Problem Chosen C | **2021 MCM/ICM Summary Sheet** | **Team Control Number** 2124506

Warning: *Vespa Mandarinia* Invasion

Summary

Vespa mandarinia (Asian giant hornet) is reported to invade Washington State and a mass of sightings have been submitted. In this essay, we will provide a series of models for analyzing public reports and prioritizing them.

Firstly, we construct spread migration model to predict *V. mandarinia's* spread. Thereinto, **Monte Carlo Simulation** is applied to calculate the probability of each divided region. The most possible area for the colony to arrive has more than 65% probability. The model then forecasts *V. mandarinia*'s colony locations in 2021 based on the given reports. Finally, colony is predicted to arrive in **Lynden City** in December 2021. The result illustrates that agriculture department has to respond actively.

Secondly, a **Convolutional Neural Networks (CNN)** is applied to recognize reported images. At first, images connected with text information by Global ID is carefully preprocessed. As positive cases are very limited, several transformation methods are used to expand dataset. Classic AlexNet structure is chosen. Finally, the model has about **85%** accuracy in testing dataset, which helps classify right reports from mistaken ones. Therefore, further images could be classified automatically.

Thirdly, we provide **Prioritizing Model** targeting image, position and text in a report**.** At first, if the image is positive, the case will be further investigated. Or, its rest targets will be scored respectively. For position, the distance between reported location and predicted location is calculated. For description text, word clouds provide positive and negative words list. **Parameter identification** is used to decide weights for score compile. If the score reaches 85 of 100, this report will be investigated manually.

If a new case is confirmed, our prediction model will update using **Bayes' Statistics**. When positive case is far from our estimated position, new case's position is chosen as new origin of our model. If it is in the range of our estimation, we will calculate newer conditional probabilities inside the most probably 4×4 block and update most possible cell. An example is given with the most possible cell moving.

At last, to judge *V. mandarinia*'s eradication, here we use **hypothesis testing**. Poisson distribution with expectation value (10) describes the spread of *V. mandarinia* in 2 years. If there are no more than 2 cases reported in 2 years, *V. mandarinia* is eradicated in Washington State with confidence coefficient of 99.5%.

A brief sensitive analysis and discussion is given at the end of essay as well.

Keywords: Monte Carlo Simulation, Bayes' Statistics, CNN, hypothesis testing

Contents

1 Introduction

1.1 Problem Restatement

It is reported that a nest of *Vespa mandarinia* on Vancouver Island, Canada (Asian giant hornet) was destroyed. *V. mandarinia* is a cruel predators and invaders who prey on European honeybees as well as damage their nests, and it also attack other insects.

In view of the destructiveness of *Vespa mandarinia*, the State of Washington has made measures as helplines and a specific website for the public to report their sightings of it. Our team is required to explain the data collected from witnesses and consider a strategy to make a best use of the public reports by prioritizing them. And to deal with this problem, we divide it into several sub-problems as below:

1. Predict the spread of *Vespa mandarinia* and state the accuracy of its prediction.
2. Classify *V. mandarinia* from other hornets and predict likelihood of mistakes.
3. Set priority for investigation.
4. Update our model with added reports and indicate updating frequency.
5. Analyze what situation *Vespa mandarinia's* vanishment can be proved

1.2 Problem Analysis

In consideration of the invasion of *V. mandarinia* in Washington, this article will focus on the analysis of reported data and the prioritization of the reports.

First of all, to predict the spread of the colony and address its accuracy, we need to build its spread prediction model. The first step is to divide the area into cells and predict the probability after moving of each cell. After figuring out the probability of each direction, we can use Monte Carlo Simulation to simulate colony's movement.

Secondly, given data and images should be used to classify right reports from mistaken ones. We hire CNN to classify images of *V. mandarinia* from other species'. As the positive sample is much smaller than the negative, several image transformations are used to expand it.

Thirdly, we indicate which reports are worthy to investigate by their image, position and text. As for image, once it is confirmed to be authentic, the report will be checked. As for the position, we will score a case by its distance from our estimated position. And for text, it will be scored with a corpus. Then, only if the weighted score is higher than a threshold value, can we believe that it needs manual check.

Fourthly, to update the model and address its frequency, we are going to combine new positive reports with predictions. The locations in reports are treated as prior distribution, with the maximum likelihood method, we can modify our models. Then, for the frequency, we can deal with reports as a whole during a limited time span.

Finally, we are required to find evidence verifying the extinction of *V. mandarinia*. The fewer new positive cases are, the less possible of its existence. When the number of new positive cases is less than a set value, it rejects our hypothesis that pest subsists.

2 Assumptions and Justifications

In order to simplify our models and combine with the reality, we set a series of assumptions listed below:

1. **The public will report immediately.** The public has paid much attention to *V. mandarinia* and it helps to determine the accurate date.
2. **The colony will spread as a whole.** Because of the habits of *V. mandarinia*, it is reasonable.
3. **Reports by the public are based on the reality.** It is rare for the public to report fake information on purpose.
4. **There is no a new colony appearing in Washington State.** We assume that the government will take some measures to prevent new colonies to enter.
5. **The hornets in the colony won't surge.** The rate of reproduction of this pest isn't high and some effective methods will be used to control its number.

3 Notations

Symbols	Descriptions	Units
L	The unit length of a cell divided in the map	m
v	The speed of the colony	$m/month$
dT	The time unit	day
w	The weight between the input and the hidden layer	
$score(d)$	The score of the distance	
$score(w)$	The score of the words in descriptions	
n	The number of new positive reports in the 2 years	
N	The threshold value in the 2 years	

4 Spread Prediction Model

Vespa mandarinia, as known as Asian giant hornet, is a kind of hornet and it shares some common attributes with other hornets. It has strong territory consciousness and it prefers to move around its hive. However, *V. mandarinia* is more aggressive and destructive [1]. Considered as an invasive species, some proofs assert that *V. mandarinia* has migrated to Washington State. In this case, it is necessary to supervise its migration.

4.1 Kinematics of Migration

As for the spread of *V. mandarinia*, we deem that the swarm will move as a whole. Some of hornets estrange from the colony and fly to other directions, which explains why some reports illustrated that *V. mandarinia* was witnessed in different regions at the approximate same time. The figure below displays some positive case records reported by publics while two outliers are ignored in this diagram. It is obvious that the colony has a tendency of moving southeast:

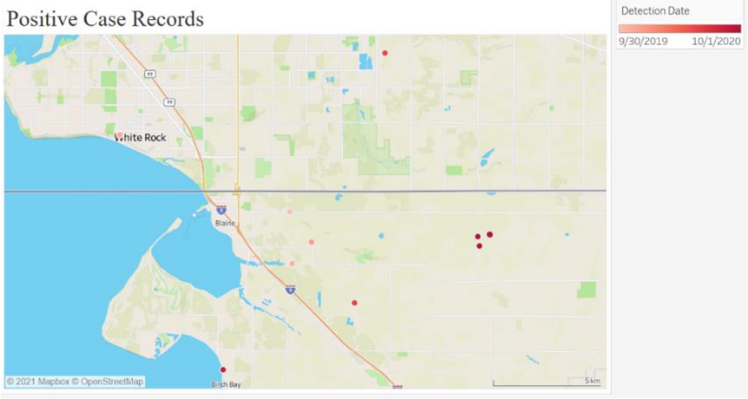


Figure 4.1 Positive Case Records

Subsequently, as the speed of the colony is determined, the condition to update its site can be figured out:

$$\begin{cases}(dx, dy) = v \times dT \\ (x_t, y_t) = (x_{t-1}, y_{t-1}) + (dx, dy)\end{cases} \tag{4.1}$$

where v is the speed of the colony, dT is the change of time interval T, dx and dy are the change in the X coordinate (East-West) and Y coordinate (North-South) respectively, t is a particular time of location change, $t-1$ is the time before time t, (x_t, y_t) is the coordinates at time t.

4.2 Speed and Active Area

In light of given attachment, *V. mandarinia* has moved 1.8 km in 2 months [2]:

$$v = 900\ m/month \tag{4.2}$$

where v is the speed of the colony.

And we divide the map into cells. The time taken for the colony to travel across a cell of the table is approximately as below:

$$dT = \frac{L}{v} = 6\ days \tag{4.3}$$

where L is the cell's length, and we take 6 days as a time unit (dT).

Based on the time unit (6 days) and 30 days in a month, we deduce that the colony will travel in a period of 5 time units per month. And because the colony can travel across a cell of table within every time unit, the possible active area then can be depicted as a table with 11 columns and 11 rows, which means that the colony of Vespa mandarinia may appear in any of the 121 cells with different probability.

4.3 Rule of Spread

Now that we have determined the scope of activity, we need to figure out the rule of its move within the possible range.

As for the travel direction, for instance, we focus on one dimension (south and north), and the positive direction is south. We deduce the equation to depict the probability among the three conditions as below:

$$\alpha_1 + \beta_1 + \gamma_1 = 1 \tag{4.4}$$

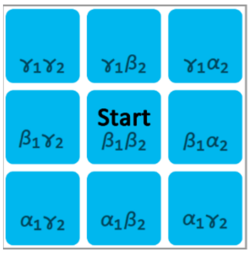
where α_1 means the probability of moving south, β_1 means that of moving north and γ_1 means that of staying stable in the south-north dimension.

Similarly, as for west-east dimension, the positive direction is east and the equation of probable move is as below:

$$\alpha_2 + \beta_2 + \gamma_2 = 1 \tag{4.5}$$

where α_2 means the probability of moving west, β_2 means that of moving east and γ_2 means that of staying stable in the west-east dimension.

Furthermore, we consider the potential that the colony will move to all directions, so, the move along the south-north dimension and the west-east dimension are independent. After that, we infer the probability of the move of the colony and it is shown as below:



$\gamma_1\gamma_2$	$\gamma_1\beta_2$	$\gamma_1\alpha_2$
$\beta_1\gamma_2$	Start $\beta_1\beta_2$	$\beta_1\alpha_2$
$\alpha_1\gamma_2$	$\alpha_1\beta_2$	$\alpha_1\gamma_2$

Figure 4.2 Probability of the Regions Around

The diagram shows the probability at each block after once movement.

4.4 Distribution Probability

In this segment, the possible active area is the table with 11 columns and rows. We deem that the colony will move along the two dimensions. We employ **Monte Carlo Simulation** [3], and set that it will move 5 times in each iteration. Codes are attached in Appendix A. Then, using frequency to replace probability and we obtain the probabilities of the destination for the colony from its central point as the start. The frequency of a hornet's move can represent the probability of the whole colony.

Additionally, since the parameters of probability mentioned in 4.3 are not set, we take data of locations from attachment as prior distribution. Then, we test a series of probabilities of moving towards each direction to indicate true value. Finally, the parameters fitting the prior distribution are determined as below:

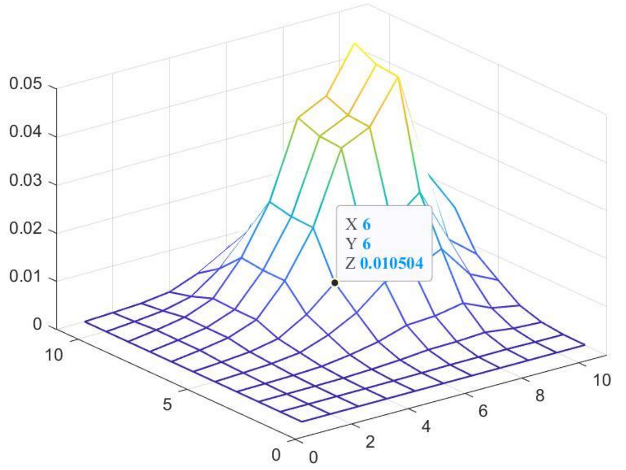
$$\begin{cases} \alpha_1 = 0.4\ (South) \\ \beta_1 = 0.2\ (North) \\ \gamma_1 = 0.4\ (Stable) \end{cases} \tag{4.6}$$

The parameters in equation 4.6 represent those in equation 4.4 (south-north dimension).

$$\begin{cases} \alpha_2 = 0.7\ (East) \\ \beta_2 = 0.1\ (West) \\ \gamma_2 = 0.2\ (Stable) \end{cases} \tag{4.7}$$

The parameters in equation 4.7 represent those in equation 4.5 (west-east dimension).

Therefore, we obtain the distribution of destination as follows:



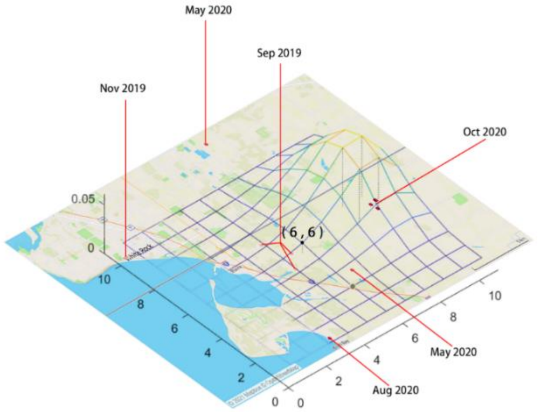


Figure 4.3 Probability Distribution

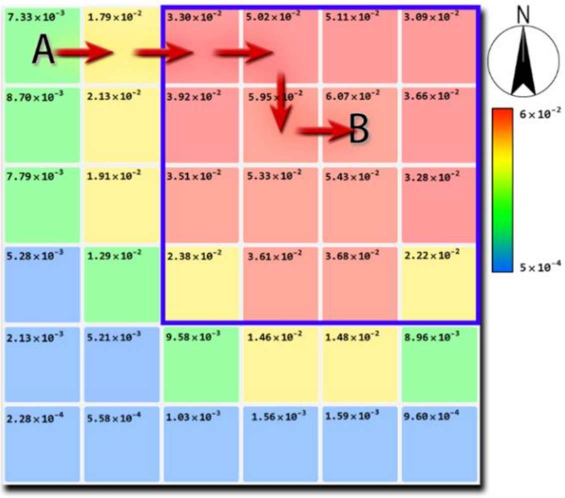
Figure 4.4 Probability Coordinate System

For Figure 4.3, the X-axis represents the east-west dimension. The Y-axis represents the south-north dimension and Z-axis represents the probability at (X, Y). Moreover, (6, 6) is the origin of the colony.

For Figure 4.4, the dots labeled with dates represent the positive cases. And the area with high probability is the same as real case reported in Oct. 2020, which verifies that parameters meet the requirements of the prior distribution.

4.5 Migration of the *Vespa Mandarinia* Colony

Based on probabilities of destination cells determined in Char. 4.4, we merge the regions with relative higher probabilities into a 4×4 region which is as follows:



7.33×10^{-3}	1.79×10^{-2}	3.30×10^{-2}	5.02×10^{-2}	5.11×10^{-2}	3.09×10^{-2}
8.70×10^{-3}	2.13×10^{-2}	3.92×10^{-2}	5.95×10^{-2}	6.07×10^{-2}	3.66×10^{-2}
7.79×10^{-3}	1.91×10^{-2}	3.51×10^{-2}	5.33×10^{-2}	5.43×10^{-2}	3.28×10^{-2}
5.28×10^{-3}	1.29×10^{-2}	2.38×10^{-2}	3.61×10^{-2}	3.68×10^{-2}	2.22×10^{-2}
2.13×10^{-3}	5.21×10^{-3}	9.58×10^{-3}	1.46×10^{-2}	1.48×10^{-2}	8.96×10^{-3}
2.28×10^{-4}	5.58×10^{-4}	1.03×10^{-3}	1.56×10^{-3}	1.59×10^{-3}	9.60×10^{-4}

Figure 4.5 Probabilities in Potential Area

The 4×4 region framed by blue is the most likely area for colony to arrive. And the total probability of the whole framed region is up to 65%.

We utilize **Monte Carlo Simulation** to update the location of the colony, which is denoted as the central point of it monthly. The migration path of the colony will be depicted. *V. mandarinia*'s habits are taken into consideration as well. *V. mandarinia*'s mobility will experience a low period for about 3 months in winter [4], hence, we deem that the colony tends to stay in the same place in the period. Thus, we adjust the parameters to reduce the predicted migration during the period and add a random term to decide whether they move.

Later on, we operate the prediction. The destinations of the colony's migration in each month are shown in the following figure:

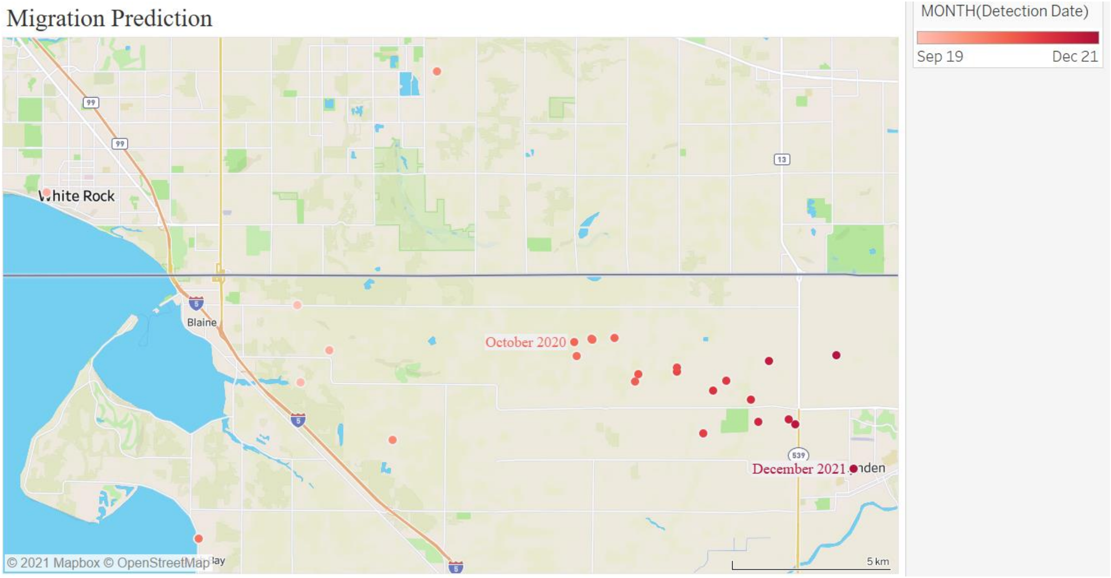


Figure 4.6 Migration Prediction

The diagram shows the migration in the coming 12 months and some reported positive cases. The darker the point is, the later the time is. The most recent case reported in Oct. 2020 (48.9834° N, 122.5825° W) is labeled. From our model, the colony has a tendency to move southeast. Finally, colony will arrive at Lynden city at the end of 2021.

5 Image Classification Model

5.1 Data Preprocessing

To begin our image recognition model training, disordered data must be preprocessed. Data is in a form of multi-model which includes images, videos and text, so we establish the connection between different model data through Global ID.

First of all, we preprocess the image data. The .xlsx file indicates that there are 15 positive reported cases, and the negative and unverified ones have 1,243 and 1,351, respectively. Files with wrong formats are checked and possible images inside are extracted manually. Thus, directly processing existing image data will cause serious data imbalance. We expand the positive images by data expansion, including image inversion, offset, color inversion, and noise addition. Code will be presented in Appendix B. By means of data augment, the positive images are expanded to 1,045.

Secondly, we tackle with the video data. In this process, we intercept the video data into multiple different frames by taking screenshots, and then convert them into image data for processing. Some of the intercepted data is shown as follows:

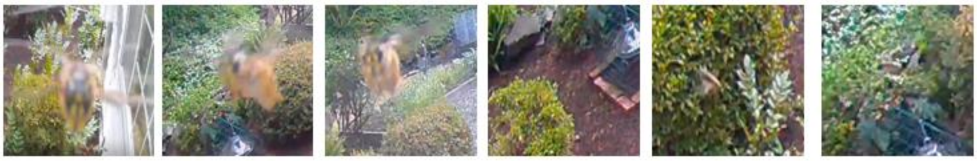


Figure 5.1 Part of Intercepted Data

Eventually, we deal with the text data, and since the text data is a textual description of the image data and video data, we establish the relationship between them with Global ID.

5.2 Image Recognition

CNN is evolution of Artificial Neural Network (ANN) by replacing matrix multiplication with convolution. We will introduce ANN and then explain how it transforms into CNN. Firstly, we have a vector formed by m inputs, the ANN has n nerve cells, one final outputs. ANN does the calculation on inputs as follows:

$$z = xw^T + b_0, w \in R^{N\times M}, b_0 \in R^{1\times N} \tag{5.1}$$

where w is the weight between the input layer and hidden layer.

Nerve cells has their own constant bias representatively and form a bias vector b_0. When the hidden layer nerve cells receive the vector z, they will apply an activation function:

$$\sigma(z) \in R^{1\times N} \tag{5.2}$$

$\sigma(z)$ is the activation function. Common options include $sigmoid, tanh$ and $ReLu$.

Then, the output layer will do calculation:

$$y = \sigma(z)\theta + b_1 \in R^{1\times 1} \tag{5.3}$$

where θ is weight and b_1 is bias vector with the same size as output.

To train the model of ANN, we have to do backpropagation which means that calculate forward to get the loss and do differential backward to get the gradient. Firstly, we define a $Loss\ Function$ here to explain how to train the model:

$$L(w, b_0, \theta, b_1) = \sum_{i=1}^{K} \frac{1}{2}[N(x_i; w, b_0, \theta, b_1) - y_i]^2 \tag{5.4}$$

Then, we update the ANN's parameters so that the loss function decreases after that. We apply gradient descent method here so we do differential with each parameter of ANN to loss function and update our weights:

$$w \to w - r\frac{\partial L}{\partial w}, b_0 \to b_0 - r\frac{\partial L}{\partial b_0}, \theta \to \theta - r\frac{\partial L}{\partial \theta}, b_1 \to b_1 - r\frac{\partial L}{\partial b_1} \tag{5.5}$$

where r is the learning rate of the model and need to be chosen by people.

Now we will talk about how CNN is different from ANN. CNN uses a filter of $n \times n$ to move and do convolution along the way to reduce the size of image and extract the characteristics. Moreover, pooling helps the model to go further by extracting key information in one area only [5].

The model structure is a simple CNN structure, AlexNet model [6]. It adds an activation function, ReLU after each convolution layer to avoid the vanishment of gradient and increase the speed of convergence. Random dropout helps the model to ignore specific nerve cell to avoid overfitting. Moreover, Local Response Normalization layer helps to improve model's precision as well. Additionally, overlapping max pooling is another impressive design of AlexNet.

The images sequence is disorganized before the train. The batch size is 4 and the epoch is 1,000. Model will store each 50 epochs. Then, we input our preprocessed data above and begin to train the model. The structure of the neural network is as follows [7]:

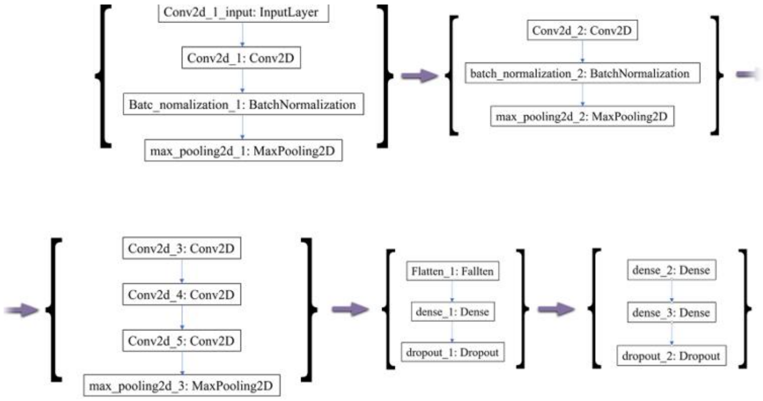

Conv2d_1_input: InputLayer
Conv2d_1: Conv2D
Batc_nomalization_1: BatchNormalization
max_pooling2d_1: MaxPooling2D
Conv2d_2: Conv2D
batch_normalization_2: BatchNormalization
max_pooling2d_2: MaxPooling2D
Conv2d_3: Conv2D
Conv2d_4: Conv2D
Conv2d_5: Conv2D
max_pooling2d_3: MaxPooling2D
Flatten_1: Fallten
dense_1: Dense
dropout_1: Dropout
dense_2: Dense
dense_3: Dense
dropout_2: Dropout

Figure 5.2 Neural Network Structure

After 1000 epochs, the training set accuracy acc reached 0.8969 and the validation set acc is 0.8471. The figure of loss and accuracy rate has been plotted as follows:

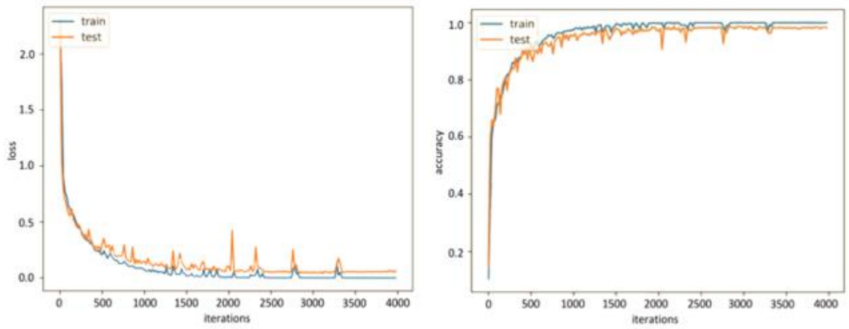


Figure 5.3 Loss and Accuracy Graphs

Diagrams of loss shows the loss function value will decease as the iteration time increases. It decreases dramatically at first and then falls down slowly. Then it will stay stable around 0.03. Moreover, the train set performance will stay beyond the test set. The other diagram shows that model's accuracy level will increase as the model is trained. The accuracy for train set and test train set will go up to 89% and 85%, respectively. Moreover, there are some glitches of loss function which may lead into glitches of accuracy. Thus, enough epochs are necessary for training.

6 Prioritizing Model

The public usually reports mistaken sightings of *V. mandarinia*. Therefore, a criterion of the authenticity is demanding and we will prioritize investigation for the most likely reports.

We focus on the three targets: image recognition, position and text. If an image is classified to be real *V. mandarinia*, the report will be checked manually. Or, we will score the rest targets respectively and if the weighted score is higher than 85, we deem that the report is authentic and it will be checked manually as well.

6.1 Image Recognition Scoring Criteria

In chapter 5, we have built a model to classify mistaken classifications based on images. Now, we put the reported photos (p) to the model and get the prediction about right or wrong classification:

If right, score (p) =100. If wrong, score (p) = 0.

Because photos are strong evidences, we trust the results of analysing images.

6.2 Position Scoring Criteria

We can score the distance between case position and the predicted position of colony from the Spread Prediction Model above. We set d as the distance (unit: *m*).

There is a certain distance between the most likely area to witness a *V. mandarinia* and its colony center. Then, the probability will decrease gradually as distance increases. So do the score(d). Afterwards, we employ a normal distribution to simulate the relationship:

$$score(d) = 100 \times e^{-\frac{(d-\mu)^2}{2\sigma^2}} \tag{6.1}$$

where μ is the mean value denoting the distance between the colony center and the most possible active area. And σ is the standard deviation. The score of the starting point (score(0)) is about 80% of the full mark. By parameter identification, the standard deviation is 9000.

Moreover, we utilize Monte Carlo Simulation in the prediction model and obtain the most regular area for the Asian giant hornets to active is about 6000 m from the central point of their colony. Hence, we deduce the relationship as follows:

$$score(d) = 100 \times e^{-\frac{(d-6000)^2}{1.62\times10^8}} \tag{6.2}$$

Then, we discuss the distance range. Based on the discussion above, we obtain the following equations:

$$\frac{d}{100\sqrt{2\pi}} \sim N(\mu, \sigma^2) \tag{6.3}$$

If that:

$$\frac{\frac{d}{100\sqrt{2\pi}} - \mu}{\sigma} < u_\alpha \tag{6.4}$$

where $u\alpha$ is the unilateral quantile ($\alpha = 0.05$ based on the confidence coefficient):

$$d < 100\sqrt{2\pi}(\mu + \sigma u_\alpha) \tag{6.5}$$

And the graph is as below:

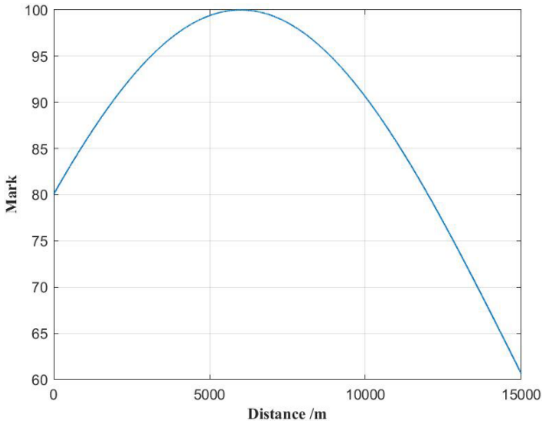


Figure 6.1 Mark and Distance

The starting score is 80 and when the distance reaches 6000 m, it gets a full score (100).

6.3 Text Scoring Criteria

Apart from the elements above, some keywords in the reports also require attention. If we construct a keywords library, and retrieve keywords based on it whenever there is a new report. With this method, we can test the reliability of the descriptions by witnesses, which lightens the workload.

We have extracted the word cloud from the attached information of reports and the following pictures display the result:

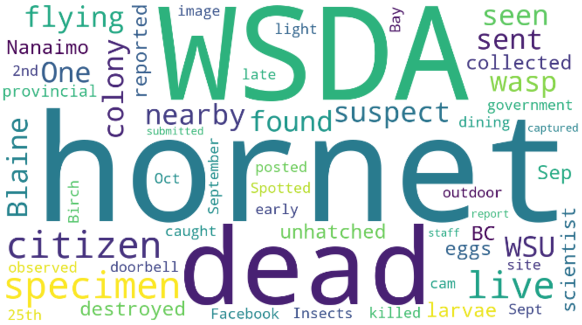


Figure 6.2 Words in Positive Reports

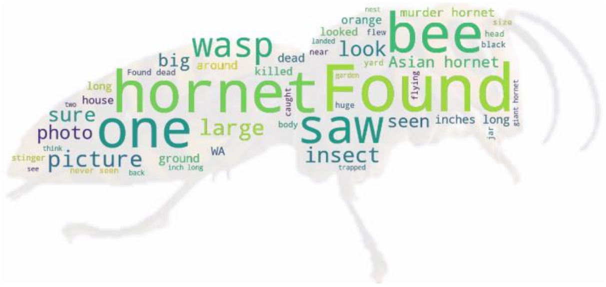


Figure 6.3 Words in Negative Reports

Figure 6.2 and 6.3 represent the words extracted form given positive and negative reports respectively.

Then, we aggregate the positive words and negative words in the following table:

Table 6.1 Positive and Negative Keywords

Positive Words	WS, specimen, colony, WSU, scientist, Blaine, captured, suspect, reported, unhatched
Negative Words	bee, murder

If there is a negative word, which lowers the reliability of a report, the score will decrease by 20. By contrast, a positive word will raise the reliability and the score will increase by 20. The range of total score is from 0~100 in this target and the result out of the range will be replaced by the nearest boundary value. In this way, we can get the score based on the description (score(w)).

6.4 Prior Report Selection Model

First of all, we score the photos in a new report. And once it is classified as an image of *V. mandarinia*, we will check it artificially. Otherwise, we will continue to score the other two targets: distance and word cloud (score(d) and score(w)):

$$score = p \times score(d) + q \times score(w) \tag{6.4}$$

where p and q are the weights and:

$$p + q = 100\% \tag{6.5}$$

To determine the weights, we focus on given positive reports. The score(d) and score(w) are combined with different p and q values and estimate the most possible values for them to make $score$ in equation 6.4 higher than 85. The result is as below:

$$\begin{cases} p = 42\% \\ q = 58\% \end{cases} \tag{6.6}$$

The flow chart of this process is as follows:

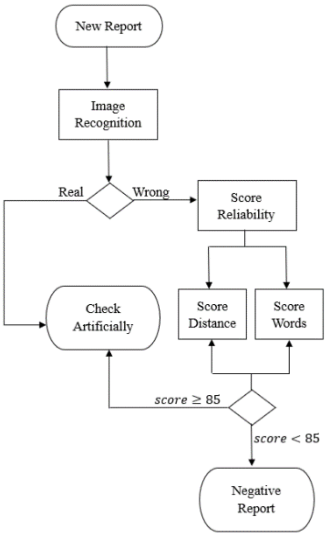


Figure 6.4 Flow Chart of Positive Reports Selection

From the flow chart above, we could assess reports from publics and choose the most suspicious cases to investigate.

7 Model Update Strategy

New confirmed case provides sampling information so that we have to update the model after a new positive case is reported. Once we have real subsequent information, we could update our model by Bayesian Statistics. In this segment, we focus on the method to update our models and how frequent we update them.

7.1 Verification of Authenticity

In the discussion above, we chose the most reliable positive reports to investigate. In line with the discussion, only a report is confirmed to be positive can it be taken into account for our further improvement of the models.

7.2 Update Based on Bayesian Statistics

In the research of the first sub-problem before, we predicted the central point and the possible distribution of the colony and got the probabilities in different regions. Now, we set the area of the distribution as **S** and the central point of the colony as **B**. Then, we begin to classify the received reports based on the geographical locations. Therefore, there are two situations: all the locations in the positive reports this month are within the area **S**; or, there is a site out of the area **S**. And we are going to discuss these two situations respectively.

7.2.1 Situation One

As for the former situation, firstly, we analyze the case that there is only **one** new positive report in the month. According to this situation, the positive location locates in the predicted area **S**, which proves that our prediction is reliable to some extent. Moreover, the new report provides plenty of real information about the colony, which helps us to modify our models. The site needed to be updated is **B** which is shown as follows:

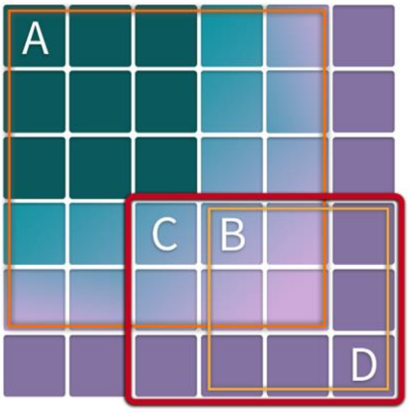


Figure 7.1 Update Regions

where **A** is the central point of colony predicted last month, **B** is the central point predicted early this month, **C** is the central point located in anywhere within the potential active area of last month, **D** is the location in the new report and it locates in the distribution area which regards both **B** and **C** as its centre. Then, conditional probability is used to estimate the most potential location between **B** and **C**.

Table 7.1 Events and Descriptions One

Events	**Descriptions**
E	The colony appears in area **D**
F_1	The colony appears in area **B**
F_2	The colony appears in area **C**
F_i (i=1,2,3…)	The colony appears in area **C**$_i$ (i=1,2,3…) and **C**$_i$ is within the area **S**

Then, F_i are mutually exclusive events and we set P(F_i) as its prior probability which reflects the probabilities of the emerging of all the potential reasons.

On the basis of the Total Probability Theorem [8], we deduce the equation as below:

$$P(E) = \sum_i P(F_i E) = \sum_i P(E|F_i) \cdot P(F_i) \tag{7.1}$$

Then, according to Bayes Formula, we can get the equation as follows:

$$P(F_i|E)=\frac{P(E|F_i)\cdot P(F_i)}{\sum_i P(E|F_i)\cdot P(F_i)} \tag{7.2}$$

where $P(F_i|E)$ is the posterior probability, which means the probability that the colony center locates in area F_i given that event *E* has occurred. The greater value the equation 7.2 gets, the more possible that the colony will arrive in area F_i.

Subsequently, take event F_1 and event F_2 for example, according to equation 7.3, if $P(F_1|E)$ and $P(F_2|E)$ are compared, we can only analyse the following formulas:

$$\begin{cases} P(E|F_1)\cdot P(F_1) \\ P(E|F_2)\cdot P(F_2) \end{cases} \tag{7.3}$$

And if equation 7.3 satisfies the following condition:

$$P(F_2|E) > P(F_1|E) \tag{7.4}$$

then, it is illustrated that the colony are most possible to arrive in area **C** given that it is witnessed in area **D**. And according to the result, we will modify our models.

Finally, we calculate the potential regions for the colony to move to and if area D is not included in the distribution area centred on the area F_i, then:

$$P(E|F_i)=0 \tag{7.5}$$

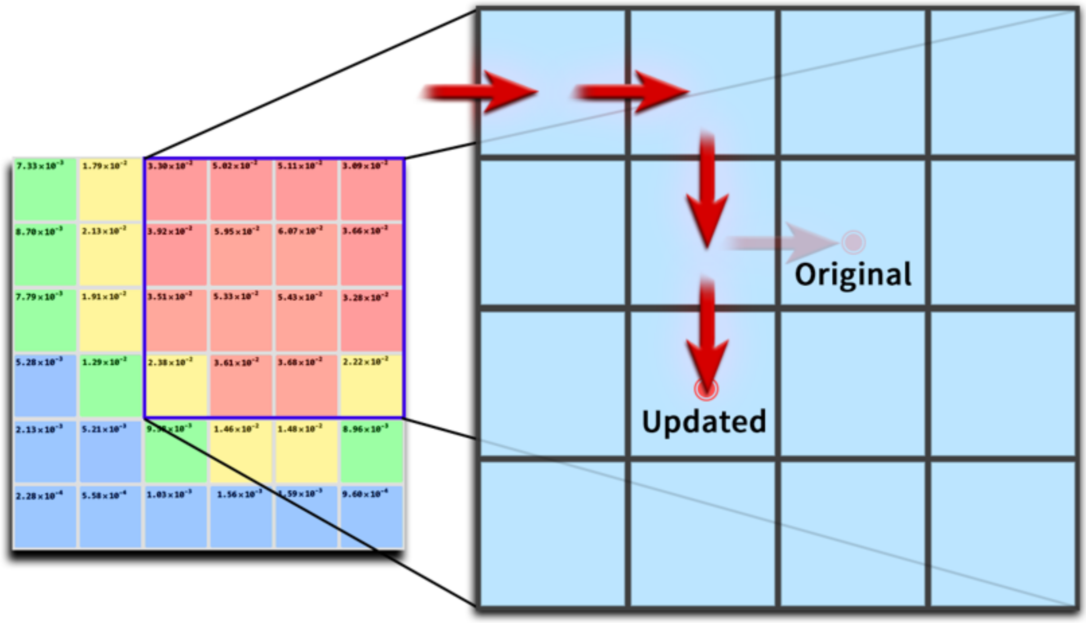
On the basis of the equations above, we can do the further calculation and comparison.

To demonstrate the steps above, here is an example. We have developed a mini-program to update new positive case verified by WSDA. Codes have been attached in Appendix C. For instance, if we receive a positive case reported on block (7,7), we will calculate the conditional probability for 4×4 block circled above:

$$p_{condi}=p_{test}\times p_{new} \tag{7.6}$$

where p_{condi} is the probability of the block, p_{test} is the probability of the maximum block and p_{new} is the probability of the block with the same relative replacement from the maximum block.

As a result, we give new conditional probabilities in the 4×4 block in figure below:



7.33×10^{-3}	1.79×10^{-2}	3.30×10^{-2}	5.02×10^{-2}	5.11×10^{-2}	3.09×10^{-2}
8.70×10^{-3}	2.13×10^{-2}	3.92×10^{-2}	5.95×10^{-2}	6.07×10^{-2}	3.66×10^{-2}
7.79×10^{-3}	1.91×10^{-2}	3.51×10^{-2}	5.33×10^{-2}	5.43×10^{-2}	3.28×10^{-2}
5.28×10^{-3}	1.29×10^{-2}	2.38×10^{-2}	3.61×10^{-2}	3.68×10^{-2}	2.22×10^{-2}
2.13×10^{-3}	5.21×10^{-3}	[illegible]	1.46×10^{-2}	1.48×10^{-2}	8.96×10^{-3}
2.28×10^{-4}	5.58×10^{-4}	1.03×10^{-3}	1.56×10^{-3}	1.59×10^{-3}	9.60×10^{-4}

Figure 7.2 4×4 Block with New Conditional Probabilities

It illustrates that the most possible point of migration moves from original block (2,3) to updated block (3,2).

Furthermore, we analyse the case that there are some more positive reports in a short period of time and the reported locations are within our predicted area **S**. If the positive reports are reported at a long-time interval, we will update our models in chronological order based on the discussion above. However, it requires further research when the dates of positive reports are close. And there are some new events included:

Table 7.2 Events and Descriptions Two

Events	**Descriptions**
E	Several Asian giant hornets appear in area **D**
E_i (i=1,2,3…)	Hornet number i appears in area **D**

Then, E_i are mutually exclusive events:

$$P(E) = P(E_1 \cap E_2 \cap \dots) = \prod_i P(E_i) \tag{7.7}$$

Afterwards, we will further deduce:

$$P(E|F_i) = P(E_1 E_2 \dots |F_i) = \prod_j P(E_j|F_i) \tag{7.8}$$

Consequently, equation 7.3 can be updated as below:

$$P(F_i|E) = \frac{(\prod_j P(E_j|F_i)) \cdot P(F_i)}{\sum_i (\prod_j P(E_j|F_i)) \cdot P(F_i)} \tag{7.9}$$

Finally, based on the equations above, we will obtain a more reliable predicted location of the colony.

7.2.2 Situation Two

As for the latter situation, there is at least one location in positive reports is out of our predicted position. It is a small probability event and it does not implicate that our models failed. It is because data are under sampled [9] and our model could update with it.

When there is only a positive location reported, then, we deem that our models have a deviation that we have to update them. Based on the useful information: the appearing location of *V. mandarinia*, we deem that the real central point of the colony is nearby. Thus, we suppose new reported case's position to be new origin. In addition, if several cases are confirmed, their geographic mass of centre will become new origin.

7.3 Update Frequency

Our model requires update to assure its accuracy, however, it is not proper to update the model frequently. Because on the one hand, it may enlarge the deviation. On the other hand, an enormous amount of costs is demanded for an update, while the budget and resources are limited.

If there is no new positive report, we will update the model monthly based on the principle in Chapter 8.

If there is a new case reported, we will wait for 15 days and if there is another new positive case, we will deal with them together. Or, we will update the model based on the only case. The 15-day is for the reason that *V. mandarinia* is destructive and the short-period reports may implicate the wide spread of the pests.

8 Disappearing Judgement Model

V. mandarinia might vanish because of the perniciousness. And in this segment, we will discuss how to evaluate the extent of its extinction.

8.1 Conclusive Evidence

New positive reports will reflect the existence of the pest. If continuous positive reports are reported monthly, the rage of this pest is transparent. The fewer new positive reports are, the fewer the hornets are. When the number of positive reports is below a certain level, it is deemed that they are eradicated.

We suppose 2 years as a unit, according to the number of positive reports within this period, we analyse whether the pest is extinct. Moreover, the pest doesn't active in winter, so the time unit ignores the three months.

8.2 *Vespa Mandarinia's* Disappearing Evidence

We make assumptions on the basis of the number of new positive reports (n) in the 2-year unit:

Table 8.1 Hypothesis of Extinction

Hypothesis	Content
H	*V. mandarinia* hasn't been eradicated.
H_0	*V. mandarinia* has been eradicated.

Supposed H is true, we assume that n obeys Poisson distribution [10]:

$$P(n=k)=\frac{\lambda^k}{k!}\cdot e^{-\lambda},(k=0,1,2\ldots) \tag{8.1}$$

where k means the possible number of the new positive reports, λ is the Poisson distribution parameter. Then we determine it as the mean value of the positive reports during the period, which is 10 based on the given data. So, the following equation can be deduced:

$$P(n\le N)=\left(1+\lambda+\cdots+\frac{\lambda^N}{N!}\right)\cdot e^{-\lambda} \tag{8.2}$$

And if $N=2$:

$$P(n\le N)=0.05 \tag{8.3}$$

which means that it is a small probability event and the the hypothesis is rejected with confidence coefficient 99.5%. So, we deem that if the number of positive reports is less than 2 in a 2-year unit, *V. mandarinia* has been eradicated in Washington State.

9 Sensitivity Analysis

Here we give a brief sensitive analysis to Spread Prediction Model mentioned in Character 4. We have decided possibility of moving toward one direction $(\alpha,\beta\ and\ \gamma)$ with respect to reported cases in real map. However, this direction is not very accurate because reported case sample is under sampled and hard to represent colony migration.

Here we add a 2.5% disturbance to the north-south moving probability and new probabilities change into:

$$\begin{cases}\alpha_1=0.4(South)\\ \beta_1=0.2(North)\\ \gamma_1=0.4(Stable)\end{cases}\Rightarrow\begin{cases}\alpha_1=0.41\\ \beta_1=0.2\\ \gamma_1=0.39\end{cases} \tag{9.1}$$

We can draw graphs as follows:

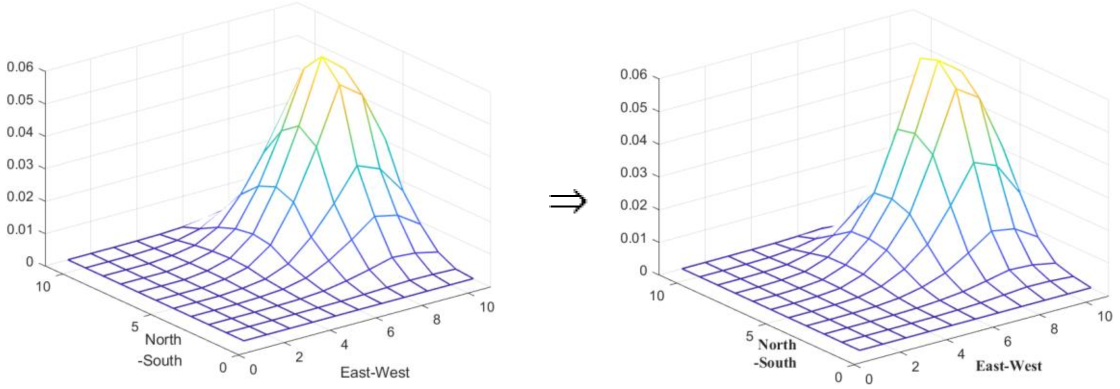


Figure 9.1 Comparison between Original and Deviated Probabilities

The diagram left is with original probabilities while the right one is with a disturbance. There is no significant biased change because of disturbance. However, the probabilities of model fluctuate because the frequencies fluctuate around the probabilities. Thus, our model is not sensitive with moving probability.

10 Model Evaluation and Further Discussion

10.1 Evaluation

The main work of this segment is to estimate both the advantages and weaknesses of our models.

First of all, the primary advantages are as follows:

1. The map is divided to make the spread prediction model fit the real data and have good stability.
2. The models include the effect of images, descriptions and locations as a whole.
3. Prior distribution and Bayes' theory are employed to update the prediction model.

Simultaneously, the weaknesses are listed as below:

1. The time interval is too long without enough time information.
2. The judgement of descriptions and locations is not objective enough.
3. The unprocessed locations are underutilized.

10.2 Further Discussion

We can shorten the time interval to make it more precise. Or, we can divide the colony into several small groups and analyze their migration.

Apart from that, the spread prediction model can be used to forecast the migration of other invasive species. The report classification model can be used in other fields.

11 Memorandum

A Research on Vespa Mandarinia in Washington State

After the colony of Vespa mandarinia in Canada was destroyed in September 2019, a panic spread quickly throughout the regions around. Afterwards, some strong evidences emerged gradually confirming that this pest exists in Washington State. In light of its destructive power, the government should make measures to destroy it. And we will offer our suggestions based on our models.

First and foremost, *Vespa mandarinia*, also known as Asian giant hornet, is a kind of hornet. Its body is black and its average length is 16 mm. It has a stinger and is extremely aggressive. And its attack may cause death of people and other bees, which will damage the environment to a large extent. Moreover, *Vespa mandarinia* is possible to migrate all year round except for the period from December to February. With the basic knowledge, we build a series of models.

Our first model is used to predict the area where the colony of *Vespa mandarinia* will move to in Washington State. We analyzed the positive reports from given reports which demonstrated the tendency of moving to the central part of the state. And then, the spread prediction model was built and predicted the most likely place for the migration of the colony at different times. By prediction, the colony will arrive in Lynden city in the end of 2021.

Next, we built the report classification model to distinguish right reports from mistaken ones. It is necessary considering the limitation of the resources and budget. Due to the confusion between *V. mandarinia* and other bees, we employed Convolutional Neural Network (CNN). Given images constituted the training set to train the network. The trained CNN is to judge whether a reported image is about *V. mandarinia*.

Subsequently, the Prioritizing Model of reports was built to prioritize investigation of the most potential reports. This model focuses on the three targets extracted from a report: image, distance and words in description. Firstly, the image is tested its authenticity by CNN. If the result confirms its validity, the report will be check artificially later. Otherwise, the model will score the other two targets respectively. Only if the total score is higher than the threshold value, can the report be further checked by relative personnel. Or, the report is believed with a low level of confidence.

Besides, our models should be flexible to new reports, which requires us to update our models with some method. Generally speaking, we will update our models monthly if there are no new reports. Once a report appears, we will modify our prediction model based on the new date and location. In view of the costs and consumption, we will deal with all the new reports in 15 days from the date of the first emerged report. In this way, the update process will be efficient and economical and our models will be more accurate.

Eventually, the evidence proving the extinction of *V. mandarinia* is demanding, since it is the probable consequence of all the control measures. By mathematical reasoning, if new reports of positive sightings are fewer than two cases in 2 years, it illustrates that *V. mandarinia* has been eradicated from Washington State with a probability over 99%.

At this point, we have introduced all of our work and results. And because our work was on the basis of the reports reported by the public, it will be helpful to appeal to the public to provide relative reports as detailed and accurate as possible. We sincerely hope that our summary can be effective.

References

[1] Matsuura, M. and S. F. Sakagami. 1973. A bionomic sketch of the giant hornet, *Vespa mandarinia*, a serious pest for Japanese apiculture. Jour. Faa. Sci. Hokkaido, 19(1): 125–162

[2] Attachment file 2021MCM_ProblemC_Vespamandarinia.pdf

[3] Hubbard, Douglas; Samuelson, Douglas A. (October 2009). "Modeling Without Measurements". OR/MS Today: 28–33.

[4] Matsuura, M. 1988. Ecological study on vespine wasps (Hymenoptera: Vespidae) attacking honeybee colonies. I. Seasonal changes in the frequency of visits to apiaries by vespine wasps and damage inflicted, especially in the absence of artificial protection. Applied Entomology and Zoology, 23(4): 428–440.

[5] Irhum Shafkat, Intuitively Understanding Convolutions for Deep Learning, https://towardsdatascience.com/intuitively-understanding-convolutions-for-deep-learning-1f6f42faee1

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12). Curran Associates Inc., Red Hook, NY, USA, 1097–1105.

[7] John Ellson et al. Graphviz and Dynagraph – Static and Dynamic Graph Drawing Tools, online published on website

[8] M. H. Degroot & M. J. Schervish, Probability and statistics Fourth Edition, Pearson ISBN 978-0-321-50046-5 Page 349

[9] Walt Kester (2003). Mixed-signal and DSP design techniques. Newnes. p. 20. ISBN 978-0-7506-7611-3.

[10] Haight, Frank A. (1967), Handbook of the Poisson Distribution, New York, NY, USA: John Wiley & Sons, ISBN 978-0-471-33932-8

Appendix A

```
ns=[]; ew=[]; for a=1:1000
    NS=6; EW=6;
    for x=1:5
        south=rand();east=rand();
        if south>0.60
            NS=NS+1;
        elseif south<0.2
            NS=NS-1;
        end
        if east>0.3
            EW=EW+1;
        elseif east<0.1
            EW=EW-1;
        end
      end
    ns(a)=NS; ew(a)=EW;
end
cou_ns=[]; cou_ew=[];
for b=1:11
    cou_ns(b)=int16(sum(ew == b,'all'));cou_ew(b)=int16(sum(ns==b,"all"));
end
x=1:10;y=1:10;z=transpose(cou_ew)*cou_ns;z=z/1000000
sum(sum(z))
mesh(z)
```

Appendix B

```
from keras.callbacks import TensorBoard, ModelCheckpoint, ReduceLROnPlateau, EarlyStopping
from keras.utils import np_utils
from keras.optimizers import Adam
from model.AlexNet import AlexNet
import numpy as np
import utils
import cv2
from keras import backend as K
K.image_data_format() == 'channels_last'
def generate_arrays_from_file(lines,batch_size):
    n = len(lines)
    i = 0
    while 1:
        X_train = []
        Y_train = []
        for b in range(batch_size):
            if i==0:
                np.random.shuffle(lines)
            name = lines[i].split(';')[0]
            img = cv2.imread(r"C:/Users/Wu Tao/Desktop/BEEcalssic/BEEcalssic/data/image/train" + '/' + name)
            img = cv2.cvtColor(img,cv2.COLOR_BGR2RGB)
            img = img/255
            X_train.append(img)
            Y_train.append(lines[i].split(';')[1])
            i = (i+1) % n
            X_train =
utils.resize_image(X_train,(224,224))

            X_train = X_train.reshape(-
1,224,224,3)

            Y_train =
np_utils.to_categorical(np.array(Y_train
),num_classes= 2)

            yield (X_train, Y_train)

if __name__ == "__main__":

      log_dir = "C:/Users/Wu
Tao/Desktop/BEEcalssic/BEEcalssic/log
s/"

      with open(r"C:/Users/Wu
Tao/Desktop/BEEcalssic/BEEcalssic/dat
a/dataset.txt","r") as f:

            lines = f.readlines()

      np.random.seed(10101)

      np.random.shuffle(lines)

      np.random.seed(None)

      num_val = int(len(lines)*0.1)

      num_train = len(lines) - num_val

      model = AlexNet()

      checkpoint_period1 =
ModelCheckpoint(
    log_dir + 'ep{epoch:03d}-
loss{loss:.3f}-val_loss{val_loss:.3f}.h5',
monitor='acc',

save_weights_only=False,

save_best_only=True,

period=3)

      reduce_lr = ReduceLROnPlateau(

monitor='acc', factor=0.5, patience=3,
verbose=1)

      early_stopping = EarlyStopping(
monitor='val_loss', min_delta=0,
patience=10, verbose=1)

      model.compile(loss =
'categorical_crossentropy',optimizer =
Adam(lr=1e-3), metrics = ['accuracy'])

      batch_size = 4

      print('Train on {} samples, val on {}
samples, with batch size
{}.'.format(num_train, num_val,
batch_size))

model.fit_generator(generate_arrays_fro
m_file(lines[:num_train], batch_size),

                  steps_per_epoch=max(1,
num_train//batch_size),

validation_data=generate_arrays_from_f
ile(lines[num_train:], batch_size),

                  validation_steps=max(1,
num_val//batch_size), epochs=100,

                  initial_epoch=0,

callbacks=[checkpoint_period1,
reduce_lr])

model.save_weights(log_dir+'last1.h5')
```

Appendix C

```
a=[]; m=input("relative position(NS):")

n=input("relative position(EW):")

b=[];for x=2:5

      for y=3:6

            x1=m-x+1;   y1=n-y+1;

            b(x-1,y-2)=a(x1,y1)*a(x,y);

      end

end

b=b/0.001306099600000

mesh(b)
```