| Problem Chosen | 2021 MCM/ICM Summary Sheet | Team Control Number |
|---|---|---|
| C | | 2122219 |

# Hornet is coming


## Summary

Lately, a colony of the Asian giant hornet (AGH) was discovered on Vancouver Island. The nest was quickly destroyed, but since that time, several confirmed sightings of the pest have occurred in Washington State. It was alarming that the AGH poses a serious threat to apiculture, and this species is considered an actionable quarantine pest. Thus, our team was saked to interpret the data provided by the public reports and help the state to prioritize these reports for additional investigation.

For the aspect 1, we affirmed that the invasion is just at the biginning, because of the very few number of positive reports compared to negative reports and the positive reports have a very concentrated disribution. Based on this deduction, we chose cellular automata (CA) method to predict the spread of AGH over time, in which we considered the positive reports' information, AGH's life habits, and environmental factors. We take one year as time step, 10 km as the diameter of a cell. Dispersal heat maps of AGH in 2019, 2020, 2025 and 2030 predicted using CA are given in the later section.

For the aspect 2 and 3, we firstly trained the BP neural network method to do classification based on the predicted heat maps and information of the given reports. After that, we used the trained network to evaluate the mistaken likelihoods of unverified reports. Then, we addressed aspect 3 based on aspect 2. Due to the harm caused by the AGH at different distances from the the report is far apart. So, we took the likelihood, the average value of the predicted heat degree within 1km of the report and the predicted heat degree occurring within 8 km but beyond 1 km respectively as the evaluation indexes and established a comprehensive evaluation model using these. Then, We prioritized the unverified reports with this model.

For the aspect 4, our update model is, first of all, we need to compare the new reports in one year with those in the previous year to determine whether the AGH trend is increasing or decreasing. Then, we decided to add whether a superposition or attenuate process to the CA model according to the trend, so that the predicted distribution can be closer to the actual situation.

For the aspect 5, we know that whether AGH can be found and reported is a probability problem. But, what we can accept is that if there is no positive report in a certain year, there is a high probability that AGH is eliminated. So, in this paper, we proposed a method to estimate this probability based on our prediction and update model of the AGH's dispersal.

Finally, we analyze the sensitivity of our model, proving that our model is accurate and robust for different cases.



**Keywords**: Cellular automata, Heat map, BP neural network, Comprehensive assessment

# Contents

# 1 Introduction

In September 2019, the Asian giant hornet (AGH) was discovered on Vancouver Island in British Columbia, Canada, but he nest was quickly destroyed. The AGH threatens the survival of other species of insects, such as European honeybees and other insects that are considered agricultural pests. The life cycle of this hornet is similar to many other wasps. Fertilized queens emerge in the spring and begin a new colony. In the fall, new queens leave the nest and will spend the winter in the soil waiting for the spring. A new queen has a range estimated at 30 km for establishing her nest.

Due to the potential severe impact on local honeybee populations, the State of Washington has created helplines and a website for people to report sightings of these hornets. Some of these sightings were Vespa mandarinias, but some were other kinds of insects. Based on these reports from the public, the state must decide how to prioritize its limited resources to follow-up with additional investigation. We are asked to explore and address the following aspects:

- Address and discuss whether or not the spread of this pest over time can be predicted, and with what level of precision.
- The error rate of witness reports is relatively high. So we should create, analyze, and discuss a model that predicts the likelihood of a mistaken classification by the files provided.
- Check the correctness of the model. Use the model to discuss how our classification analyses leads to prioritizing investigation of the reports most likely to be positive sightings.
- Address how our could update your model given additional new reports over time, and how often the updates should occur.
- Use the model to constitute evidence that the pest has been eradicated in Washington State.

# 2 Analysis of the Problem

Firstly, to predict the spread of AGH, we should know clearly about what information can we get from the datas given along with the problem. We counted the number of reports in different lab status, we can affirm that the invasion is just at the biginning, because of the very few number of positive reports compared to negative reports and the positive reports have a very concentrated disribution. Based on this deduction, we choose cellular automata (CA) method to predict the spread of AGH over time, in which we consider the positive reports' information, AGH's life habits, and environmental factors.

Secondly, in our view, evaluating the likelihood of a mistaken classification and investigation priority level of reports are of the same essence and the last is based on the former. In many evaluation models, such as AHP model, the relationship between evaluation objectives and evaluation indexes is often linear, and the weight of evaluation indexes determined by them is often difficult to avoid the problem of being too subjective, which often leads to the evaluation results that cannot accurately reflect the actual situation. At the same time, we have a lot of real reports or disaggregated data to draw on, and this data are inherently informative. In view of this, we choose

BP neural network model to do the evaluation. Then, we can address aspect 3 based on aspect 2. Due to the harm caused by the AGH at different distances from the the report is far apart. So, we will take the likelihood, the average value of the predicted heat degree within 1km of the report and the predicted heat degree occurring within 8km but beyond 1km respectively as the evaluation indexes and establish a comprehensive evaluation model using these. Then, We can prioritize the unverified reports with this model.

Thirdly, considering that positive reports will decrease or even disappear under the influence of human factors. So, we need to compare the new reports in one year with those in the previous year to determine whether the AGH trend is increasing or decreasing. Then, we decide to add whether a superposition or attenuate process to the CA model according to the trend, so that the predicted distribution can be closer to the actual situation.

Fourthly, we know that whether AGH can be found and reported is a probability problem. What we can accept is that if there is no positive report in a certain year, there is a high probability that AGH is eliminated. So, in this paper, we will propose a method to estimate this probability based on our prediction and update model of the AGH's dispersal.

The framework of our models in this paper is shown in Fig. 1.

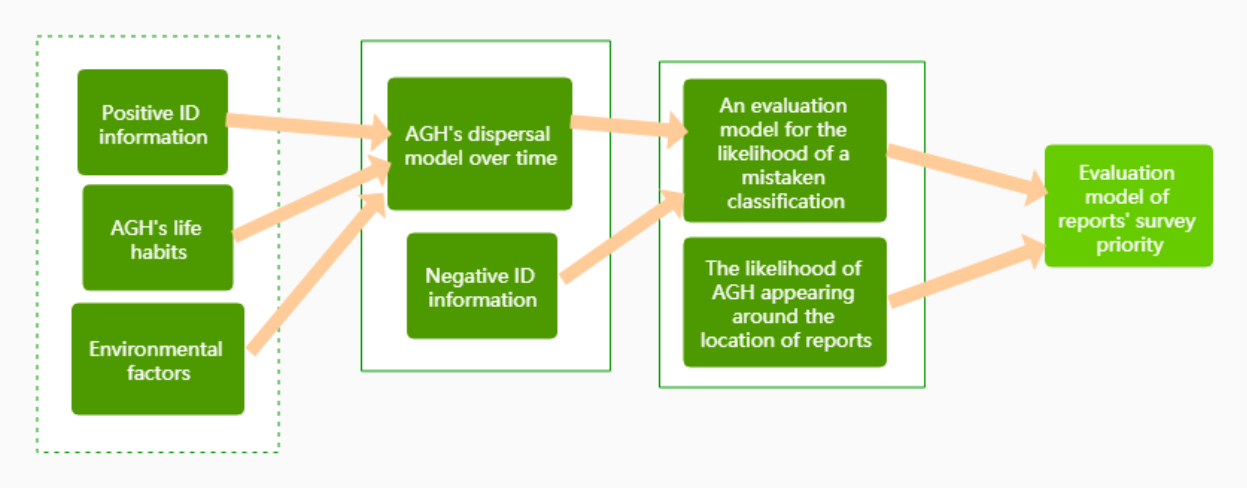



Figure 1: Modeling framework of this paper

# 3 Model Assumptions

- The differences in AGH's seasonal activity can be ignored and the distribution of AGH is the same all year round, which is the distribution in the end of the year.
- The initial AGH invasion sites are within 38km of the sites of Positive ID in 2019.
- The effect of topography on AGH diffusion can be ignored.

# 4 Symbol description

| Symbol | Definition |
|---|---|
| $x_i$ | The $ith$ evaluation index |
| $y$ | The report or a evaluation target |
| $H$ | The heat map |
| $\tau$ | Threshold |
| $sgn(h)$ | $sgn(h) = \begin{cases} 1, & h > 0 \\ 0, & h = 0 \\ -1, & h < 0 \end{cases}$ |
| $h$ | The heat degree in heat map |
| $P_{error}$ | The mistaken possibility of a report |
| $Num_{negative}$ | The number of negative reports |
| $Num_{positive}$ | The number of positive reports |
| $L_{priority}$ | The priority level |
| $P^{Y}_{eradicated}$ | The possibility of AGH being eradicated in $Yth$ year |

# 5 Dispersal Model of AGH

Whether or not the spread of this pest over time can be predicted? For this question, our answer is yes, which is based on our meticulous analysis. Firstly, to predict the spread of AGH, we should know clearly about what information can we get from the datas given along with the problem. We counted the number of reports in different lab status, which is shown in Fig. 2. From this figure, we can affirm that the invasion is just at the biginning, because of the very few number of positive reports compared to negative reports. Besides, when we look at the distribution of the reports in different lab status, we can have more confidence to say so, because the positive IDs are very concentrated.

When we look more carefully, we can also see that the initial sites of invasion is likely to be more than one. As shown in Fig. 3, the nearest distance between Vancouver positive report and all positive reports in Washington in 2019 is more than 80 km. However, we know that the queen's largest diffusion radius is 30km and the worker bees' biggest activity radius from the nest is 8km, therefore, AGHs in Washington are not migrated from vancouver in 2019. That means there was

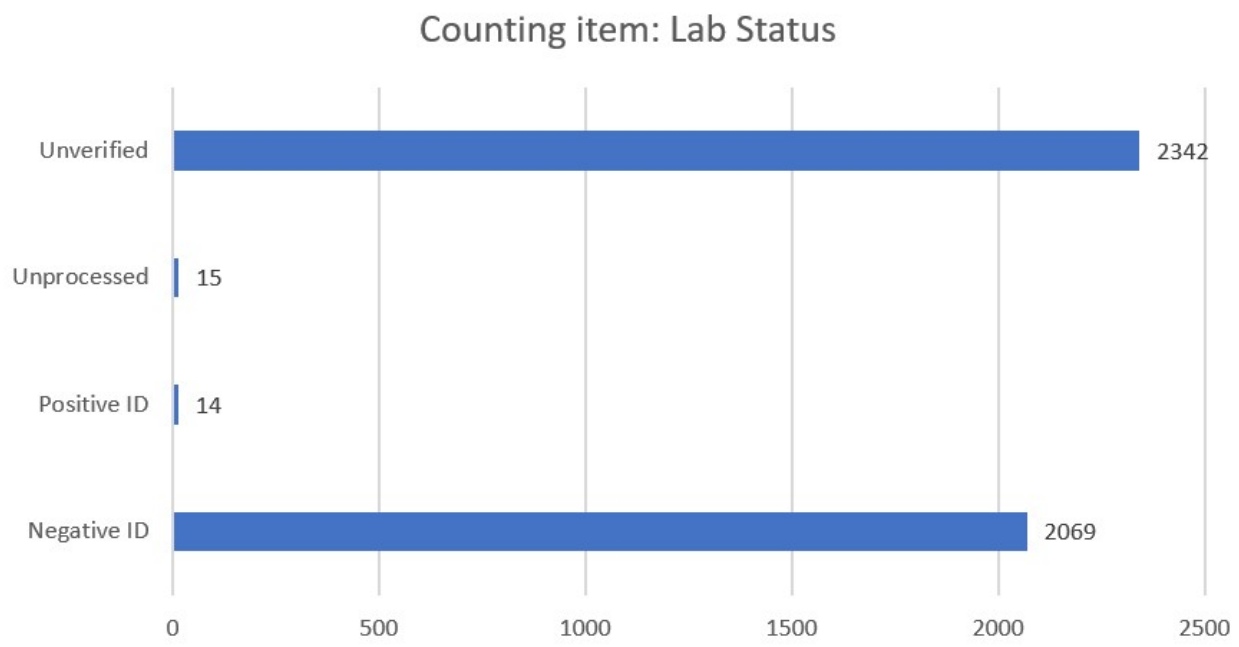



Figure 2: The number of reports in different lab status

probably more than one initial invasion site. In addition, since the maximum distance between all locations of positive reports in 2020 and all locations of positive reports in Washington State in 2019 is 29.9 km, less than 38 km, which means all locations of positive reports in 2020 are within the AGH diffusion range of all positivity reported locations in 2019. Thus, we have reasons to hypothesize that the initial AGH invasion sites are within 38km of the sites of Positive ID in 2019. Then, take enviroment factors into consieration, we build a dispersal model to predict the spread of AGH over time. After a lot of discussion and analysis, we think cellular automata (CA) mehtod is the best choice that is suitable for this case.

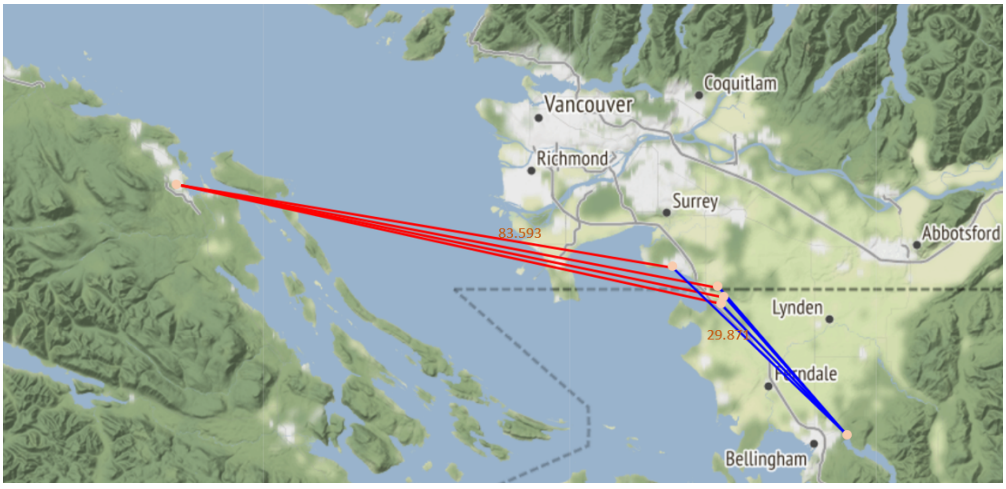



Figure 3: Distances among positive IDs

## 5.1 CA model establishment

CA are a scheme for computing using local rules and local communication. Typically a CA is defined on a grid, with each point on the grid representing a cell with a finite number of states. A transition rule is applied to each cell simultaneously. Typical transition rules depend on the state of the cell and its (4 or 8) nearest neighbors, although other neighborhoods are used. CAs have applications in parallel computing research, physical simulations, and biological simulations. This section will consider how to apply CA to this problem.

To apply this method, initially, we should discrete the time and grid the space. Taking the geographical factors such as latitude and longitude into account, we have improved the CA algorithm. As the AGHs' dispersal depends on the migaration of the new queens, which migarate by year, we take one year as time step. We take 10 km as the diameter of a cell.

The existing points marked as positive are used as the input points for this algorithm. Let's assume that these points are the source points. Information on AGHs' migration shows that the maximum distance a AGHs' colony can travel in a year is about 30 kilometers. At the same time, if a point is farther away from the source point, its probability of being a migration destination will be lower. To reflect this, we assume that the probability of the migration destination being within a range of 0 to 15 km from the source is 0.6, and that the probability of the migration destination being within a range of 15 to 30 km is 0.4. For the four directions of east, west, south and north, we follow the above steps to determine the distance and perform migration simulation. Secondly, factors such as light, temperature, moisture, air, and soil will affect the destination of migration. So we obtained the climate influencing factors corresponding to different latitudes and longitudes as the influencing factors for determining the destination of cell migration [1]. For the use of this factor, we first calculate the average climate factor in the area as 128. Considering that the climate impact factors obtained are discretely distributed, so for each source point, we use the largest climate impact factor within 2 km as the climate impact factors of the source point. When climate impact factor is greater than 128, we regard it as a high probability of successful migration. For the part less than 128, we set the probability of successful migration as climate impact factor divided by 128.

After determining the above influencing factors, we simulate a process of continuous expansion of the source point as the cell increases over time. At this time, we get the accurate latitude and longitude after each migration( make a destination point map based on this). When the expansion is completed, we calculate each element The heat of the cell (ranging from 45 N to 50 N and 120 W to 125 W). The diameter of each cell is 0.1 degree (It's approximately 10 km). We use the number of points whose distance to the center point of each cell is less than 2 km as the heat agree and make a heat map based on this( make a heat map based on this).

The rules of evolution: first, set the relevant initial parameters, and then input the initial positive point, and then move on with the increasing time and under the above influence factors. After each round of migration (one round in a year), information is record. Repeat the process until the target year is reached. Finally, we use this information to draw relevant charts. The workflow is shown in Fig. 4.

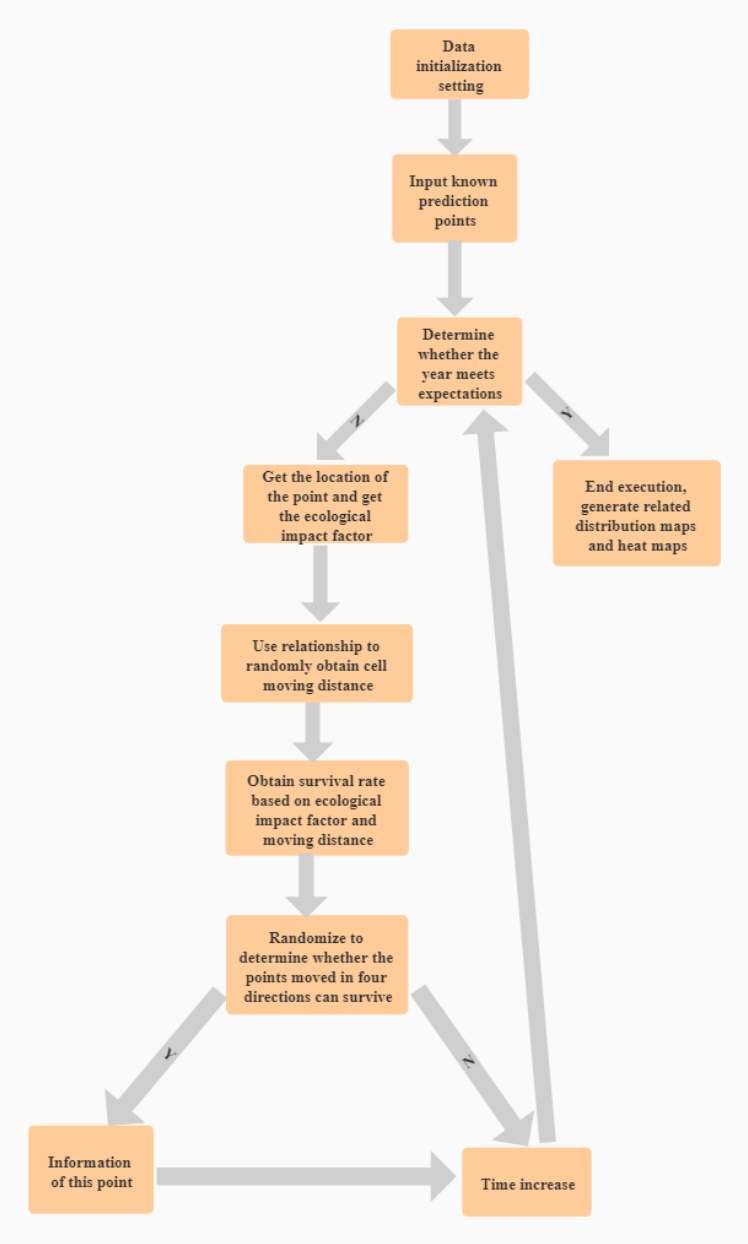
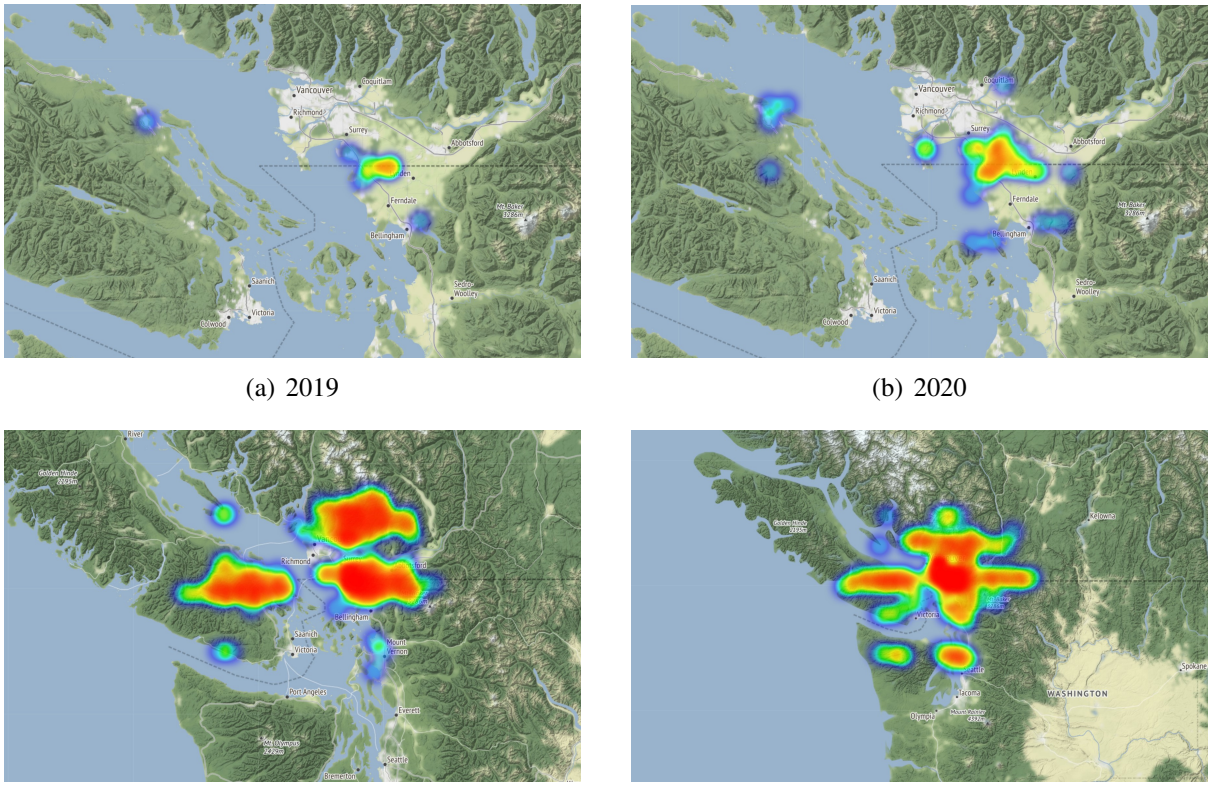



Figure 4: Flow chart of CA method

## 5.2 Prediction results of the spread of AGH over time

We used the dispersal model to draw the predicted dispersal heat maps in 2019, 2020,2025 and 2030, the results are shown in Fig. 17. From the figure, we can clearly see the range and regional differences of AGH diffusion in each year, on the basis of which we can carry out better prevention and control of AGH.




(a) 2019 (b) 2020

(c) 2025 (d) 2030

Figure 5: Dispersal heat maps of AGH predicted using CA mehtods

# 6 Evaluation models of reports

In our view, evaluating the likelihood of a mistaken classification and investigation priority level of reports are of the same essence and the last is based on the former. So, in this section, we will firstly introduce an uniform evaluation method, and then apply it to solving problem 2, finally solve problem 3 based on the solution of problem 2.

## 6.1 BP neural network model establishment

In many evaluation models, such as AHP and fuzzy comprehensive evaluation model, the relationship between evaluation objectives and evaluation indexes is often linear, and the weight of evaluation indexes determined by them is often difficult to avoid the problem of being too subjective, which often leads to the evaluation results that cannot accurately reflect the actual situation. At the same time, we have a lot of real reports or disaggregated data to draw on, and this data are inherently informative. In view of this, we hope to find an evaluation method based on the existing data information, that is to say, the mapping form and coefficients from the evaluation indexes to the evaluation target are determined by the existing data. After sufficient thinking and discussion, we choose BP neural network [2] model to do the evaluation.

In machine learning, backpropagation (backprop, BP) is a widely used algorithm for training feedforward neural networks. Generalizations of backpropagation exists for other artificial neural

networks (ANNs), and for functions generally. These classes of algorithms are all referred to generically as "backpropagation". In fitting a neural network, backpropagation computes the gradient of the loss function with respect to the weights of the network for a single inputoutput example, and does so efficiently, unlike a naive direct computation of the gradient with respect to each weight individually. This efficiency makes it feasible to use gradient methods for training multilayer networks, updating weights to minimize loss; gradient descent, or variants such as stochastic gradient descent, are commonly used. The backpropagation algorithm works by computing the gradient of the loss function with respect to each weight by the chain rule, computing the gradient one layer at a time, iterating backward from the last layer to avoid redundant calculations of intermediate terms in the chain rule [3]. It's working principle is shown in the schematic diagram Fig. 6.

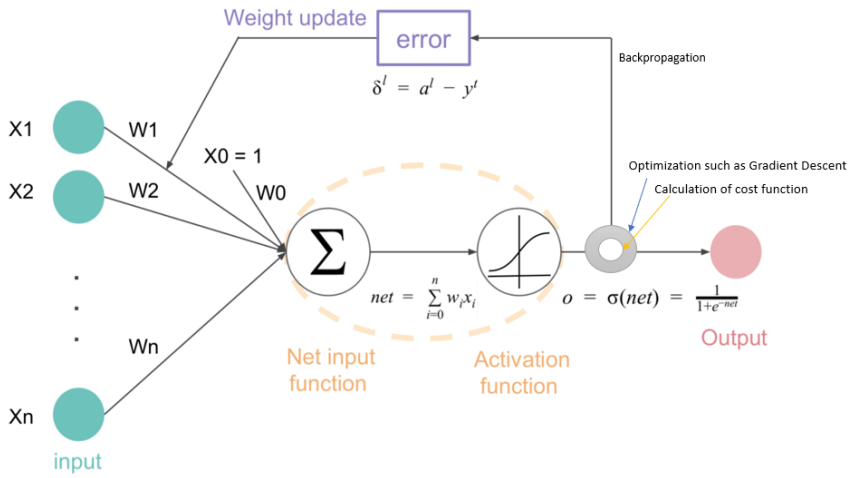



Figure 6: Schematic diagram of backpropagation (Quoted from here)

In the scenario of this paper, the BP neural network is used to establish the evaluation model, which should follow the work flow as shown in Fig. 7. Firstly, the determination of the evaluation indexes $x_i$ depends on the selection of the evaluation target $y$. It should be considered comprehensively and carefully as far as possible, and key indexes should not be omitted, otherwise large errors will occur, which will make the model useless. Then, to get the map

$$y = f(x_1, ..., x_n), \tag{1}$$

we should design an appropriate architecture of the neural network like Fig. 8. To do so, we take each evaluation index as a node of input layer, and input layer is a node that represents the evaluation target. In the hidden layers, We add the Sigmoid activation function layer so that the neural network can train the nonlinear mapping. More details about the architecture will be given in next two sections. As for the training process, we mainly use some features of existed Positive ID reports and negative ID reports combining with the prediction results of AGH' dispersal.

## 6.2 The likelihood of a mistaken classification

What indexes are related to a report misclassified? In our opinion, one of the most important indexes should be the heat degree $h$ at the report location of the heat map $H$ forecasted in this year.

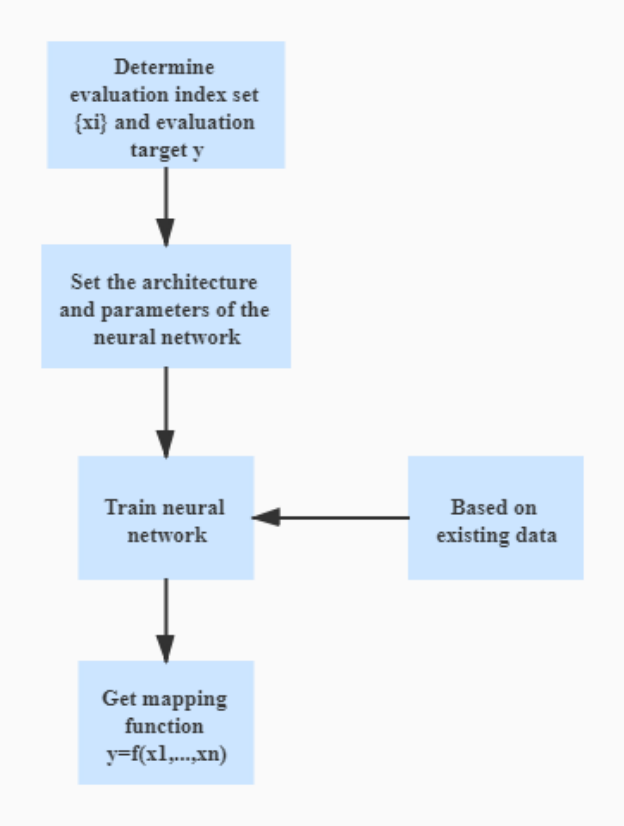



Figure 7: Flow chart of the evaluation model

The larger $h$ is, the more likely AGH is to occur here, that is, the more likely the classification is to be correct here, namely, $h$ is negatively correlated with the evaluation target $P_{error}$ ($P_{error} \in [0, 1]$). Second, we know that there are a lot of AGH-like wasps in Washington State, which can be misleading and that is why so many false reports are being made. On the opposite, the place where the error report appears indicates that different wasps may cause the misjudgment here, and the more the error reports exist, the greater the possibility of different wasps appearing, and the easier it is for people to misclassify. Therefore, the number of error reports $Num_{negative}$ within a certain range from the position of the report to be evaluated is also taken as an important evaluation index. Similarly, correct reporting number $Num_{positive}$, which is negatively related to $P_{error}$, is also should be taken into consideration. The calculation equations of $Num_{negative}$ and $Num_{positive}$ are as follows.

$$Num_{negative} = \sum_{i}^{N} P_{error}^{(i)}, d\left(y^{(i)}, y\right) < \tau_1, \tag{2}$$

$$Num_{positive} = \sum_{i}^{N} \left(1 - P_{error}^{(i)}\right), d\left(y^{(i)}, y\right) < \tau_2, \tag{3}$$

where $N$ represents the total number of negative reports and positive reports, $y^{(i)}$ represents $ith$ report, $y$ represents the report to be evaluated, $P_{error}^{(i)}$ is 0 when the $ith$ report is positive and 1 when negative, $d\left(y^{(i)}, y\right)$ is the distance between report $y^{(i)}$ and report $y$, $\tau_1$, $\tau_2$are two thresholds.

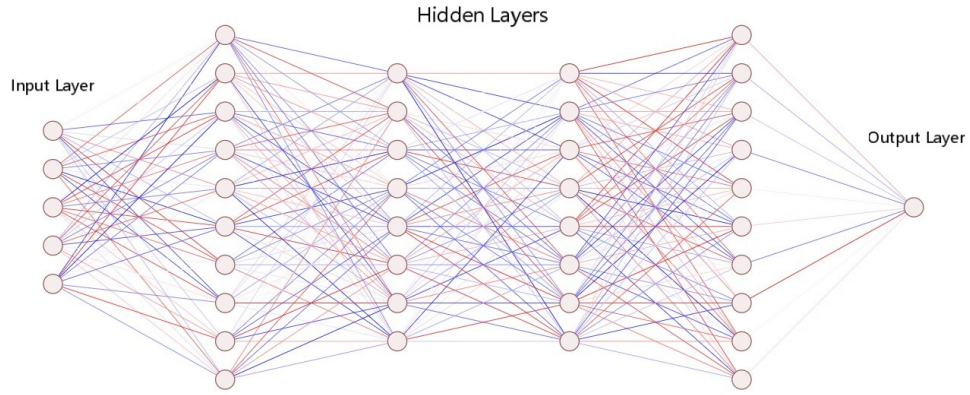



Figure 8: Example architecture of BP neural network

In addition, many factors, such as environment and date, have been taken into account when predicting the spread of AGH, so they are not used as new evaluation indexes to avoid duplication.

Then, we began to train, and our neural network architecture was shown in Fig. 9. First, we obtained the index values corresponding to each positive and negative report. Where, $h$ comes from the predicted AGH heat map of the corresponding year, we set $\tau_1 = \tau_2 = 5km$. Then, we can calculate $Num_{positive}$, $Num_{negative}$. $P_{error}$is 0 when the report is positive, and 1 when negative. We have a total of 2069 report data for training, and we selected 1448(70%) reports as the training set and 621(30%) reports as the test set. For this problem, we are equivalent to using BP neural network to make a classification, so the possibility we want is not the output result, but the probability $P_{error}$ that the output result is error .

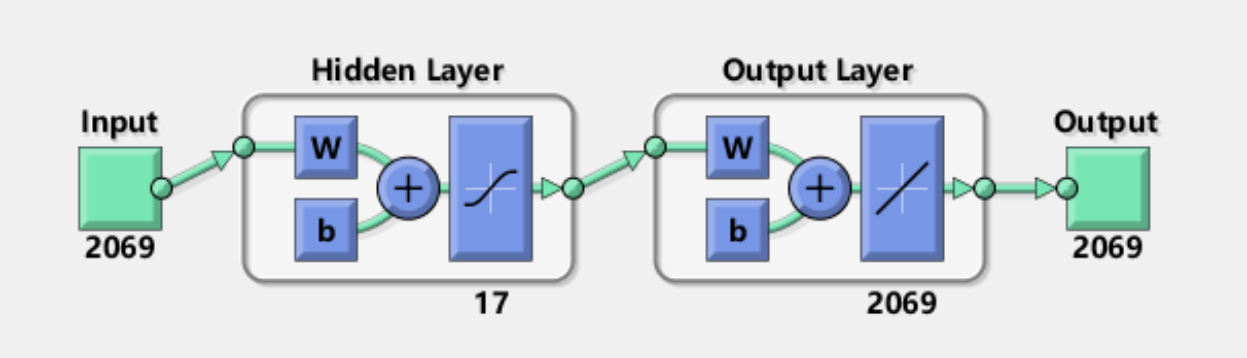
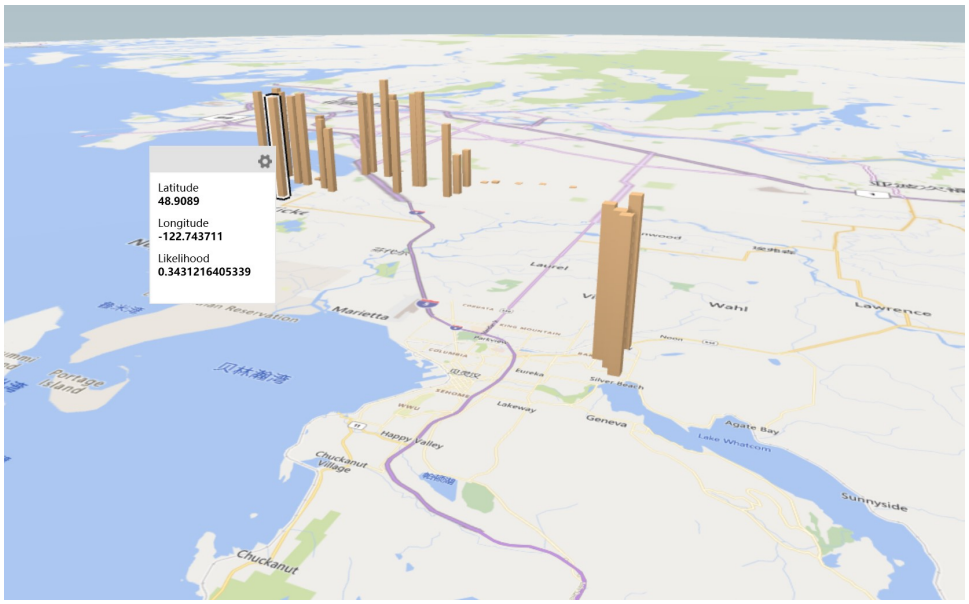


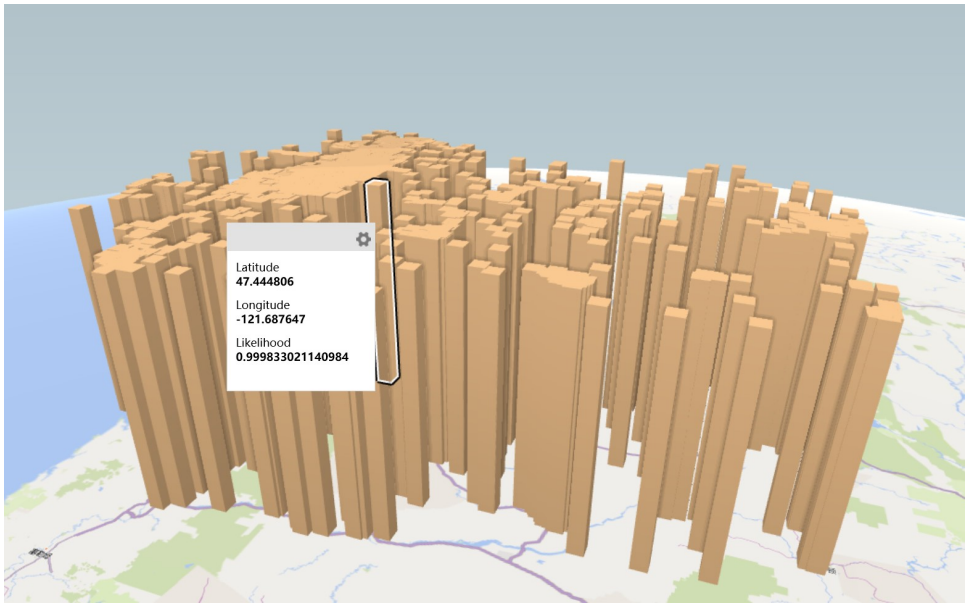

Figure 9: Architecture of our neural network

BP neural network is a multi-layer feedforward network trained by error back propagation algorithm. It can learn and store a large number of input-output patterns and its learning rule is to use the steepest descent method, through back propagation to continuously adjust the weight and threshold of the network to minimize the error square sum of the network. In addition to the above three evaluation indicators, we define the gradient descent function parameter alpha as 0.7, and the hidden layer size is set to [17,8]. We first call the preprocessing function to standardize three

evaluation parameters. Then we perform 10,000 iterations, and update the parameters after each backpropagation to get the model. In the end, we will get a probability in the interval [0,1]. In this model, we assume that the probability is greater than 0.5 as negative, and the probability is less than 0.5 as positive. Then use the test set to verify the obtained model. Comparing the obtained results with the expected results, we obtain a training model with an accuracy of 0.9984. In the last part, we use the unverified reports as the prediction set to determine whether the point is positive or negative. The obtained results are shown in Fig. 10, which are very in line with our expectations.




(a) Positive




(b) Negative

Figure 10: Classification results of the unverified reports

## 6.3 Investigation priority level of reports

To evaluate the investigation priority level of the report $y$, $P_{error}$ attained by BP neural network is certainly one of the most important evaluation indexes that can't be ignored. Besides, we have learned from the attachment that AGHs only fly 0.5 1.25 miles (12 km) on average (and never more than 5 miles (8 km)) from the nest in search of food and there is some evidence that hornets do the worst damage to honey bee colonies that are less than 0.5 miles (1 km) from the nest and that, while nests further away may be molested by one or a few hornets they are not generally slaughtered. Therefore, the harm caused by wasps at different distances from the report $y$ is different, and the closer the distance is, the greater the harm is, and the higher the investigation priority level should be. Therefore, we need to quantify this as an evaluation index. Considering that the presence of AGH within 1km may not only injure people but also damage local beekeepers' income, while the presence of AGH within 8km but beyond 1km will not harm beekeepers' income. Therefore, our index quantification method is to calculate the average value of the predicted heat degree $h^1$ within 1km of the report $y$ and the predicted heat degree $h^2$ occurring within 8km but beyond 1km respectively.

$$h^1 = \frac{\sum h_i}{\max h \cdot \sum sgn\left(h_i\right)}, \quad d\left(y_i, y\right) \leqslant 1km, \tag{4}$$

$$h^2 = \frac{\sum h_i}{\max h \cdot \sum sgn\left(h_i\right)}, \quad 1km < d\left(y_i, y\right) \leqslant 8km, \tag{5}$$

Where $h_i$ represents the heat degree of the $y_i$ report's position on the heat map forecasted in the corresponding year, $\max h$ is the maximum heat degree of the heat map forecasted in the corresponding year, $sgn(h_i)$ meets Eq. 6, the two expressions are divided by $\max h$ respectively is to eliminate the dimensional effect.

$$sgn\left(h\right) = \begin{cases} 1, & h > 0 \\ 0, & h = 0 \\ -1, & h < 0 \end{cases} \tag{6}$$

With these three indicators, the Ministry of Agriculture will assign appropriate weights to these three indicators according to the degree of importance attached to them, and then, they can get the priority level $L_{priority}$ they want using this comprehensive assessment model.

$$L_{priority} = w_1(1 - P_{error}) + w_2 h^1 + w_3 h^2 \tag{7}$$

where $w_1, w_2, w_3$ are the weights and should satisfy $w_1 + w_2 + w_3 = 1$, $L_{priority} \in [0, 1]$, the bigger $L_{priority}$ is, the more the report should be pay attention to.

To test the model, we set $w_1 = 0.4, w_2 = 0.3, w_3 = 0.3$, and caculated all priority levels of all unverified reports and the reslut is shown in Fig. 7. From this figure, we can see clearly about the whole situation of priority levels of different reports. It should be noted that some of the unverified reports overlap in location, so we have accumulated the probabilities at some locations, resulting in the fact that some of the columns in the figure are particularly high.

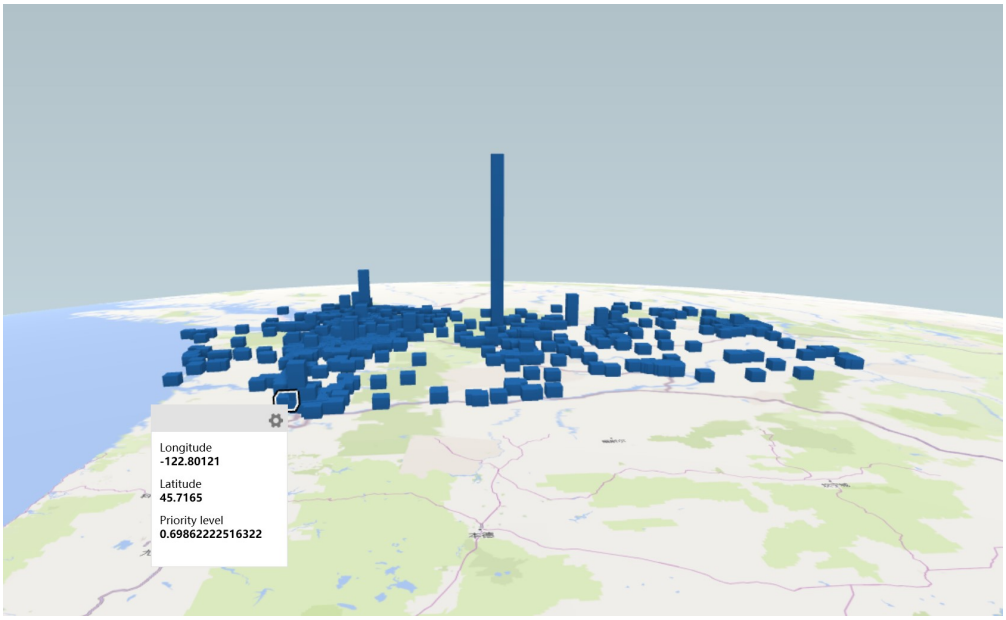



Figure 11: Priority levels of unverified reports

# 7 Model update and evidence of AGH being eradicated

As for how to update the model with new reports, we have actually done this already. When using CA to predict the distribution of AGH, we used the positive reports in 2019 as the seed points to simulate for one year to get the heat map $H_{2019}$, and there occurs many new positive reports in 2020, indicating that the corresponding positions where these new reports appear have a high probability of AGH existing. Therefore, we used these points as seed points again, and used CA to simulate for one year to get $H'_{2020}$. This is superimposed with $H''_{2020}$ obtained by two years of seed point simulation in 2019 to obtain the final heat map of AGH's distribution in 2020 $H_{2020}$. Thus, it can also update when additional new reports appear like this. However, considering that positive reports will decrease or even disappear under the influence of human factors, it is not reasonable to treat them all in this way. So, we should make a classification on this, and the specific process is shown in the Fig. 12. The updated heat map can be obtained by the formula Eq. 8. Since our distribution prediction takes 1 year as a time step, we can update it annually.

$$H_Y = \begin{cases} H'_Y + H''_Y, n_1 > n_2 \\ H'_Y + H''_Y \cdot e^{n_1 - n_2}, n_1 \leqslant n_2 \end{cases} \tag{8}$$

What indicates that AGH is eradicated? Is it when the state no longer receives positive reports in one year? Our answer is no, because whether AGH can be found and reported is a probability problem, there is a large contingency. Therefore, even if there is no positive report in a certain year, AGH may still exist. However, what we can accept is that if there is no positive report in a certain year, there is a high probability that AGH is eliminated. Therefore, what we need to do is estimate this probability. Based on our prediction and update model of the AGH's dispersal, this is actually

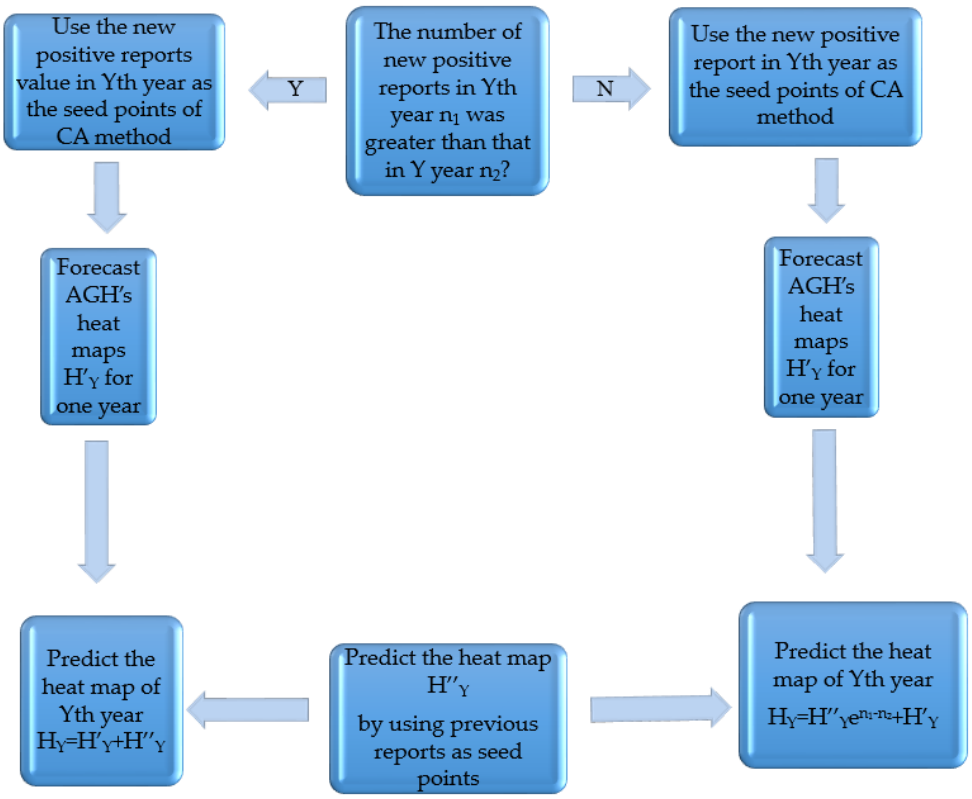



Figure 12: Flow chart of model update

very easy to quantify. The quantification formula is as follows.

$$P^{Y}_{eradicated} = 1 - \frac{\sum h_i^Y}{\sum h_i^{Y-1}} \tag{9}$$

where $P^{Y}_{eradicated}$ represents the possibility of AGH being eradicated in $Yth$ year, $h_i^{Y-1}$ represents the $ith$ heat degree of $H_{Y-1}$, $h_i^Y$ is similar.

Combined with Eq. 8 and Eq. 9, we can see that when the number of positive reports goes down, the probability of AGH being eradicated goes up every year. When there are no positive reports year by year, $1 - P^{Y}_{eradicated}$ goes down exponentially, consistent with the common sense.

# 8 Sensitivity analysis

Sensitivity analysis method is a common method to solve problems in modeling. For the establishment of this model, we used sensitivity analysis both in determining the impact of distance from the source point on migration results and the impact of climate factors on migration results. The basic idea is to change the conditions and then produce the experimental results. By comparing the experimental results, determine the optimal test plan.

## 8.1 Sensitivity analysis of probability determination in CA

First of all, we execute it according to the algorithm that we designed to reduce the probability with distance. The entire algorithm execution process can be shown in the Fig. 13. P is the probability of the influence caused by the distance. The larger the P, the greater the probability that the generated point falls within 15 kilometers. On the contrary, the greater the probability that the generated point falls within 15 kilometers.

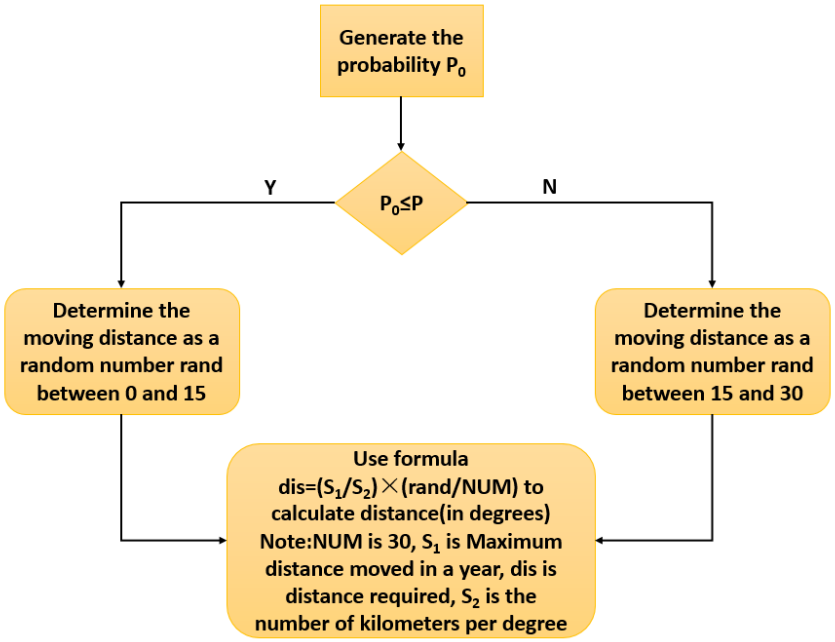
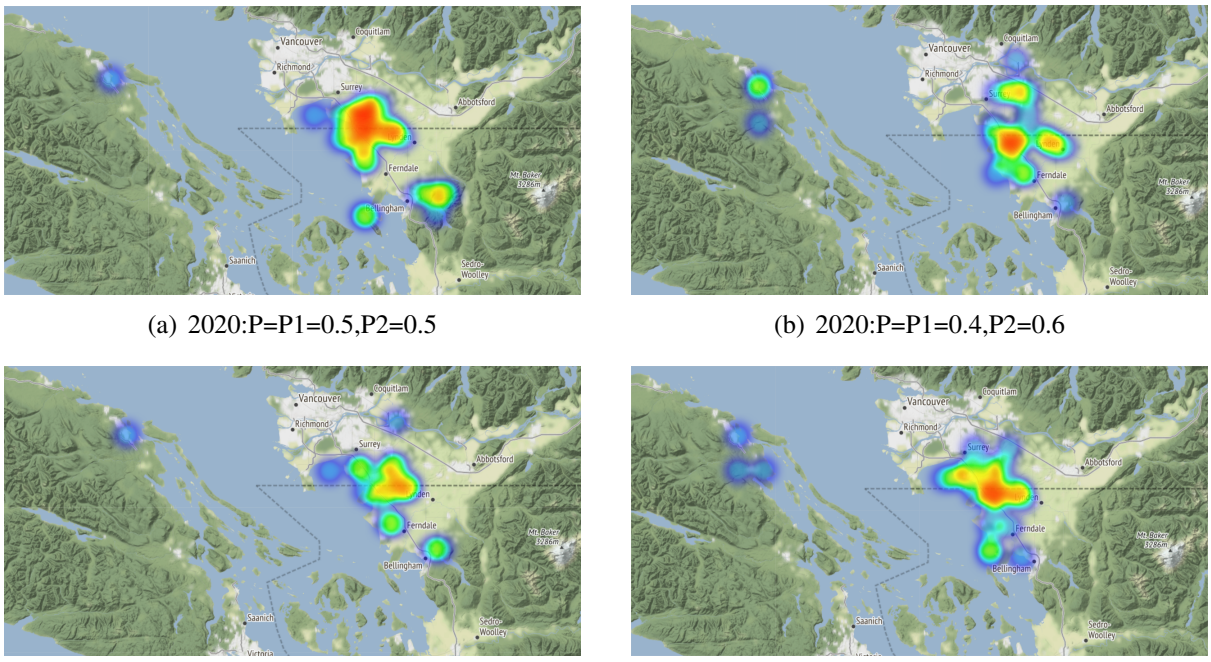



Figure 13: Flow chart of calculating migration distance

At the beginning we chose four P: 0.2, 0.3, 0.4 and 0.5. We set P1=P, P2=1-P according to the algorithm shown in Fig. 13 for two years of simulation and generate four heat maps as follows. Observing (a), we can find that when we set P to 0.5, the cell diffusion is no longer affected by distance. The result is like a circle centered on each source point. This is somewhat different from biological migration in nature. Observing (c) and (d), due to the low probability of propagation beyond 15km, the movement distance and expansion speed of the cell will be severely restricted. It can be known from the image analysis that the cell diffusion effect is not particularly good at this time. This also violates the laws of migration in biology. Observing (b), We found that the results of the experiment better reflect the influence of distance on biological migration. In summary, we use P=0.4, that is, P1=0.4, P2=0.6. As the best solution to this problem.

## 8.2 Sensitivity analysis on the climate effect factor

First, we introduce a variable as F. When the calculated climate influence factor is greater than this number F, we make it happen with a high probability. On the contrary, we set the probability of its occurrence as e divided by 128. The greater the value of F, the greater the influence of biological conditions on environmental constraints. Variable e is the maximum value of the surrounding climate influence factor. The flow of the entire algorithm refer to Fig. 15.




(a) 2020:P=P1=0.5,P2=0.5

(b) 2020:P=P1=0.4,P2=0.6

(c) 2020:P=P1=0.3,P2=0.7

(d) 2020:P=P1=0.2,P2=0.8

Figure 14: Heat maps of differnt probabilities

Similar to the above idea, we set F to 0,128 (the average value of the entire climate impact factor data), 800, and 1000. Then we performed two-year simulation on the previous model.Similar to the above idea, we set F to 0,128 (the average value of the entire climate impact factor data), 800, and 1000. Then we performed two years of simulation on the previous model and got the following heat maps. Observing (a), F=0 means that regardless of the influence of climate influence factors, the diffusion map he gets will not be constrained by geographical location. This is against the basic principle of biological adaptability. Observing (c) and (d), When F is too large, the probability of biological survival will be greatly reduced. According to image analysis, AGH will not have a good expansion effect after two years. This also is related to biological migration Violated by the law. So it is unreasonable to adopt a larger F. Observing (b), We found through many experiments that when F is set to an average value (128), we can get a good biological simulation effect.

# 9 Strengths and weaknesses

In terms of advantages, the content of this paper progresses layer by layer, each part is closely related, and the model framework is reasonable. Especially in our modeling process, we have fully mined the data information given. At the same time, before the establishment of the model, we have made a horizontal comparison of many model methods of the same type, and we have selected the better method. Moreover, in the process of modeling, we have also carried out reasonable and clever transformation and simplification to the practical problems, such as the evolution law setting when using the CA model. Nevertheless, we should also see that there are still some defects in our method. For example, because it is difficult to obtain the relevant altitude data on the open platform, we did not consider the influence of altitude in the simulation. In addition, due to the

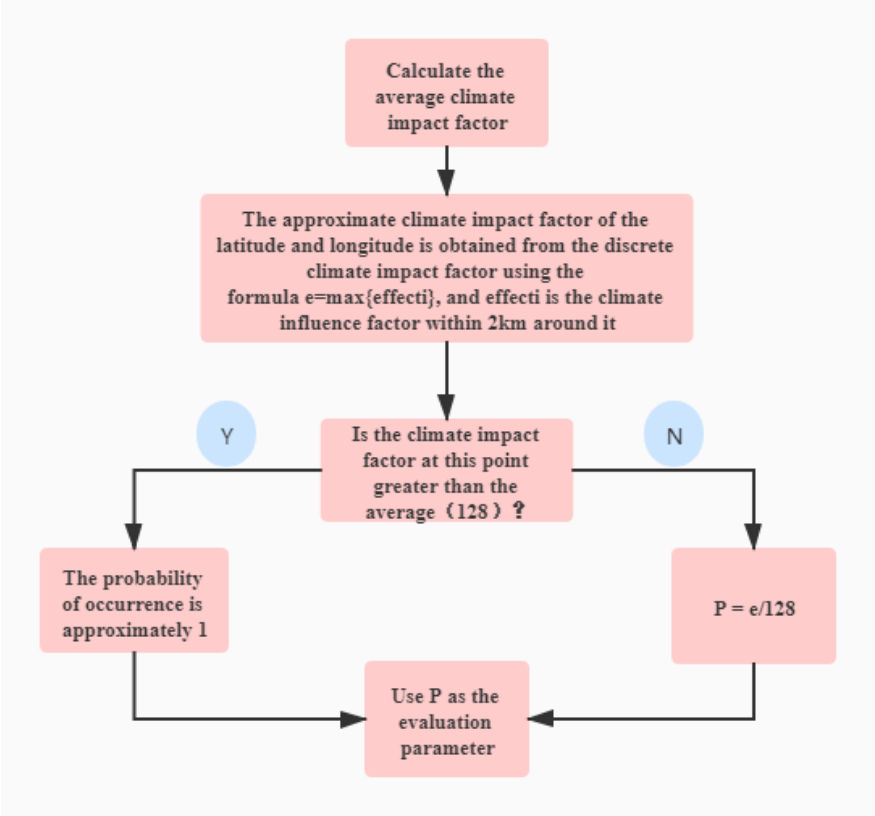



Figure 15: Climate impact factor evaluation algorithm

high spatial accuracy set by us, the complexity of the later operation is very high. In addition, there are too few positive reports in BP training, which may lead to some problems in classification.

# 10 Conclusions

To interpret the data provided by the public reports and help the state to prioritize these reports for additional investigation, we proposed a dispersal model to predict the distribution of AGH over time using CA mehtod, established two evaluation models to evaluate the likelihood of a mistaken classification and the investigation priority level of the report based on BP neural network and comprehensive assessment, put forward a model update method to use the new reports to improve our prediction results , and finally we proposed a method to estimate the probability that AGH is eradicated based on our prediction and update model of the AGH's dispersal. All in all, this paper make full use of the information hidden in the data, and the methods are more colse to the reality and common sense. Certainly, we have to acknowledge that our mehtods have some weaknesses, and these are the points we want to improve in the future.

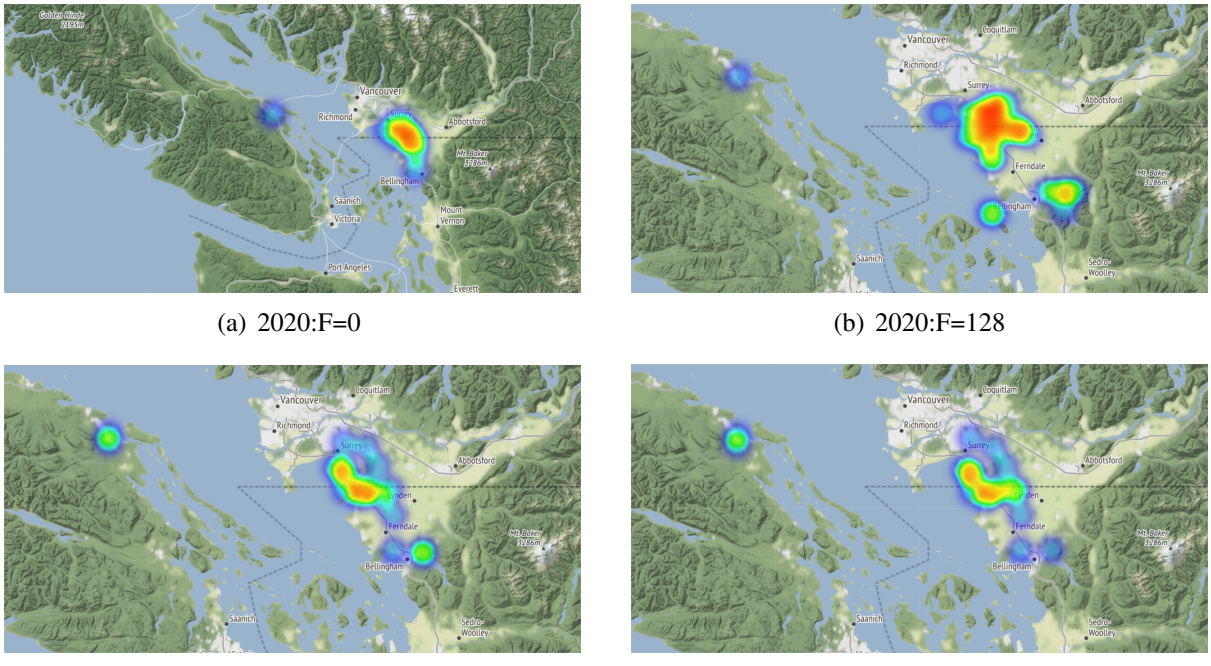
# References

[1] Gengping Zhu, Javier Gutierrez Illan, Chris Looney, David W. Crowder. Assessing the ecological niche and invasion potential of the Asian giant hornet. Proceedings of the National Academy of Sciences Oct 2020, 117 (40) 24646-24648. Doi: https://doi.org/10.1073/

pnas.2011441117

[2] Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J. Learning representations by back-propagating errors. Nature. 8 October 1986, 323 (6088): 533536. Doi: https://doi.org/10.1038/323533a0

[3] Wikipedia: Backpropagation




(a) 2020:F=0

(b) 2020:F=128

(c) 2020:F=800

(d) 2020:F=1000

Figure 16: Heat maps of differnt climate impact factor

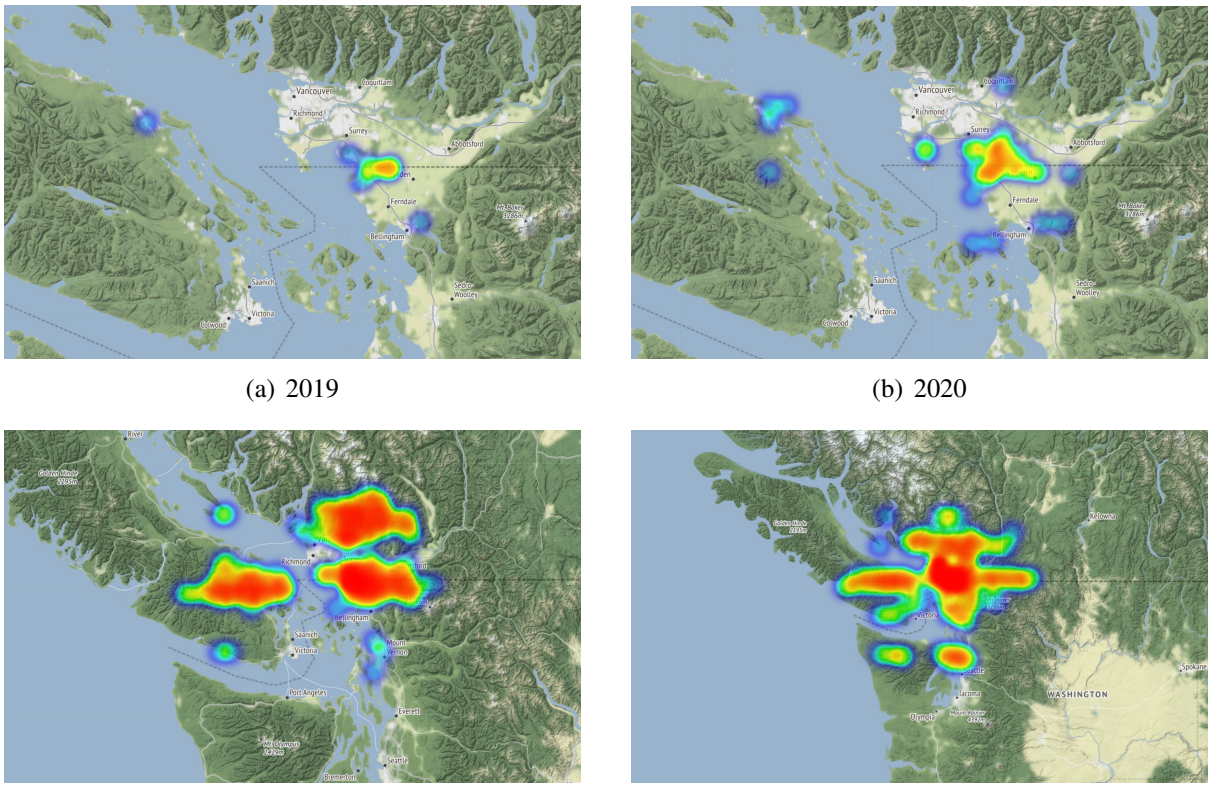
# Appendices

## Appendix A First appendix

### Memorandum

Lately, a colony of the Asian giant hornet (AGH) was discovered on Vancouver Island. The nest was quickly destroyed, but since that time, several confirmed sightings of the pest have occurred in Washington State. It was alarming that the AGH poses a serious threat to apiculture, and this species is considered an actionable quarantine pest. Thus, our team was saked to interpret the data provided by the public reports and help the state to prioritize these reports for additional investigation.

First of all, we have established the Dispersal Model of AGH. CA are a scheme for computing using local rules and local communication. To apply this method, initially, we should discrete the time and grid the space. As the AGHsdispersal depends on the migaration of the new queens, which migarate by year, we take one year as time step. We take 0.1km as the side length of a grid. Let's look at the prediction of AGH diffusion over time.




(a) 2019 (b) 2020

(c) 2025 (d) 2030

Figure 17: Dispersal heat maps of AGH predicted using CA mehtods

Secondly, in our view, evaluating the likelihood of a mistaken classification and investigation

priority level of reports are of the same essence and the last is based on the former. In many evaluation models, such as AHP model, the relationship between evaluation objectives and evaluation indexes is often linear, and the weight of evaluation indexes determined by them is often difficult to avoid the problem of being too subjective, which often leads to the evaluation results that cannot accurately reflect the actual situation. At the same time, we have a lot of real reports or disaggregated data to draw on, and this data are inherently informative. In view of this, we chose BP neural network model to do the evaluation. Then, we addressed aspect 3 based on aspect 2. Due to the harm caused by the AGH at different distances from the the report is far apart. So, we took the likelihood, the average value of the predicted heat degree within 1km of the report and the predicted heat degree occurring within 8km but beyond 1km respectively as the evaluation indexes and establish a comprehensive evaluation model using these. Then, We prioritized the unverified reports with this model, which is shown in Fig. 18.

Thirdly, considering that positive reports will decrease or even disappear under the influence of human factors. So, we compared the new reports in one year with those in the previous year to determine whether the AGH trend is increasing or decreasing. Then, we decided to add whether a superposition or attenuate process to the CA model according to the trend, so that the predicted distribution can be closer to the actual situation.

Fourthly, we know that whether AGH can be found and reported is a probability problem. What we can accept is that if there is no positive report in a certain year, there is a high probability that AGH is eliminated. So, in this paper, we proposed a method to estimate this probability based on our prediction and update model of the AGH's dispersal.

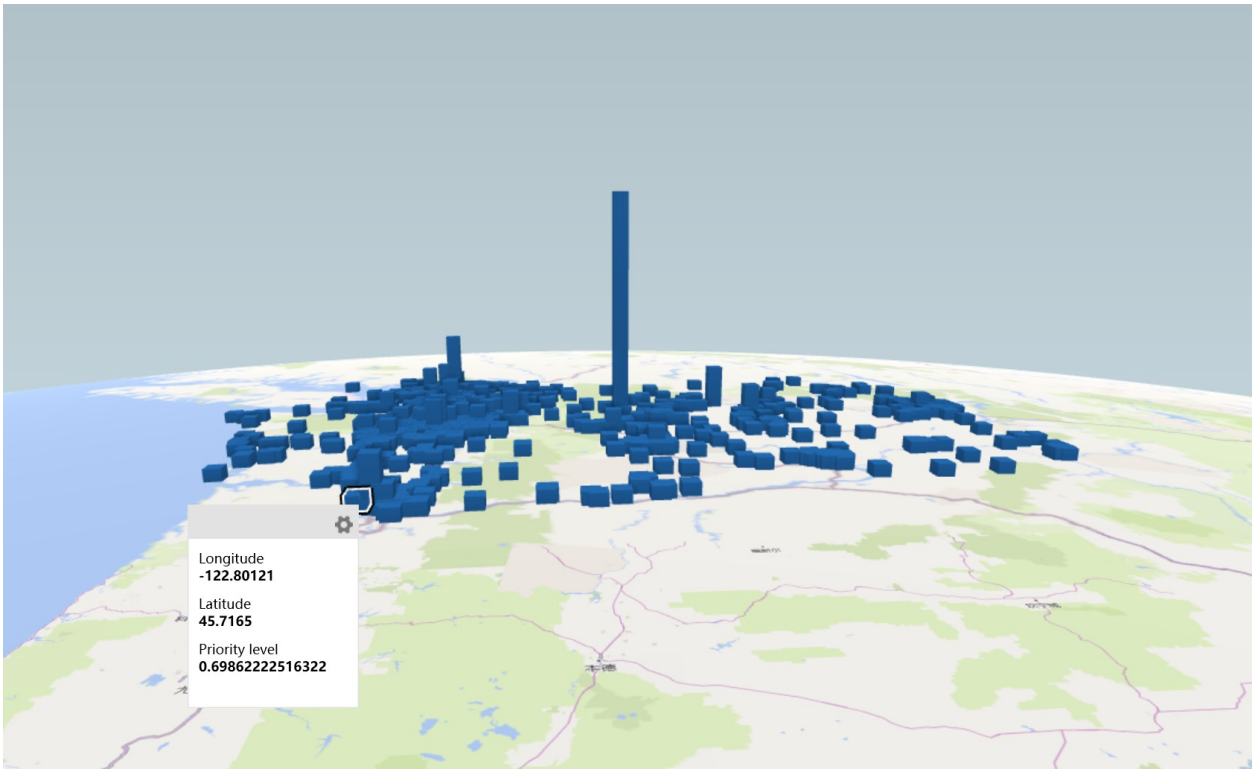



Figure 18: Priority levels of unverified reports

# Appendix B Second appendix

Here are simulation programmes we used in our model as follow.

### Core code for the CA Model and Investigation Priority Model:

```
#Some core code
#Simulation code
starttime = 43466#2019-01-01
newdate = []
#N:number of processing years
for j in range(len(positive)):
  newdate.append(0)
  for i in range(1,365 * N):
    for j in range(len(positive)):
      newdate[j]+=1
      if newdate[j] % 365 == 0:
        if random.randint(0,9) <= 3:
          ran_num = random.randint(0, 30.0)
        else:
          ran_num = random.randint(0, 15.0)
        vv = ran_num / 30 * v
        if isOK([positive[j][0] + vv, positive[j][1]]) == 1:
          positive.append([positive[j][0] + vv, positive[j][1]])
          newdate.append(0)
    #...The same is true for other directions
def isOK(loc):
  ran=random.randint(0,128) #128 is the average
                  # value of biological ecological factors
  temp=-1
  for i in range(len(la_we)):
    if get_distance_hav(la_we[i], lo_we[i], loc[0], loc[1]) <= 2:
      temp=max(temp,effect[i])
  if ran > temp:
    return 0
  return 1;
ans = []
for i in range(len(la_final)):
  num = 0.4 * float(likehood[i]) + 0.3 * (max2 - float(posfinal[i]))
           / max2 + 0.3 * float(negfinal[i]) / max1
  ans.append([la_final[i],lo_final[i],num])
final = np.array(ans)
data = pd.DataFrame(final)
writer = pd.ExcelWriter('final.xlsx')
data.to_excel(writer, 'final', float_format = '%.5f')
```

### Core code for the BP neural network:

```
# ()
def g(z,deriv=False):
    if deriv:
        return z*(1-z)
    return 1/(1+np.exp(-z))
#
```

```
def model(x,theta1,theta2,theta3):
    z2 = np.dot(x,theta1)
    a2 = g(z2)
    z3 = np.dot(a2, theta2)
    a3 = g(z3)
    z4 = np.dot(a3, theta3)
    a4 = g(z4)
    return a2,a3,a4
# (BP)
def BP(a1,a2,a3,a4,theta1,theta2,theta3,alpha,y):
    delta4 = a4 - y
    delta3 = np.dot(delta4,theta3.T)*g(a3,True)
    delta2 = np.dot(delta3,theta2.T)*g(a2,True)
    deltatheta3 = (1/len(y))*np.dot(a3.T,delta4)
    deltatheta2 = (1/len(y))*np.dot(a2.T,delta3)
    deltatheta1 = (1/len(y))*np.dot(a1.T,delta2)
    theta1 -= alpha*deltatheta1
    theta2 -= alpha*deltatheta2
    theta3 -= alpha*deltatheta3
    return theta1,theta2,theta3
#
def gradDesc(x,y,alpha=0.7,max_iter=10000,hidden_layer_size=(17,8)):
    m,n = x.shape
    k = y.shape[1]
    theta1 = 2 * np.random.rand(n,hidden_layer_size[0]) - 1
    theta2 = 2 * np.random.rand(hidden_layer_size[0],hidden_layer_size[1]) - 1
    theta3 = 2 * np.random.rand(hidden_layer_size[1],k) - 1
    j_history = np.zeros(max_iter)
    for i in range(max_iter):
        a2, a3, a4 = model(x,theta1,theta2,theta3)
        j_history[i] = costFunc(a4,y)
        theta1, theta2, theta3 = BP(x, a2, a3, a4, theta1, theta2, theta3, alpha, y)
    return j_history,theta1, theta2, theta3
```