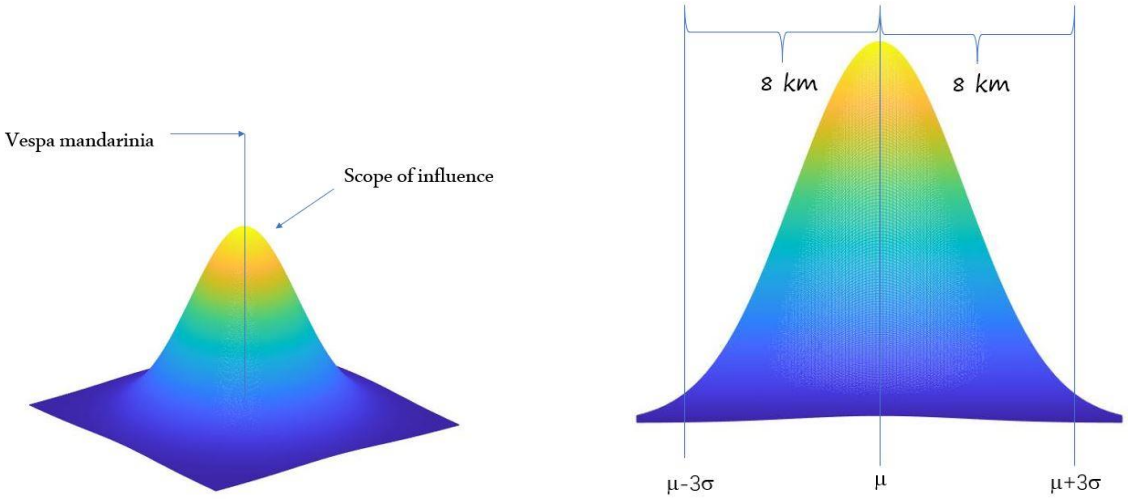
Table 1: The meaning of L_i

Value of L_i	Means
0	There is no Vespa mandarinia
1	Vespa mandarinia exists

In order to predict the expansion of Vespa mandarinia more accurately, we set the size of a single cell to be smaller when dividing the cell. Therefore, in the established two-dimensional cell space, the state of a single cell is not just determined by its Moore neighborhood, it depends on the expanded Moore neighborhood. By defining the probability matrix of the activity range of Vespa mandarinia, it can help describe the expansion rules under this situation.

Define the scope of Vespa mandarinia, the activity is centered around the location where the positive ID has been confirmed, and the activity is carried out in a range of 8 kilometers. The closer to the activity center, the higher the probability of their appearance. Therefore, the

occurrence probability of their activity locations is a two-dimensional normal distribution, and the center of the matrix is the mean value of the normal distribution, which also corresponds to the honeycomb location on the actual map.



(a) Two-dimensional normal distribution of activity range

(b) Marginal distribution of two-dimensional normal distribution

Figure 2: Vespa mandarinia's scope of influence

The random variable distribution of probability density satisfies the following formula:

$$f(x,y) = (2\pi\sigma_1\sigma_2\sqrt{1-\rho^2})^{-1} e^{[-\frac{1}{2(1-\rho^2)}(\frac{(x-\mu_1)^2}{\sigma_1^2})-\frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}]} \tag{1}$$

μ_1, μ_2,σ_1,σ_2,ρ are constants, the center point is (μ_1, μ_2).Due to the complexity of the two-dimensional normal distribution, we abstract the marginal distribution of the two-dimensional normal distribution for consideration.The marginal distribution can be expressed as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

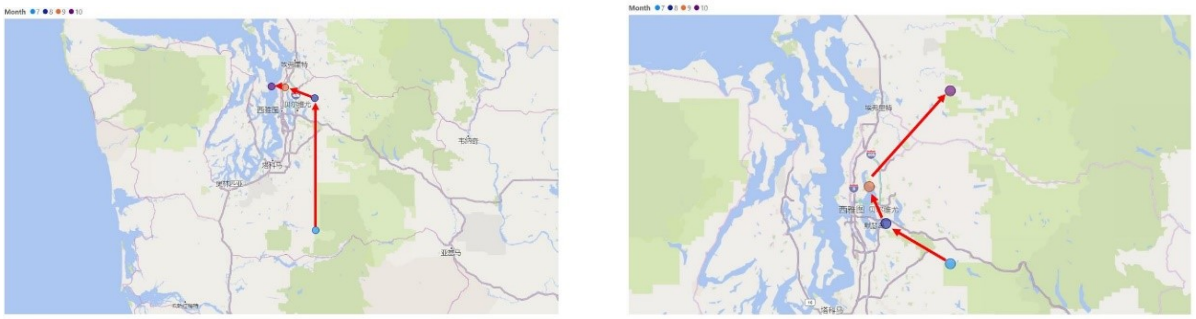
In the normal distribution, the proportion of the area falling in the interval $(\mu-3\sigma, \mu+3\sigma)$ is 99.73%. Therefore, there is a corresponding relationship $3\sigma = 8$. The side length of a single cell in the cell space is 0.09. Therefore, the size of the probability matrix S_i is 177*177.

$$S_i = \begin{bmatrix} f_{1,1} & \cdots & f_{1,177} \\ \vdots & \ddots & \vdots \\ f_{177,1} & \cdots & f_{177,177} \end{bmatrix} \tag{3}$$

Different biological groups have different laws of survival and death. Their death is determined by environmental conditions, race life, etc. We uses D to represent the probability of biological death events, $D = (D_1, D_2, \cdots, D_n)$. According to the activity range probability matrix S_i, the propagation and expansion of the wasp is simulated. This distribution shows us the possibility of bee colonies in other areas under the influence of existing colonies.

4.2.2 Biological migration vector

Due to the special habits of living things, under the line of food, climate, altitude, and terrain, they may gradually migrate toward an ideal living environment. Therefore, we introduce a new variable to describe this trend, called the biological migration vector.



(a) (b)

Figure 3: Schematic diagram of overall migration trend of bee colony

We use $\vec{V_t}$ to describe the magnitude and direction of this vector. Because the position of the wasp's hive does not change according to the change of seasons, $\vec{V_t}$ describes the trend of the new hive deviating from the original hive position, not the change of the same hive position.

Define the position center of the wasp colony observed every three months in the existing data every three months as P_t, and get $P = (P_1, P_2, \cdots, P_t)$, and the migration vector of bee colony $\vec{V_t} = P_t - P_{t-1}$, thus get the biological movement trend $V = (\vec{V_1}, \vec{V_2}, \cdots, \vec{V_t})$. Considering that past data will also have an impact on the current bee colony, a certain weight is assigned to the migration vector of different time nodes. According to the principle that the closer the time the greater the impact, the migration vector in the current state $\vec{V_{predict}}$.

$$\vec{V_{predict}} = \alpha^1 * \vec{V_t} + \alpha^2 * \vec{V_{t-1}} + \cdots + \alpha^{t-1} * \vec{V_1} \tag{4}$$

Combine $\vec{predict}$ with the probability density S_i of the emergence of a new colony to obtain the migration probability density $M_{i,t}$ at point i at time t.

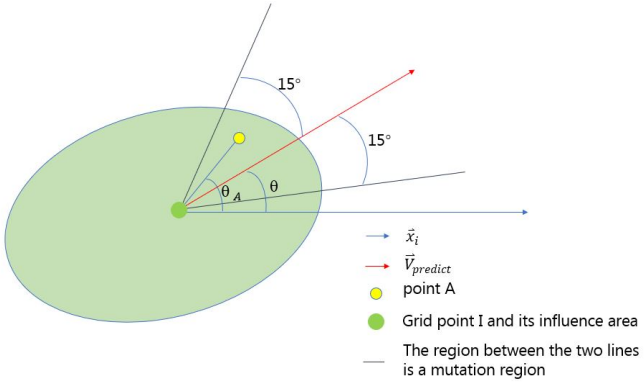


Figure 4: Two-dimensional presentation of mutation

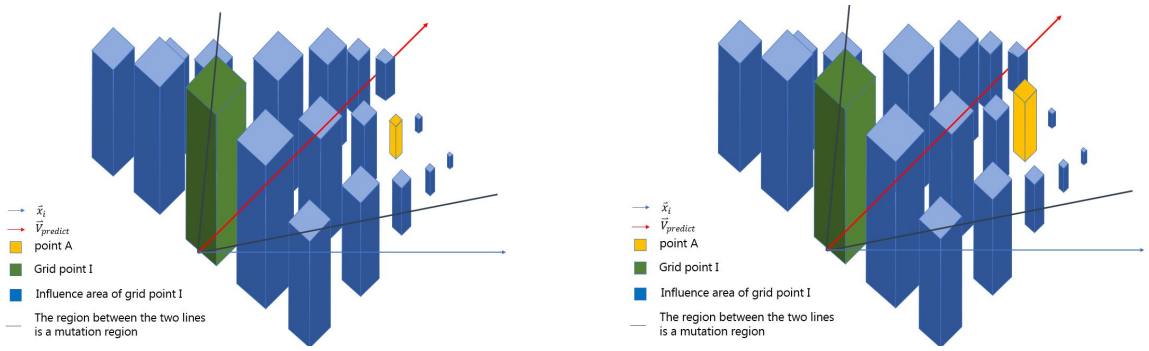
Step:1 Define the reference direction $\vec{x}$ in the map.

Step:2 Taking the point i in the grid as the origin, make a coordinate axis $\vec{x_i}$ in the direction of $\vec{x}$. Define the angle between $\vec{V}_{predict}$ and $\vec{x_i}$ as θ, and the length of $\vec{V}_{predict}$ as L.

Step:3 Select $[\theta - 15°, \theta + 15°]$ as the angular range where sudden changes may occur, and perform sudden change processing on the activity probability within the range of $[\theta - 15°, \theta + 15°]$ centered on point i. The R rule for mutation was established as follows

(a) Randomly select a point A in the range, the angle between $A$and $\vec{x_i}$ is θ_A, and the distance between point A and point i is L_A.

(b) The probability of mutation at point A is $\frac{1}{\sqrt{|L_A - L|}}$. If a mutation occurs, the activity probability of Vespa mandarinia at this point doubles.



(a) Before the change (b) After the change

Figure 5: 3D catastrophe effect picture

After the mutation, we will get the biological migration probability density matrix $M_{i,t}$ after considering the biological habits, which can really determine the direction and number of Vespa mandarinia's reproduction.

4.3 Integration of MRV models

After integrating the above rules, we obtained the model MRV, which describes the reproduction and expansion of Vespa mandarinia. The model is described as follows:

Table 2: Symbol description

Symbol	Means
Map_t	The state of the whole grid at time t
S	$S = (S_1, S_2, \cdots, S_j)$ $S_J$denotes the probability that organisms exist in the surrounding environment. $S_j \in f(x, y)$
V	The set of biological migration vectors
R	The rule of distribution probability mutation
D_n	The probability of death of species n, for Vespa mandarinia, $D_{vm} = 0.1$

$$MRV = (Map_t, L, S, V, R, D) \tag{5}$$

The process of using the MRV model to predict the reproduction and expansion of Vespa mandarinia is as follows:

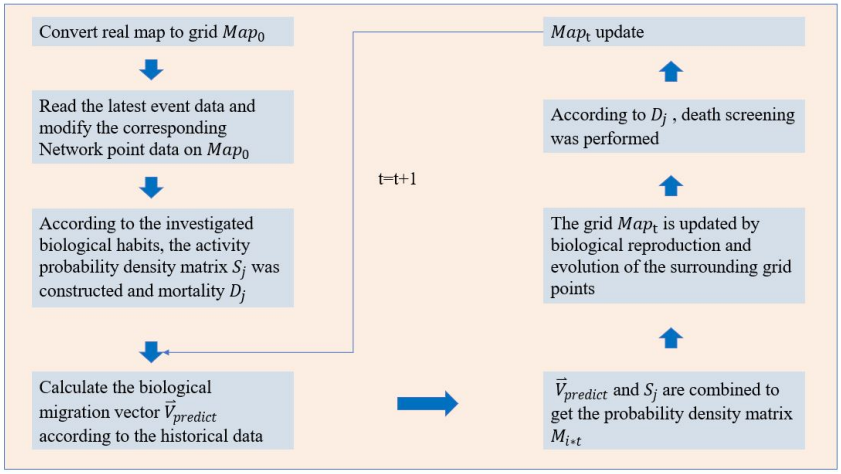


Figure 6: The calculation process of the MRV model

4.4 Model validation and results

Model validation

When using many time series models, such as ARMA and ARIMA, the time series are required to be stationary, so stationarity tests are generally required when studying a period of time. The more commonly used strict statistical test method is Augmented Dickey-Fuller test (ADF), also called unit root test.

In an autoregressive process: $y_t = by_{t-1} + a + \epsilon_t$, if the lag term coefficient b is 1, it is called the unit root. When the unit root exists, the relationship between the independent variable and the dependent variable is deceptive, because any error in the residual sequence will not decay as the sample size increases, that is to say, the influence of the residual in the model is permanent of. This kind of regression is also called pseudo-regression. If the unit root exists, this process is a random walk). The ADF test is to determine whether the sequence has unit roots: if the sequence is stable, there is no unit root; otherwise, there will be a unit root.

The H0 hypothesis of the ADF test is that there is a unit root. When we look at the result of whether the sequence is stationary, we generally look at the p_value value of the second part first. If the p_value value is less than 0.05, it proves that there is a unit root, which means that the sequence is stationary. If p_value is greater than 0.05, it proves non-stationary.

Table 3: Test results of original sequence

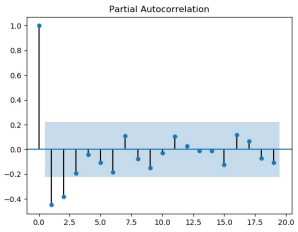
vector_lng of golden digger wasp	value		vector_lat of golden digger wasp	value
Test Statistic Value	-10.82801362		Test Statistic Value	-7.523629479
p-value	1.74E-19		p-value	3.73E-11
Lags Used	1		Lags Used	3
Number of Observations Used	76		Number of Observations Used	74
Critical Value(1%)	-3.519480535		Critical Value(1%)	-3.521980318
Critical Value(5%)	-2.900394509		Critical Value(5%)	-2.90147011
Critical Value(10%)	-2.587498428		Critical Value(10%)	-2.588072155

If p_value is close to 0.05, the critical value must be used for judgment. That is to say, if p_value is close to 0.05, the value of the first part of the statistics will be compared with the

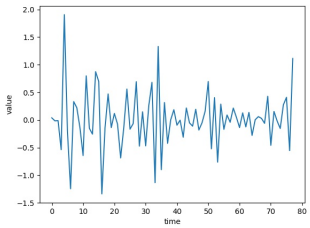
critical value of the fifth part. The value of the statistic is smaller than the critical value, which proves that the sequence is stationary, otherwise it is non-stationary.

After calculation, it is proved that $p = 1.736806486079973 \times 10^{-19} \ll 0.0001$. The result of the ADF test is 99.99% certainty to reject the null hypothesis, that is, 99.99% certainty that the original sequence has no unit root. In other words, the original sequence is a stationary sequence, and models such as ARMA can be used.

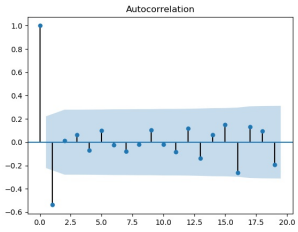
At the same time, the value of the statistic is -0.82801362467521, which is less than the 1% critical value -3.5194805351545413, which also shows that the null hypothesis can be rejected with 99% certainty. The original sequence is a stationary sequence, and ARMA and other models can be used.



(a) Partial autocorrelation of vector_lng of golden digger wasp



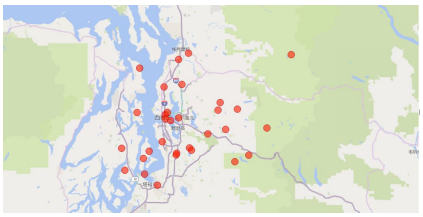
(b) The sequence diagram of vector_lng of golden digger wasp



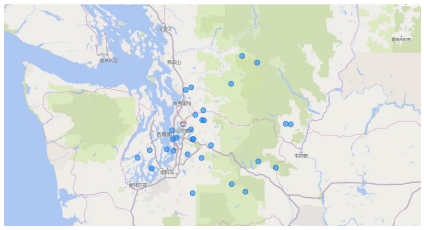
(c) Autocorrelation of vector_lat of golden digger wasp

Figure 7: Changes of parameters in the original sequence over time

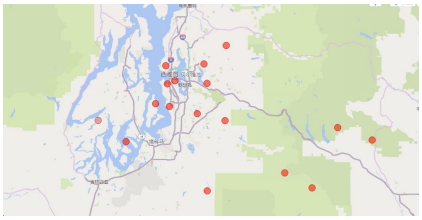
Through the time series, we can predict the migration vector status of n+1 month based on the data before n months. After running the model, we can find the migration vector and actual value of the species with the label golden_digger_wasp. It is basically consistent, and the actual expansion direction of the colony is also basically consistent with the predicted direction.For example: from July to August, the overall activity center of golden_digger_wasp has a clear trend of expanding to the southwest.



(a) August_golden_digger_wasp



(b) August predect

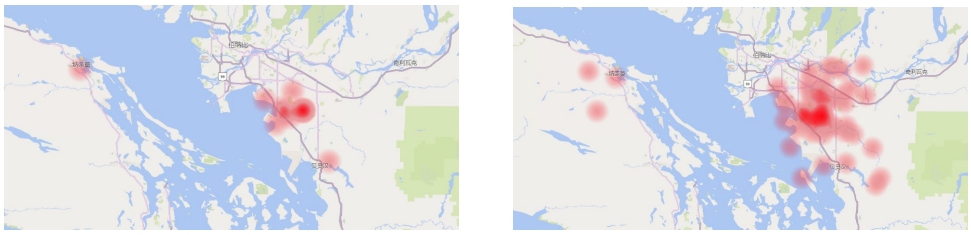


(c) July_golden_digger_wasp

Figure 8: Golden digger wasp expansion from July to August

Model results

One year after the time of the latest report, we can find that when the government or citizens did not take measures to restrict wasps, Vespa mandarinia expanded very rapidly. Their population increased from 14 to 55, and the number of Vespa mandarinia. The breeding area of the herd does not move significantly, but spreads around.

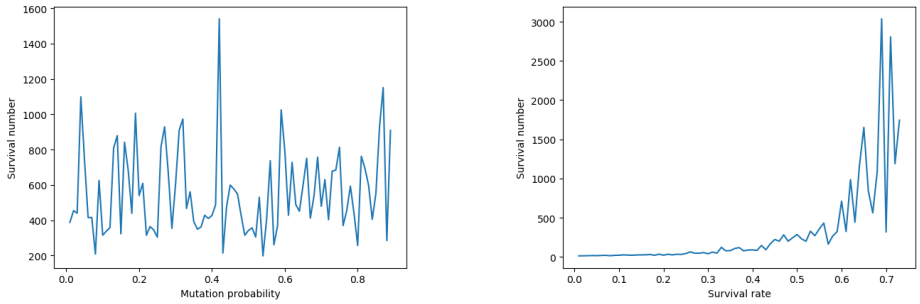


(a) Real data (b) Forecast data

Figure 9: Predict the outcome of Vespa mandarinia

Sensitivity analysis

In this part, we tested the sensitivity of MRV model in survival rate, mutation rate and reproduction number.



(a) (b)

Figure 10: Sensitivity analysis

We iterate the mutation rate from 0.01 to 0.9 with the step size of 0.01, and get the corresponding number of reproduction each time. Through figure 1, we can find that the mutation rate has no obvious effect on the number of reproduction.

Next, we iterate the survival rate from 0.01 to 0.74 in steps of 0.01, as shown in Figure 2. It is found that the survival rate plays a decisive role in the reproduction quantity. When the survival rate is greater than 0.4, the reproduction quantity will increase sharply.

4.5 Model update and extinction of Vespa mandarinia

Model update

Through continuous testing, we iterated the values of p and q in the ARMA model, and found that when p=3 and q=3, the prediction effect of the ARMA model is the best.Through our

observations, when the time interval exceeds 3 months, the ARMA forecast will have serious problems, and the forecast results will deviate too much from the actual results.

Extinction of Vespa mandarinia

We believe that the conditions for judging the survival of Vespa mandarinia are as follows:

1. From March to mid-November of the year, if Vespa mandarinia has not been detected for more than three consecutive months and less than six months, it can be considered that its number has decreased.
2. 2. From March to mid-November of a year, if Vespa mandarinia has not been detected for more than six consecutive months and less than one year, it can be considered as scarce.
3. 3. Vespa mandarinia is not detected throughout the year, which means that it is on the verge of extinction.
4. 4. If Vespa mandarinia is not found for three consecutive years, it is considered that it has been eliminated.

At the same time, according to our model, when the survival rate is only 0.05 and the mortality rate is as high as 0.95, Vespa mandarinia will not be able to reproduce and will gradually die.

5 Model II: VTP-Event screening model based on vision, text and place

Among the thousands of incident reports given, only 14 laboratories rated the results as positive, and most of the reports came from erroneous monitoring. By analyzing the Notes, location latitude and longitude and image data contained in these reports, we developed a VTP model to classify sightings and help screen out the most likely positive sightings from the new reports.

5.1 Establishment of model

5.1.1 Image Identification

With the development of computer science and technology, deep learning is becoming more and more popular in various scientific research fields, especially with the improvement of hardware computing power, deep learning can not only ensure accuracy but also increase recognition efficiency, and greatly reduce human consumption. In order to filter out the most likely sighting reports through image data, we choose deep learning methods as the basis for image recognition.

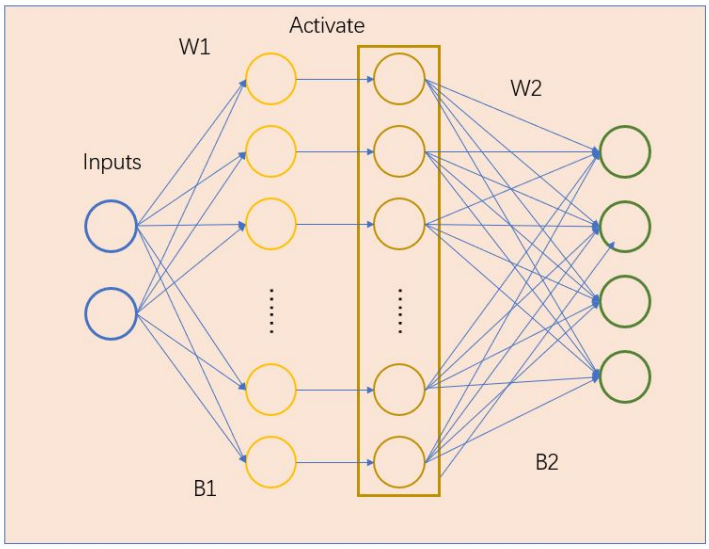


Figure 11: Neural Networks

Neural network is one of the most basic tools in deep learning. A simple layer of neural network can be expressed as:

$$H = Act(X * W_1 + b_1) \tag{6}$$

Among them, Act function represents the activation function.

The convolutional neural network performs well in the field of image processing, and we show its structure as follows:

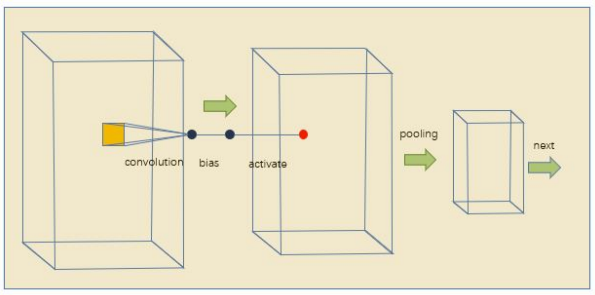


Figure 12: Convolutional neural network structure

It can be expressed in mathematical form as:

$$cons_{x*y} = \sum_{i}^{m*v} w_i * v_i + bias \tag{7}$$

In this question. The problem that we need to solve is related to insects. The individuals are small and the characteristics are not obvious enough. Therefore, we choose two-stage neural network for image recognition. Take the faster r-cnn network as the main network for image detection. The overall structure of the neural network built by the team is as follows:

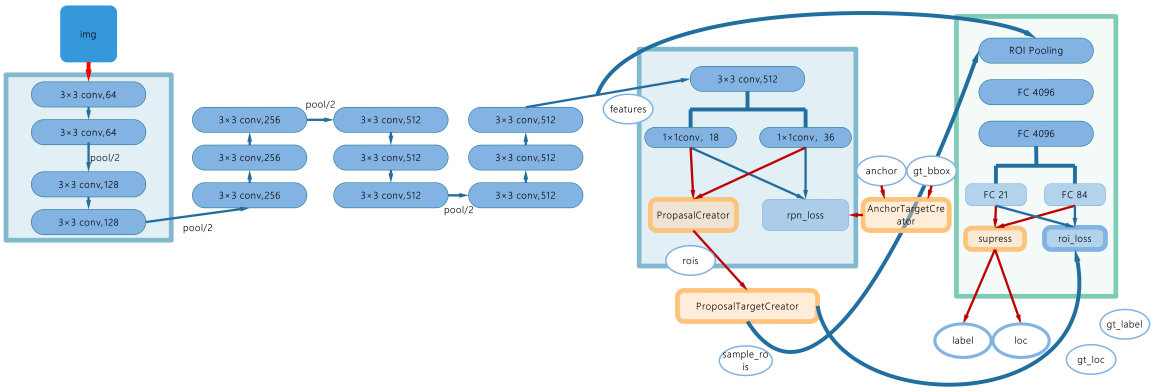


Figure 13: Image recognition and detection diagram

In the Conv layers section, we chose Resnet50 as the backbone feature extraction network. Compared with VGG16[5], Resnet50[6] solves the problem of gradient disappearance through the residual network structure, which can deepen the number of network layers. Through label extraction in data preprocessing, we have extracted a total of 12 categories of insect types, namely:

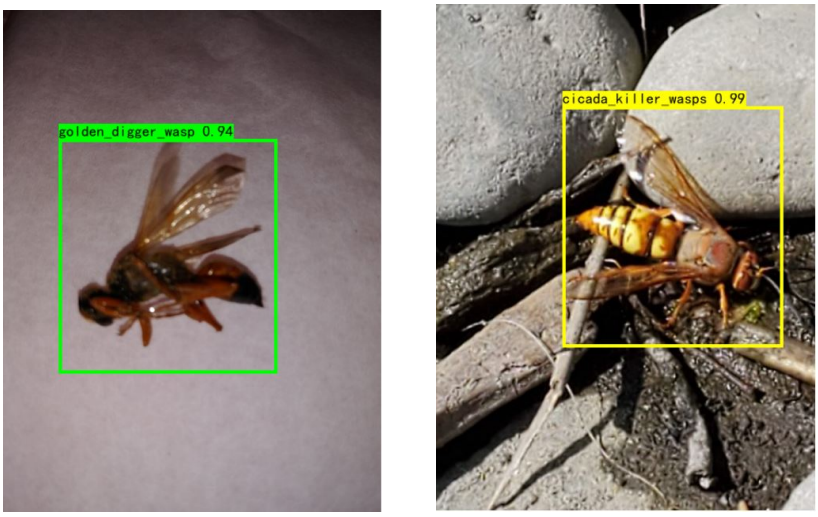
Table 4: Types of insects extracted

beetle	Bumble_bee	Cicada_killer_wasps	Fly
Golden_digger_wasp	Jerysalem_criket	Paper_wasp	Sawfly
Vespa_mandarinia	Wood_wasp	Yellw_jacket	other

Among them, many events with less data are summarized in other, and even no conclusion is given in Lab Comments.

Perform data enhancement on the image data of these 12 categories, add noise interference, cropping, rotation, brightness increase and decrease, etc., and put them into the neural network for training to obtain a trained neural network. In the training process, we freeze part of the network for training first, and then expand all the networks for training, so that the network training effect is better. Perform image recognition on the data provided by the user. On the one hand, it returns the correct classification result to the user. On the other hand, we will get 12 categories of scores and use them as one of the basis for the final score.

The following is an example of the results of the trained neural network when recognizing different pictures.



(a) (b)

Figure 14: Schematic diagram of image recognition and detection

$$vision\_scores = [vs_1, vs_2, vs_3, vs_4, vs_5, vs_6, vs_7, vs_8, vs_9, vs_{10}, vs_{11}, vs_{12}] \tag{8}$$

5.1.2 Text extraction

In some reports, citizens did not provide image evidence, but only text descriptions, and there were a large number of invalid images. Because the sharpness was too low or the object being photographed could not maintain the original state, it could not be used as a basis for judgment.Therefore, it is necessary to process and use the Notes part of the report.

Summarize the text features of the 12 classes, then summarize the relevant words in the Notes section, and calculate the corresponding scores.

Table 5: Text features of Vespa mandarinia

Live hornet captured by WSDA staff	citizen scientist	yellow heads
black thorax	Brown striped abdomens	Black striped abdomens
Yellow striped abdomens	underground	skin necrosis and hemorrhag

We count the keywords in Notes, and calculate its text_scores according to count. Similarly, text_scores also has 12 scores.

Table 6: Correspondence between count and score

Count	Score
0	0
1	0.1
2	0.3
3	0.7
>3	1

$$text\_scores = [ts_1, ts_2, ts_3, ts_4, ts_5, ts_6, ts_7, ts_8, ts_9, ts_{10}, ts_{11}, ts_{12}] \tag{9}$$

5.1.3 Place score

Most of the creatures appearing in the report are social creatures and have their own fixed areas of activity. This also means that if a certain type of bee is often found at point A, then the new report claiming that it was found at point A is more credible than claiming that it was found in other places far away from point A.

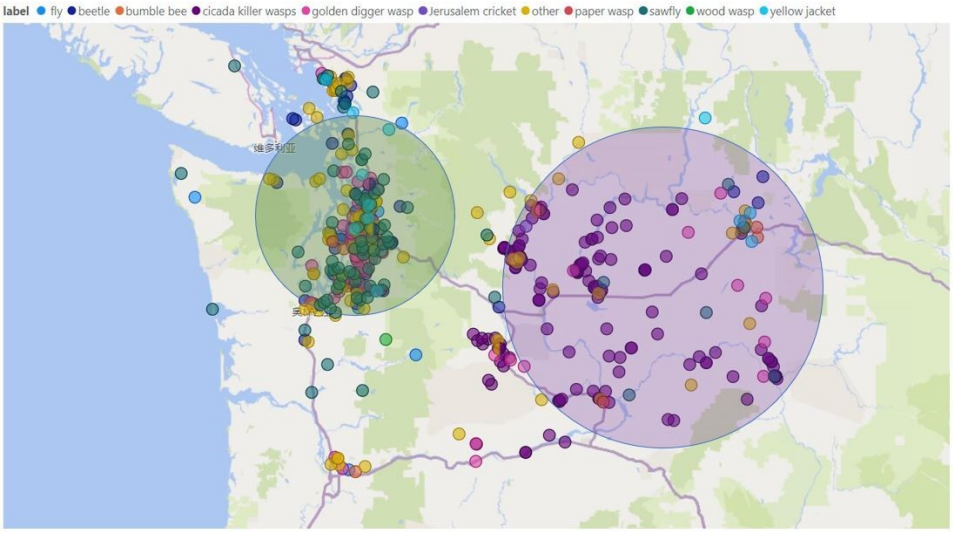


Figure 15: Where all the insects were found

For example, golden digger wasps and cicada killer wasps prefer to be in the southeast, while wood wasps are more concentrated in the northwest.

Therefore, we obtain place_scores from the average position P_bee_i where the 12 kinds of creatures appear in the report, by comparing the distance between the reported occurrence location and the average position.

$$place\_scores = [ps_1, ps_2, ps_3, ps_4, ps_5, ps_6, ps_7, ps_8, ps_9, ps_{10}, ps_{11}, ps_{12}] \tag{10}$$

$$ps_i = \frac{1}{\sqrt{1 + (lat - P\_bee_i * lat)^2 + (lng - P\_bee_i * lng)^2}} \tag{11}$$

5.1.4 Model integration

We used image data, text data, and location data to score the report in three dimensions. Based on the three scores, a certain weight ρ_i is assigned to establish a VTP model. The scores of the three dimensions here are all $\frac{1}{3}$.

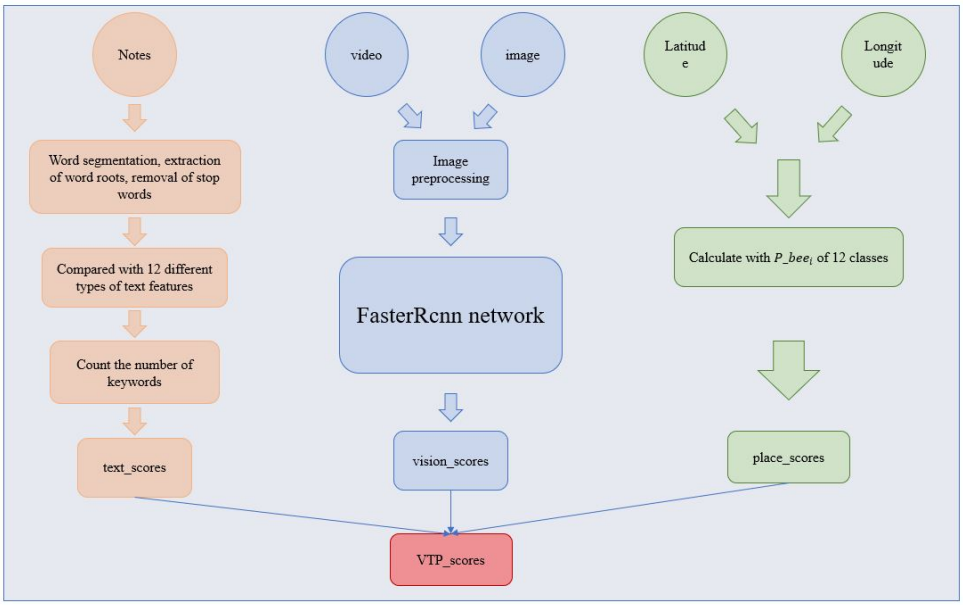


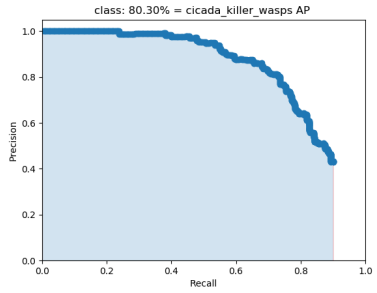
Figure 16: The structure of the VTP model

5.2 Model application and analysis

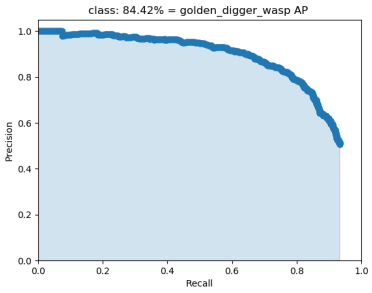
The test of Image recognition

Among the three indicators used as reference in the VTP model, the image recognition part is the most complicated and has a higher error rate. We set a series of parameters to evaluate and analyze the results of image recognition. According to some parameters of the structure, we evaluate the accuracy of the model in the image recognition process, and visualize the parameters.

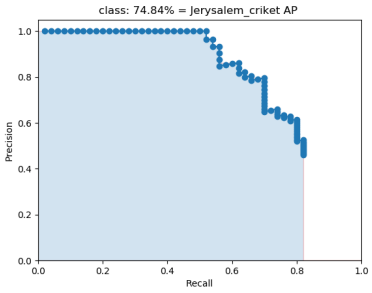
1. **P-R curve**
   Precision refers to the correct percentage of the predicted number, and recall refers to the number of all correct samples found. Generally speaking, when the precision is high, the precision is often low; when the precision is high, the precision is often low. By plotting the precision on the vertical axis and the recall rate on the horizontal axis, we get the precision-recall rate curve, referred to as "P-R curve".



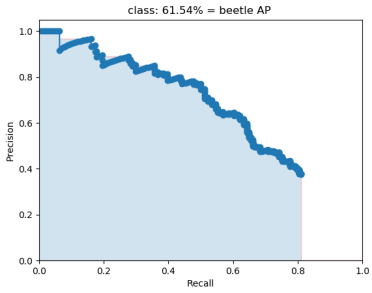
(a) cicada_killer_wasps



(b) golden_digger_wasp



(c) Jerysalem_criket



(d) beetle

Figure 17: AP of all categories

2. **IoU**
   IoU is the ratio of the intersection and union of the predicted frame and the true sample frame. Assuming that the IoU corresponding to the prediction box is greater than a certain threshold (generally, the commonly used IoU threshold is 0.5), we can say that the prediction box is correct and can be divided into TP(True Positive). On the contrary, if IoU is less than the threshold, then the prediction box is wrong, or it is an FP(False Positive). Sample frames that are not detected are included in FN(Flase Negative).

3. **AP and mAP**
   AP: The approximate area under the PR curve is a value between 0 and 1. It can also be used to measure the effect of the model. If the AP of the model is larger, the model is better, which makes it easier to compare different models. For each category, calculate the AP according to the above method, and take the average of the AP of all categories to be mAP.

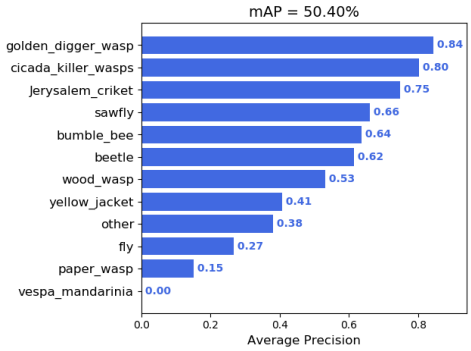
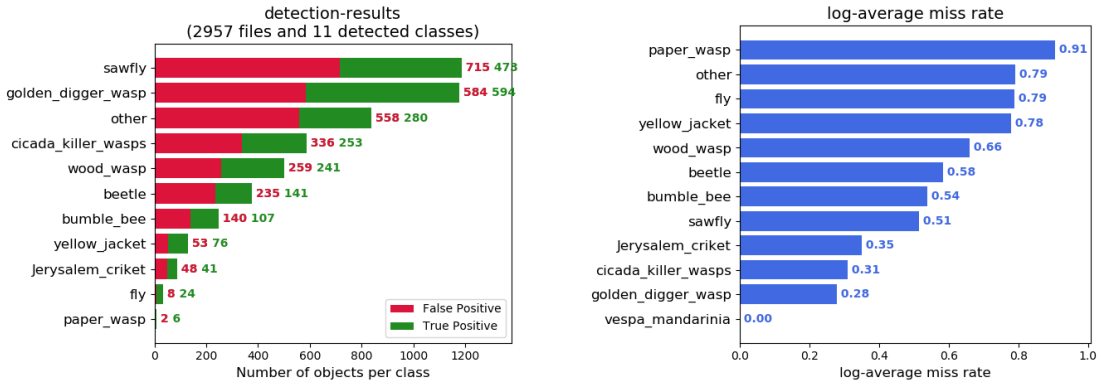


Figure 18: AP of all categories

4. **log-average miss rate**
   The log-average missed detection rate is used to measure the performance of the prediction model and is the logarithmic average of the missed detection rate of the predicted sample and the real sample.



(a) detection-results-info (b) log-average miss rate

Figure 19: Detection results and log-average miss rate

By verifying the image recognition module and referring to the constructed parameters, we can draw the following conclusions:

- The overall mAP value of image recognition is 50.40%, indicating that the prediction effect of the model is better.
- The model has the best prediction performance for golden_digger_wasp: this tag has the largest mAP value and the smallest log-average miss rate.

  The ground truth of jerysalem_criket is less but the log-average miss rate is small and the mAP value is large. The image recognition quality of this label is also excellent. Preliminary judgments may cause this phenomenon to have two reasons: one is that there are more ground truths due to the randomness of the report, and the sample clarity is high; the other is that the samples of this category have strong differences, obvious characteristics, and easy to distinguish.

  Normally, when the log-average miss rate of the sample category is high, the mAP will be low; when the log-average miss rate is low, the mAP will be high.
- When the ground truth of most sample categories is high, mAP will be high. But there are exceptions. Since the samples in the other category are samples with unobvious characteristics and excluded from the clear category, the overall characteristics are not concentrated, so even if the ground truth of the other category is high, its false detection rate is still large and the mAP value is small. This is normal.
- The prediction result of vespa_mandarinia, the research objective of this question, is 0. It is speculated that the model may underfit the prediction of this type of sample due to the insufficient number of samples in the existing training set.

5.3 Effective recommendation strategy

Of the more than 4000 reports, only 14 were identified as positive, which was mixed with a large number of irrelevant data. In order to help the staff of Washington Department of agriculture screen data, we established the following screening strategies based on VTP model:

1. Firstly, the notes part of the report is extracted and the text is calculated_Scores. If the report is submitted by wsdastaff and citizen scientist, it will be pushed directly to the staff of the Washington Department of agriculture.

2. Image recognition, calculate vision_scores. In the image recognition part, because Vespa mandarinia has too little image data, the neural network has poor training effect on Vespa mandarinia, so we only need to eliminate the events with scores of categories other than Vespa mandarinia vs_i greater than 0.5, and the rest Report retention.

3. For the data retained in the image recognition part, the place_scores are calculated by using the latitude and longitude of the time and place when uploading.

4. Calculate their VTP_scores on the retained data, and recommend the higher scores to the staff of the Washington Department of Agriculture.

6 Strengths and weaknesses

Strengths

- Based on a correct and clear understanding of the meaning of the questions, this article uses reliable mathematical and statistical methods, establishes a scientific and reasonable model, and conducts rigorous verification, with a solid and reliable mathematical foundation;

- The model established in this paper is closely related to the actual situation. For example, the solution of this paper refers to the real life cycle and reproduction mode of wasps, and has high generalization and universality;

- This article contains a series of solutions for problem solving and conclusions. Images can describe and solve problems more intuitively.

Weaknesses

- Because the number of correct samples of the research object vespa_mandarinia is too small, there is a phenomenon that no such samples are recognized during image recognition, and the model is under-fitting

- The sample area needs to be marked manually before image recognition, which results in large amount of work, time consumption and high cost of manpower and time;

7 Memo to WSDA

Memorandum

To: WSDA

From: MCM Team#2106561

Subject: An Unnecessarily Complicated Title

Date: February 9, 2021

In September 2019, Canada reported an incident of the discovery of Vespa mandarinia in its territory. In December of the same year, traces of Vespa mandarinia were discovered in Washington State, USA. As an invasive biological species, Vespa mandarinia will cause great damage to the local agriculture and ecological environment. At the same time, it has also caused middle-aged men to be attacked or even killed. For this reason, WSDA requires the public to report the sightings of Vespa mandarinia in time for pest control Threat. If WSDA cannot take timely measures to curb its development, then in 2021 Washington State will have an additional square kilometer of affected area. Our team has established two models to predict the reproduction status of Vespa mandarinia, classify and pre-process related incident reports from the public, and give specific measures to eliminate Vespa mandarinia.

Through the MRV model established by us, we can know that the number of bee colonies will increase from 14 to 55 in one year without any treatment to the invading Vespa mandarinia, and the total affected territory area of the United States is as high as 4300 square kilometers.

But the government and citizens are not waiting to die. Citizens will actively report to wsda after witnessing suspected creatures. According to citizens' report and further investigation, the current situation of Vespa mandarinia in the United States has been basically understood. New reports are pouring in, but it's too difficult for a limited staff to deal with such a large number of reports. In order to deal with this problem, we developed a VTP evaluation model based on image, location and text. After the preliminary screening of the reports, some reports with high probability were presented to the staff, which greatly saved the limited time. But the more difficult problem is to prevent the breeding and expansion of bee colony. According to our MRV model, the colony is always close to the most suitable environment when it expands. Therefore, we can find the most likely place to produce a new colony around the 14 known locations of wasps.

Proposal

- Vespa mandarinia's colony leaves only the queen in the winter, when the defensive and aggressive of the whole hive are quite low. Therefore, winter should be the key time to clean up wasp hives, and at the same time, we should also pay attention to thoroughly inactivate the residues of the hives at high temperature.
- Through the MRV model, we can roughly predict the geographical location of the next new colony. In order to be able to observe and deal with the hives in time for a long time, we can set up temporary workstations in the places with high density of new colonies, which are used to store the tools to eliminate pests and as the assembly point of workers.
- To strengthen the publicity and education of the public, although the number of reports that staff need to judge has been greatly reduced with the help of the system, there will still be a

small number of unqualified reports through the system. Therefore, it is very important to enhance the public's knowledge of biological invasion prevention, not only for the protection of the country's ecological security, but also for the people to make the right choice when facing wasps instead of being injured.

- Strengthen the national restrictions on the import of goods, especially wood and other "dangerous goods" that are easy to carry exotic species into the country.

References

[1] Washington State Department of Agriculture. 2020 Asian Giant Hornet Public Dashboard.

[2] Wolfram, Stephen. A new kind of science. Vol. 5. Champaign, IL: Wolfram media, 2002.

[3] Rortais, Agnes, et al. "A new enemy of honeybees in Europe: The Asian hornet Vespa velutina." Atlas of Biodiversity Risksfrom Europe to globe, from stories to maps. Sofia & Moscow: Pensoft 11 (2010).

[4] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." arXiv preprint arXiv:1506.01497 (2015).

[5] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[6] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016