Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2101603

Comprehensive consideration for efficient pest control

Summary

The State of Washington has created helplines and websites to report sightings of Vespa mandarinia, now the state wants to implement efficient pest control with limited resources.

For problem one, we discretize the data through geographic information. We divided the map into the 20362 pieces of square units. Then we use the Maximum Entropy Method to analyze 20 kinds of environment factors such as precipitation, temperature, altitude, etc., and generate possible distributions of this pest. Based on these methods, by adopting the natural growth model, we build up cellular automata to simulate the natural growth and spread process of Vespa mandarinia and generate a possible spatial distribution, based on current distribution. We figure out the swarm tend to move to the northwest and northeast.

For the second problem, we first apply data preprocessing on files and notes including image segmentation by CLAHE and GrabCut. Then we establish a Bayesian Belief Network to estimate the conditional probability for a true report where the parent nodes are environment influence from problem one, the child nodes are insect features. The likelihood of a mistaken classification can also be calculated. Body features are extracted by linear classification in RGB space. Nest and length features are extracted from report notes using a flexible regular expression. The average true report likelihood of a positive case is 0.7407, and for the negative case, the number is only 0.3379 which means the model's performance is reliable.

For problem three, we apply our model to existing data, it will filter out 66.55% negative report and 50.98% of the unverified report under the 0.5 probability threshold of the true report. And for the unprocessed report, named them by reverse chronological order, reports 5 and 7 have high priority, reports 1, 2, 11, 15 have medium priority for the specialist to check.

For problem four, the update of cellular automata (CA) and maximum entropy should only conduct when the report is positive. For CA we adopt the hypothesis testing to reduce the difficulty and frequency of model update, and the efficiency of the update strategy also shows some stability. For maximum entropy, under the assumption that there is no significant change in climate, we use the PCA method to process the geographic data and determine whether the newly added point is an outlier by calculating the Cook's Distance of it, to decide whether to update our model. The update of the Bayesian Belief Network simply updates statistic data.

Finally, based on the model of the first question, we make a certain estimate of the number of populations through the distribution of the population at a certain time. If the model shows a significantly low level of population density while there are no new reports that appeared in a year. then we have confidence to say that it has been wiped out.

Keywords: Bayesian Belief Network; Cellular Automata; MaxEnt;

Contents

MEMORANDUM

To: Washington State Department of Agriculture
From: Team #2101603
Date: Feb. 8*th*, 2021
Subject: Confirmation, computation and classification of Vespa mandarinia

Dear agricultural specialists, we are honored to inform you about our achievement for Vespa mandarinia spreading analysis, sighting report preprocesses, mistaken classification modeling, priority rule, and eradication evidence.

We extracted all reports marked as Positive from all 4441 reports and obtained geographic information on the distribution of these pests. Then we obtained 20 climate variables including altitude, precipitation, and temperature from all over the world through the authoritative WorldClim climate database. Combining climate and geographic information, we used the widely used Maxent niche model to analyze the habitat suitability of such pests in various regions. We were surprised to find that the model reflects that this pest is suitable for survival along the coast of Vancouver Island and northwestern Washington. This result is also consistent with some existing news reports. Therefore, it is necessary to increase pest control in these areas.

We constructed a prediction mechanism for the Vespa mandarinia by simulating the movement and habit of Vespa mandarinia. This model helps us know the possible position and the areas that may be affected in the future. As time goes by the accuracy of our model may be decreased, so we do not suggest you predicting the situation after more than one year. Additionally, based on our model, there are signs show that a group of the Vespa mandarinia are very likely to enter Canada from the northeast or northwest of Washington. Therefore, an important issue is to strengthen the prevention of insect disasters in the northern region and strengthen cooperation with relevant Canadian departments.

To achieve a better prediction effect, we recommend that you update the colony position parameters recorded in the system on time. Besides, our system is highly fault-tolerant for the discovery and prediction of bee colonies. When necessary, we only recommend that you update the system parameters in the following situations：

- There is more than one year since the last update of the system parameters.
- There are positive reports more than 7.3 km away from our predicted colonies.

We established a report processing method and a likelihood estimation model to classify the reliability of reports for you to focus on valuable information and prioritize these public reports. It is a kind of Bayesian Belief Network:

$$P(\mathbf{x} \mid \mathbf{e}) = \alpha P\left(\mathbf{e}^{\mathcal{C}} \mid \mathbf{x}\right) P\left(\mathbf{x} \mid \mathbf{e}^{\mathcal{P}}\right)$$

where $P(\mathbf{x} \mid \mathbf{e})$ represents the likelihood of real Vespa mandarinia report under evidence $\mathbf{e}$. Parents nodes $\mathbf{e}^{\mathcal{P}}$ are environment influence and children nodes $\mathbf{e}^{\mathcal{C}}$ are report insect feature. When $P(\mathbf{x} \mid \mathbf{e}) \geq 0.677$, this report has high priority and the medium priority threshold is 0.592.

To avoid the waste of resources, we build a report processing method to quickly filter out the low-quality report. It contains a file preprocess to extract insect images and key phrases in reports.

To better estimate the likelihood and prioritize investigation of the reports, we establish the Bayesian Belief Network, follow our model you can:

- Take full consideration of environmental influence such as location, climate, and previous true report location.
- Calculate the conditional probability with limited information. For example, when only knows the head color of the report insect, our model is able to give credible likelihood estimation.
- Automatically pigeonhole reports with different features such as high priority or high-quality image.

Then we analyze the unprocessed data, with the feature extraction and Bayesian Belief Network, report 5 and 7 have high priority, report 1, 2, 11, 15 have medium priority for specialists to read and reply.

To maintain the long-term validity of the model, our model will also need more relevant samples to update itself. For the ecological niche model, you can compare each newly-added Positive sample with the existing sample (for detailed comparison methods, please refer to the report). As long as the degree of deviation is not too great, you can continue to use the current model to calculate the suitability of species in various places until one day a new sample appears in an unprecedented place.

We sincerely hope that the trouble caused by the Vespa mandarinia to the local area can be resolved as soon as possible!

I. Introduction

1.1 Background

Biological invasion is a serious and intractable problem for any country. If left untreated, these biological invaders can colonize, reproduce, and overwhelm the native ecosystem. In more serious cases, the destruction of the ecological environment will also hit the local agriculture and animal husbandry, and ultimately affect the economic benefits of the society. The Asian bumblebee is the world's largest bumblebee, and since it was first spotted on Vancouver Island in September 2019, there have been numerous sightings of the species on Vancouver Island and in nearby Washington State, most of which have sadly turned out to be other species of bees. Limited resources are being wasted due to the lack of sound prediction of bumblebee distribution and good identification of sighting reports. In conclusion, proper identification of sighting reports and prediction of bumblebee distribution is an urgent problem to be solved.

1.2 Our works

- Task 1
  We construct an environment suitability map with Maximum Entropy method based on the geographic information and constructed a predict system based on the environment information and Cellular Automata. Combining these two aspects we predict the possible spread direction of the pest.
- Task 2
  Establish the workflow of preprocessing and build up Bayesian Belief Network to calculate conditional probability. Set environment influence from task 1 as parent nodes and body, nest feature as child nodes.
- Task 3
  Apply the model on existing data, filter out less likely report and analyze the performance. Set priority for unprocessed data.
- Task 4
  Using hypothesis testing, we established an update strategy for cellular automata. With PCA and Cook's Distance, we devised a strategy to update Maximum Entropy. Update Bayesian Belief Network based on statistic data.
- Task 5
  Based on Maximum Entropy method and Cellular Automata. We constructed an estimate method to estimate the density of the Vespa mandarinia. Combining with its habits, we set a principle to determine whether it has been eradicated.

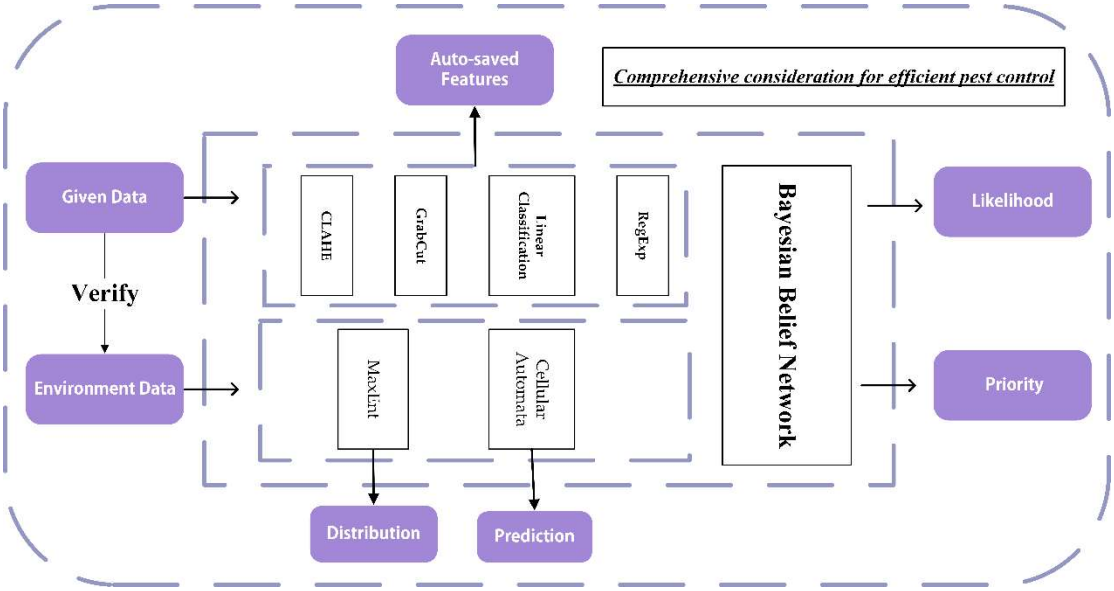

Figure 1: Full workflow

II. Restatement of the Problem

Based on the background information provided in the problem document and the associated constraints, we address the following questions:

- For the spread of this pest, give a model that can be interpreted and obtain the accuracy of the prediction
- Use existing dataset files as well as image files to create a model that can analyze and discuss the possibility of predictive misclassification.
- Use the obtained model to analyze how to prioritize the reports that are most likely to be positive sightings
- Based on the existing model, determine how to update the model with increased reporting and how often it should be updated.
- Based on the existing model to discuss how to show that the pest has been eradicated in Washington State

III. Basic assumption

1. The climate will not change significantly in a short time.
2. Nest will not be too far from the positive sightings.
3. Queens have roughly the same fertility.

IV. Symbols

Symbol	Definition
la	Latitude
lo	Longitude
D	Cook's Distance
C	Cellular
d	Distance form a nest
p	Death rate

V. Models

5.1 Analysis and Solving of Question One

5.1.1 Model Preparation

We hope to figure out the movement of swarm and predict its possible next destinations. As a typical habit of Vespa mandarinia, the direction of this swarm is often determined by the movement of queen. Moreover, the materials provided indicate that the hornet has limited flight capability and Vespa mandarinias' activity range is restricted by the location of the hive (i.e. queen), which means the **positive reported position cannot be too far from a nest**. So, an intuitional idea of researching the swarm's movement is trying to simulating the movement of queen. In addition, environmental parameters such as temperature and humidity are also important for Vespa mandarinia's surviving. In consideration of these effects, we propose this model to **predict a possible distribution of Vespa mandarinia in its active period**. Now we introduce cellular automaton (CA) and maximum entropy (MaxEnt) method.

Model preparation for cellular automata

Cellular Automata (CA) was proposed in the early 1950s by J. von Neumann, the father of computers, to simulate the self-replication function of living systems. Since then, Stephen Wolfram has conducted in-depth research on the theory of cellular automata. For example, he conducted several meaningful research on the models generated by all 256 rules of one-dimensional elementary cellular machines, and divided the cellular automata into four types: stationary, periodic, chaotic, and complex.

Cellular automaton is based on discrete spatial layout and discrete time intervals to divide cells into finite states. The evolution of an individual cell's state is only related to its current state and the state of its local neighbors. L-system, lattice gas model, Lattice Boltzmann Method, traffic flow model, etc., are all embodiments of cellular automata, which have important theoretical significance and practical application value.

Model preparation for Maximum Entropy algorithm

The Maximum Entropy algorithm combines the spatial explicit occurrences of the target species with a set of environmental predictor variables to estimate the habitat suitability of the species. From the article[1], we know that given the localities $x_1, \ldots . x_j$ chosen from a set of discrete grid cells X under an unknown probability distribution π and features $f_1, \ldots, f_n$ where $f_i : X \to R$, it is practicable to seek an approximation $\hat{\pi}$ under which every f_i's expectation is very close to the empirical average of f. Here we define the entropy of distribution p on X

$$\mathrm{H(p)} = -\sum_{x \in X} p(x) \ln p(x) \tag{1}$$

The principle of maximum entropy states that when we only have partial knowledge of the unknown distribution, we should choose the probability distribution that conforms to this knowledge but has the maximum entropy value.

5.1.2 Model Establishment

Data preparation

Here we selected all the reports of sightings that were confirmed to be Positive from the document and extracted the corresponding longitude and latitude data of this place. After confirming the validity of the data, the relevant ecological and environmental factors were downloaded from the Worldclim database [2] including 20 related variables such as precipitation, climate, and altitude in each period. Table 1 shows the detailed information.

Table 1:20 environment variables

Environment variables	Definition
bio1	The annual average temperature

bio2	Monthly mean of diurnal temperature difference
bio3	Isotherm(bio2 / bio7 × 100)
bio4	Variance of temperature change
bio5	The highest temperature of the hottest month
bio6	The lowest temperature of the coldest month
bio7	Range of annual temperature variation
bio8	Mean temperature in the wettest season
bio9	Mean temperature of driest season
bio10	Mean temperature of the warmest season
bio11	Mean temperature of the coldest season
bio12	Average annual rainfall
bio13	The rainfall in the wettest month
bio14	The rainfall in the driest month
bio15	Variance of rainfall changes
bio16	The rainfall in the wettest season
bio17	The rainfall in the driest season
bio18	Average rainfall in the warmest season
bio19	Average rainfall in the coldest season
Alt	Altitude

Solution of MxEnt algorithm

The Vespa Mandarinia distribution record data and 20 environment variable data are imported into the Maxent3.3 model. The model is established with 75% of the actual distribution data selected randomly, and the remaining 25% of the actual distribution data are used to validate the model. The other parameters are the default values, and the predicted values are continuous raster data (0 ~ 1). The obtained ASC raster data was converted and imported into ArcGIS for analysis, and the suitability of Vespa Mandarinia in Washington State and its surrounding areas (as shown in Figure 2) was obtained. The map was divided into units of 100 square kilometers to get the suitability of the corresponding position of each unit.

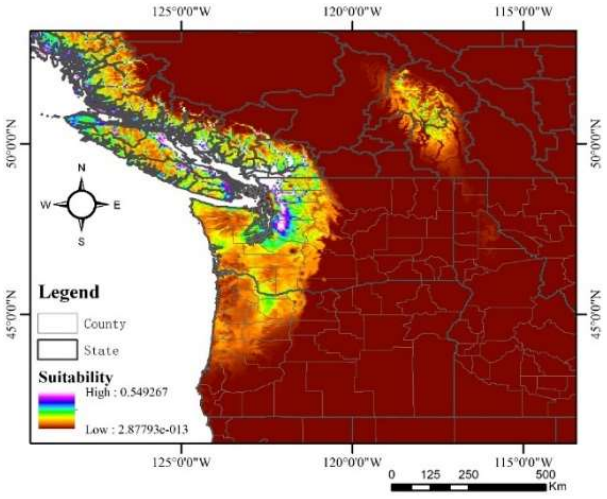

Figure 2: The result of MaxEnt Algorithm

Validation of the results

The accuracy and reliability of the MaxEnt model prediction can be verified by using the ROC curve of the training set and test set data. Figure 3 shows that the ROC curve of the data set is far away from the ROC curve of the random prediction model, and the corresponding AUC value of the training data and the test data is 0.989 and 0.983 respectively, which is significantly higher than the AUC value of the random prediction model (0.5). This indicates that the MAXENT model can better predict the potential distribution of Vespa Mandarinia in China by combining the existing distribution data of Vespa Mandarinia with the filtered environment variables, and the prediction results are reliable.

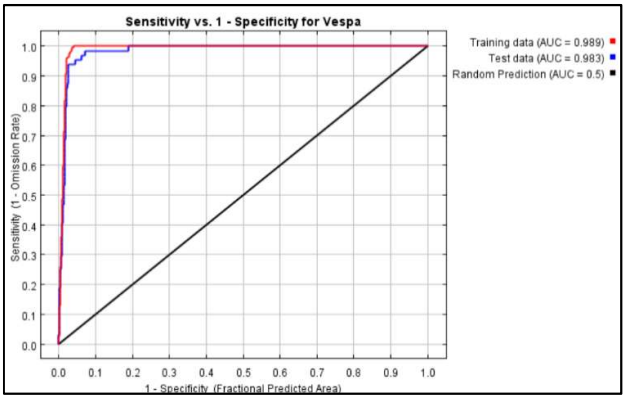

Figure 3: Algorithm Sensitivity

Solving Cellular Automata

In this section we are going to do preparations for the following parts. Since the coordinate data is continuous, to simplify our work we decided to integrate data first. Considering the habits and flight capacity of *Vespa mandarinia* we set the diameter of each square unit be 1 km to discretizate the area between the longitude: (-130°, -116°) and the latitude (45°, 53°) according to the result of Maxent. Now, we denote the discrete map as C.

Each element C_i of C satisfies: [12]

$$\begin{cases} C_i = 1, if\ there\ is\ Vespa\ mandarinia's\ nest \\ C_i = 0, otherwise \end{cases}$$

And the neighborhood centered at C_i is called the Moore neighborhood of C_i denoted as M_i.

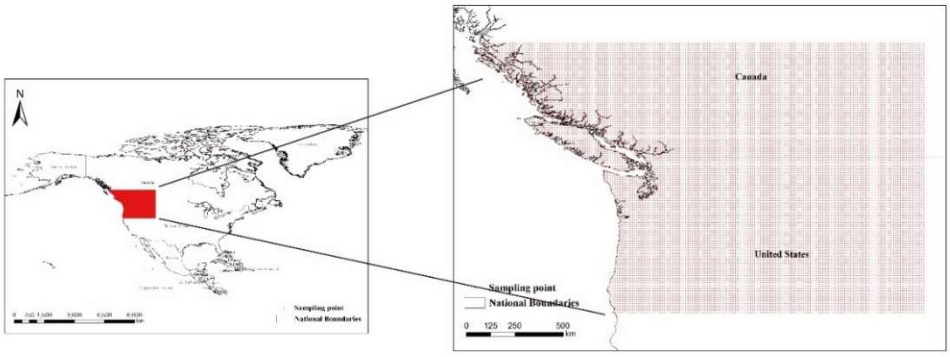

Figure 4 Cellular Distribution

This process is designed to **simulate the movement of the queen** which determines the movement direction of the colony. According to the provided material, the activity of *Vespa mandarinia*'s colony has obvious seasonality (They only move the nest once a year during spring), and the hornet's range of activity is typically limited (They only fly 1-2km on average never fly more than 8km away from the nest). **Thus, we assume that the undiscovered nest(s) is near the positive sightings, and the movement of Asian giant hornet follows Two-dimensional Gaussian distribution during active period**. Two important tasks are to determine the position and the number of colonies.

In this section we mainly talk about how to simulate the movement of swarms. **T**he key of this process is to **simulate the movement of the queen** which determines the movement direction of the colony. According to the provided material, the activity of *Vespa mandarinia*'s colony has obvious seasonality (They only move the nest once a year during spring), and the hornet's range of activity is typically limited (They only fly 1-2km on average never fly more than 8km away from the nest). **Thus we assume that the undiscovered nest(s) is near the positive sightings, and the movement of Asian giant hornet follows Two-dimensional Gaussian distribution during active period**. Two important tasks are to determine the position and the number of colonies.

Step 1: Now we are to figure out the position of colony, and we consider the positive sightings detection date. Since the hornets' activity has obvious periodicity, they only migrate once a year. The positive sighting position are mainly located at latitude (48.7775°, 49.1494°), longitude: (-123.9431°, -122.4186°) which indicates that its distribution is spatially dense. So an algorithm based on the spatial information might be helpful to locate the activity center of *Vespa mandarinia* We calculated the dendrogram of these point. Then we decided to use k-means where $\mathrm{k} = 3$ to give a possible located region of the nest.

Step2: Then we are going to build the cellular automaton. [hz1] First we should figure out how the population of *Vespa mandarinia* grows, then we determine the movement pattern and rule of a cell. [13][14]

Finally, we introduce the natural growth model.

Population growth often obey this PDE called Logistic equation

$$\begin{cases} \dfrac{dx}{dt} = rx\left(1 - \dfrac{x}{x_m}\right) \\ x(0) = x_0 \end{cases}$$

where x_m is the theoretical maximum population.

For discrete cases we have an update formula: [15]

$$\mathrm{y}_{k+1} = y_k + r * y_k\left(1 - \frac{y_k}{N_m}\right)$$

Since the population reaches its peak in August, by fitting we have $\mathrm{r} = 0.988$.

Then we are going to define the movement rule of cellular.

Combined with the traditional cellular rule, we give these two rules[hz2]:

1. If cell C1 was 1, and C1 becomes 0 after iteration, it means that the original queen died after reproducing.
2. If the cell C1 was originally 0, the probability that it has p becomes 1. The specific probability p is calculated as follows:

   $$\mathrm{p} = \mathrm{k1} \times \mathrm{k2} \times \mathrm{E}^{k3}$$

   where E is the environment livability at C1.

K2 depends on the number N of cells currently 1 in the Moore neighborhood of C1, and the specific values are as follows:

Table 1

N	0	1	2	3	4	>4
K2	0	0.0010	0.0015	0.0020	0.0005	0.0001

After fitting with the population growth model, we have $k3 = 1.0,\ k2 = 1.7$.

To stimulate the spatial distribution of *Vespa mandarinia*, we conducted 100 times of our experiment and documented the result for each time including the counted cellular and the position of it. Here is part of our experiment result:

Table 2

Experiment Round	1	2	3	4
Totally Counted Cell	9	13	15	12

After 100 time of experiment, we got the position information of each time. After counting the time of each cell, we can finally get the possible distribution of *Vespa mandarinia*.

By using the data of 2019 we can get a predicted distribution of 2020, we can check the distribution of 2020 and get the accuracy of our model. After 100 times experiment we finally get the accuracy of our model reaches 86.2161%.

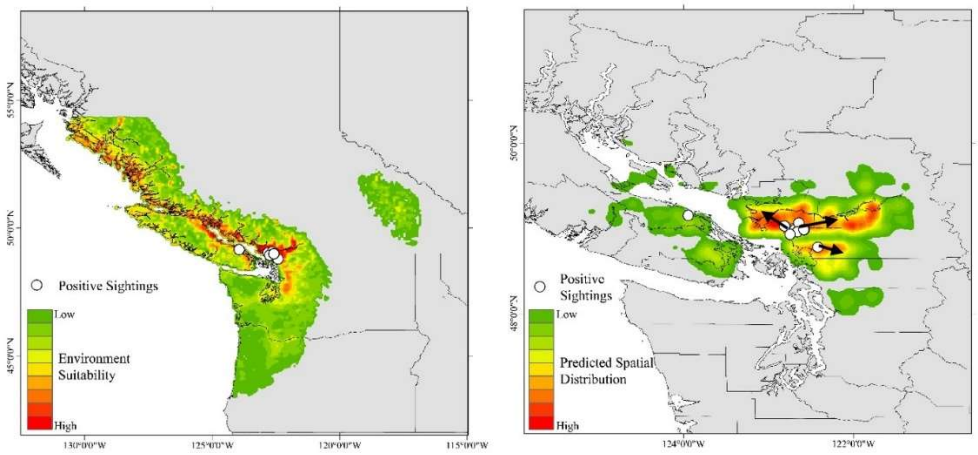

Figure 5 Predicted Distribution

From this simulated distribution map, we can easily figure out that the movement of *Vespa mandarinia* in the United States have three trends. One heads northwest and one heads northeast Both of them will probably get into Canada. The last one seems to stay in the U.S.

Sensitive Analysis

Now we are going to look at the influence of the initial point's position on our model accuracy

Now we can easily check this by counting how many points are contained in model's predicted region.

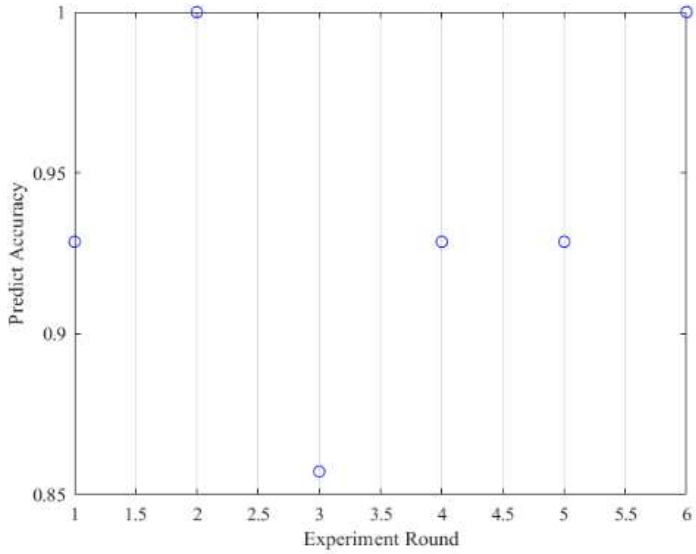

Figure 6 Predict Accuracy vs Experiment Round

So according to this figure the influence of the initial point is very limited on our model.

5.2 Analysis and Solving of Question Two

5.2.1 Data Pre-processing

With the data analysis of data set file provided and the image files provided in the reporting system, data of *Vespa mandarinia* is typically a kind of "label sparse data", which means only a small part of reports are real.

However, this phenomenon reflects the anxiety of government agencies and the waste of resources. Some files do not need to be exported to read and reply, and the processing order is supposed to be set base on the information it provides.

With this motivation, it is necessary to implement preprocessing on report notes and report files.

First, the preprocessing of files will make them easy to read, and automatically clean invalid data such as low insect proportion picture or fuzzy feature and then record them at a high/low quality image dataset. To achieve this goal, the file preprocessing will start from changing files into images using keyframe extraction and format conversion, at the same time discard non-image files. Then, we will implement mix data annotation image segmentation to filter out low insect proportion picture in a gesture to pick useful reports at a lower cost. Finally, images will be placed at two different datasets waiting for the following analysis.

The whole process of file pre-processing is shown in the figure below:

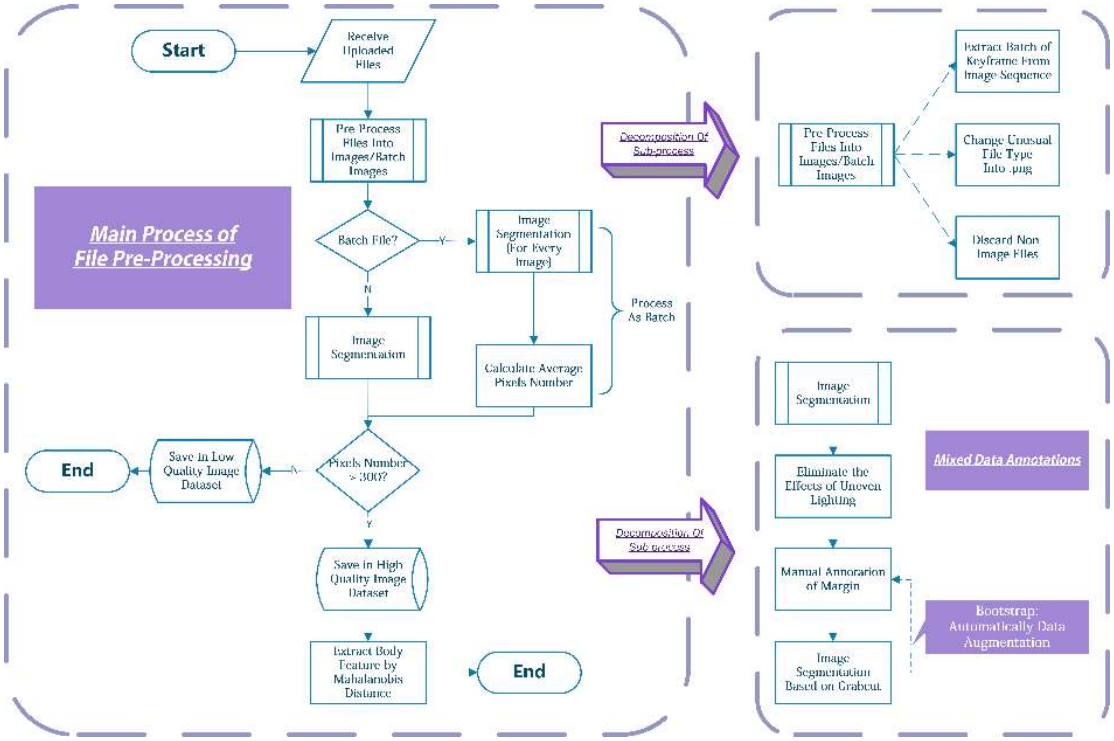

Figure 7.Flow chart of file pre-processing

The pre-processing of notes is more straightforward: to automatically extract useful information, we will automatically implement regex match after manually reading and labeling reports notes. The regex match results are helpful to reflect nest feature and body length feature.

These are motivations and central processes. The algorithms in the data pre-processing will be introduced in the following. To begin with, image segmentation algorithm is based on GrabCut which is an automated algorithm for RGB images' segmentation based on manually data annotation. [3] [4] This algorithm is based on iterative energy minimization. In the Gaussian mixture model, parameters are defined as θ, they are taken to be a full covariance gaussian mixture with K components. For pixels array $\mathbf{z} = (z_1, \ldots, z_n, \ldots, z_N)$, we add a label reflects either it is from the background or foreground model according to $\alpha_n = 0$ or 1.

So the Gibbs energy $\boldsymbol{E}$ of image segmentation is $\boldsymbol{E}(\alpha, \mathbf{k}, \theta, \mathbf{z}) = U(\alpha, \mathbf{k}, \theta, \mathbf{z}) + V(\alpha, \mathbf{z})$, where V is the smoothness term(C is the set of pairs of neighboring pixels):

$$V(\underline{\alpha}, \mathbf{z}) = \gamma \sum_{(m,n)\in \mathbf{C}} \mathrm{dis}(m,n)^{-1}[\alpha_n \neq \alpha_m]\exp - (2\langle (z_m - z_n)^2 \rangle)^{-2}$$

And U represent the penalty term of one pixel is labeled as background or foreground:

$$U(\underline{\alpha}, \mathbf{k}, \underline{\boldsymbol{\theta}}, \mathbf{z}) = \sum_{\boldsymbol{n}} \sum_{i=1}^{K} \pi_i g_i\left(\alpha_n, k_n, \underline{\boldsymbol{\theta}}, z_{\boldsymbol{n}}; \mu_i, \Sigma_i\right)$$

Here $\sum_{i=1}^{K} \pi_i = 1, 0 \leq \pi_i \leq 1$, and $g_i(x; \mu, \Sigma)$ represents the gaussian distribution:

$$g(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]$$

The algorithm of GrabCut to minimize $\boldsymbol{E}$ is as follows: [3]

Algorithm 1 GrabCut: Segmentation by iterative energy minimization

Input: T_B,T_U,α,$\mathbf{z}$

Output: $\mathbf{k}$,α,θ

1: **initialize:** Set $\alpha_n = 0$ for $n \in T_B$ and $\alpha_n = 1$ for $n \in T_U$

2: **generation:** Background and foreground GMMs initialised from sets $\alpha_n = 0$ and $\alpha_n = 1$ respectively.

3: **do:**

4: Assign GMM components to pixels: for each n in T_U,

$$k_n := \arg\min_{k_n} D_n(\alpha_n, k_n, \theta, z_n)$$

5: Learn GMM parameters from data z:

$$\underline{\theta} := \arg\min_{\theta} U(\underline{\alpha}, \mathbf{k}, \underline{\theta}, \mathbf{z})$$

6: Estimate segmentation: use min cut to solve:

$$\min_{\{\alpha_n : n \in T_U\}} \min_{\mathbf{k}} \mathbf{E}(\underline{\alpha}, \mathbf{k}, \underline{\theta}, \mathbf{z}).$$

Until: convergence

7: **End do**

At the initial state, T means user manually give a label of background T_B or foreground T_U. The following figure shows the effect of the program, which performs well under RGB space:

Figure 8.

With more data are annotated, the GMM model can be initialized more correctly, which is called “bootstrap” method. Image segmentation at RGB space will help in the following feature extraction such as “head’s color”. Later we calculate the body feature by measuring these features using weighted absolute distance [9]. After eliminating light disturbance using CLAHE[11], we train the Mahalanobis distance classifier by manually labeled data in RGB space.

For the preprocessing of notes, the algorithm is flexible regex match. We can extract many features from notes such as insect number, size, color, number of observations, location of sightings… The most useful data is the body length and nest feature which is hard to be extracted from picture.

To extract these features flexibly, we first apply synonym replacement by derivative dictionary[10], we set up the regex using key words location method. For example, body length can be extracted before “inches”, “feet” and we build a dictionary to extract nest location including “garden”, “forest”， “balcony”， “tree”…… The following chart shows some of key phrases’ frequency.

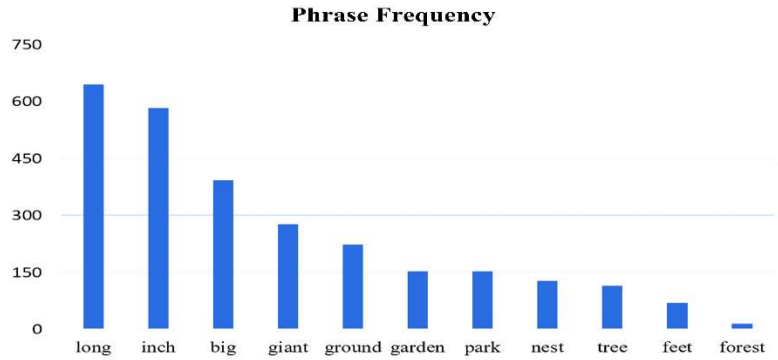

Figure 9.

5.2.2 Model Preparation

It is not an easy task to identify whether the report is a record of Vespa mandarinia, which involves several complicated factors. To simplify the problem, we assume that only two factors will affect the report, one is "Nature influence" and the other is "report feature". Then we build a Bayesian Belief Network (BBN) according to those relationships.

The mathematical model $BN=(G,\theta)$ of Bayesian networks consists of two parts: network structure G and network parameter θ. The network structure $G=\langle V,E\rangle$ is a directed acyclic graph with n nodes, and $V=\{V_1,V_2,\cdots V_n\}$ represents a node set in a directed acyclic graph that is a finite non-empty set, which has a one-to-one correspondence with the variable set $U=\{X_1,X_2,\ldots,X_n\}$. $E=\{V_k\rightarrow V_l,V_j\rightarrow V_j,\cdots\}$ represents the dependencies between nodes, where $k,l,i,j\in[1,n]$, if there is a directed edge from node Y to node X, then node Y is called the parent node of node X, node X is called a child node of node Y. The set of all the parents of node V_i in node-set V is represented by Pa_i. The set of all child nodes is represented by Ce_i, and the set of non-descendant nodes is represented by nd_i. The parameter $\theta=\{\theta_1,\theta_2,\cdots,\theta_n\}$ represents the conditional distribution table corresponding to the nodes in node set V, where $\theta_i=P(X_i\mid Pa_i)$ is the conditional probability distribution of node V_i, and if V_i has no parent node, the distribution is its edge distribution. [5] [7]

In Bayesian Belief Network, the value of nodes can be discrete or continuous. Now we use the continuous model.

5.2.3 Model Establishment

With model simplification, we only take environment influence and report feature into consideration to create Bayesian Belief Network. It is reasonable to set environment influences as parent nodes and report feature as child nodes for the report type according to their relationship in real world.

As instance, suppose we have a Bayesian Belief Network where parent node of X is {A， B}, child node of X is {C， D}, when estimating the probability of X, the parent node and the child node of X should be treated differently. Suppose we have evidence e: values of

variables at nodes other than X, $\mathbf{e}^{\mathcal{C}}$ means evidence from child node and $\mathbf{e}^{\mathcal{P}}$ means evidence from parent nodes, the Confidence Belief of x can be calculated: [5]

$$P(\mathbf{x} \mid \mathbf{e}) = \alpha \underbrace{\prod_{j=1}^{|\mathcal{C}|} P\left(\mathbf{e}_{\mathcal{C}_j} \mid \mathbf{x}\right)}_{P(\mathbf{e}^{\mathcal{C}}|\mathbf{x})} \underbrace{\left[\sum_{\text{all } \mathcal{P}_{mn}} P(\mathbf{x} \mid \mathcal{P}_{mn}) \prod_{i=1}^{|\mathcal{P}|} P\left(\mathcal{P}_i \mid \mathbf{e}_{\mathcal{P}_i}\right)\right]}_{P(\mathbf{x}|\mathbf{e}^{\mathcal{P}})}, \alpha = P(\mathbf{e})^{-1}$$

By the models we establish at problem one (Maxent and Cellular Automata), we can get the Average Appearance Possibility under N hypothesis of nest distribution (APP_N) as one of nature influence. We also count the report month (M) as another nature influence. These two factors are set as parent nodes. Then with the help of background information from Pennsylvania State University Extension that describes the insect [8]. We extract body feature: Head Feature (HF); Thorax Feature (TF); Abdomen Feature (AF) and feature from notes: Nest Feature (NF), Length Feature(LF) as child nodes. Parameters are set by program (APP_N) or frequency statistics. So, the Bayesian Belief Network for the likelihood of a mistaken report (X) will be set as:

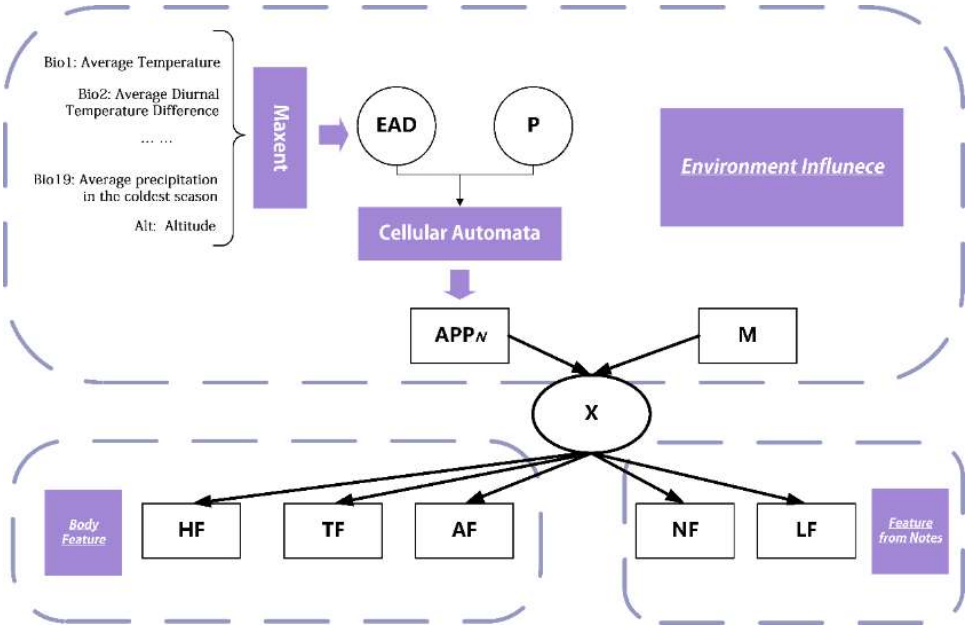

Figure 10 Flow chart of Bayesian Belief Network

5.2.4 Analysis and Conclusion of The Model

Since most reported sightings mistake other hornets for the Vespa mandarinia, our model based on report data's notes; position; report time, and the image file after pre-processing to extract report features and calculate conditional probability $P(\mathbf{x} \mid \mathbf{e})$, the mistaken likelihood is simply $L(\mathbf{x} \mid \mathbf{e}) = 1 - P(\mathbf{x} \mid \mathbf{e})$.

Based on the file preprocessing and background information from Pennsylvania State University Extension that describes the insect [8], we extract body feature, nest feature and length feature from images and report notes using distribution characteristics of trichromatic components and flexible regex match. For the parameters in Bayesian Belief Network, we use frequency statistics to record mistaken reports. Though in this "label sparse data" case, the mistaken report number is huge (2069/4400), it is easy to destroy the reliability of frequency statistics, so when updating statistics, we add the feature resolution as updating weight. In conclusion, the parameters of

feature calculation include three action:

- Credible feature & impossible for Vespa mandarinia: Classified as mistaken report, record other features in frequency statistics.
- Credible feature & possible for Vespa mandarinia: Calculate conditional probability with environment influence. Update frequency statistics with feature resolution as weight.
- Vague feature: Discard this feature in Bayesian Belief Network.

The full workflow of our model for question two is as the flow chart:

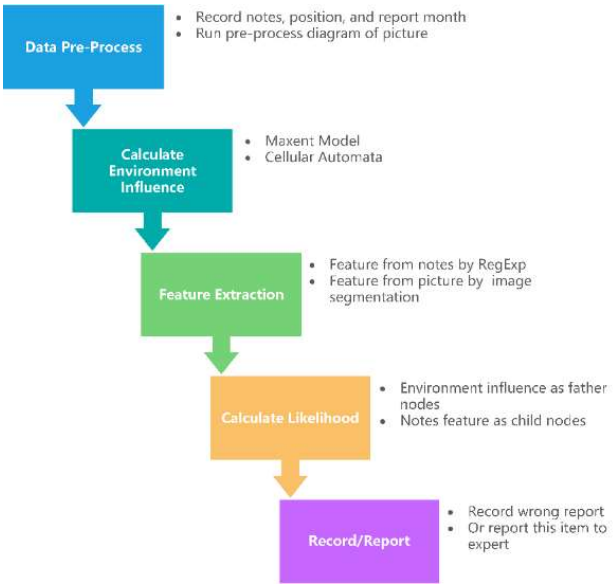

Figure 11 Full workflow

We apply this model on data we have, some of the likelihoods of a true classification are as follows: Average likelihood of positive case is 0.7407, and for negative case, the number is only 0.3379. The within the class variance of true report is 0.003595, for negative case, it is 0.02324. The variance between classes is 0.011304. This means the model's performance is stable in positive case, and it can separate most of negative reports under 0.5 likelihood. Some of the likelihood results of true report by Bayesian Belief Network (BBN) can be seen at the following chart, which shows the performance of this model is stable under most of cases:

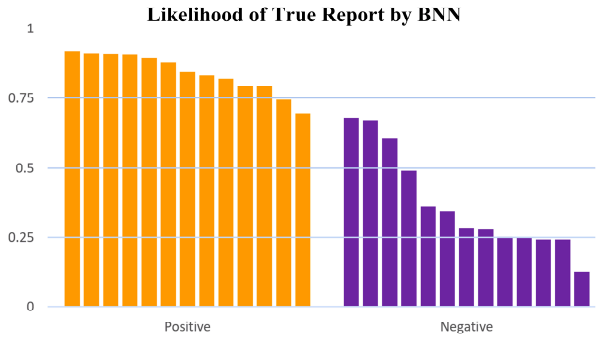

Figure 12

For mistaken likelihood, we only need 1 minus the possibility of true report: $\mathrm{L}(\mathbf{x} \mid \mathbf{e}) = 1 - P(\mathbf{x} \mid \mathbf{e})$.

5.3 Analysis and Solving of Question Three

Using the data and labels we have, we establish the full process of calculate the likelihood of a mistaken classification including file preprocessing, feature extraction and Bayesian Belief Network model. Our motivation is set up a user-friendliness and reliable model to save limited resources of government agencies from meaningless report and give a strategy to expert for ranking new reports. To achieve this goal, we test the model performance of filtering the meaningless report and ranking non-label report of their priority.

First, we apply the model at reports we have, and calculate the likelihood of a mistaken classification, and set the mistaken threshold at 0.5, the results is as the following chart:

Table 3

Class	Threshold	Number	Discard Number	Filter rate	Avg. L
Negative	0.5	2069	1377	66.55%	0.6621
Unverified	0.5	2342	1194	50.98%	0.5144

The average number in class of negative is 0.6621, which means experts shouldn't have read and replied these messages. For a summary, our model will filter 66.55% negative reports automatically, and for the unverified reports, we can filter 50.98% of them.

When new data comes into system, experts are supposed to know the cases' priority, with the distribution of existing data, we set up three kind of priority: high priority, medium priority and low priority according to top 20% and 50% of existing data's likelihood. The likelihood threshold is 0.677 and 0.592. Then we apply our model on the unprocessed report and figured out the priority of these 15 cases. Due to the global ID of report are too long to explain in article, we name them from item one to fifteen order by report time. Though report 1 and 4 are coincide in notes and report 1 doesn't upload image file, so as report 14 and 15, but here we suppose no one of them should be discard because it's a usual case in real system. We first calculate the environment influence part of these 15 reports:

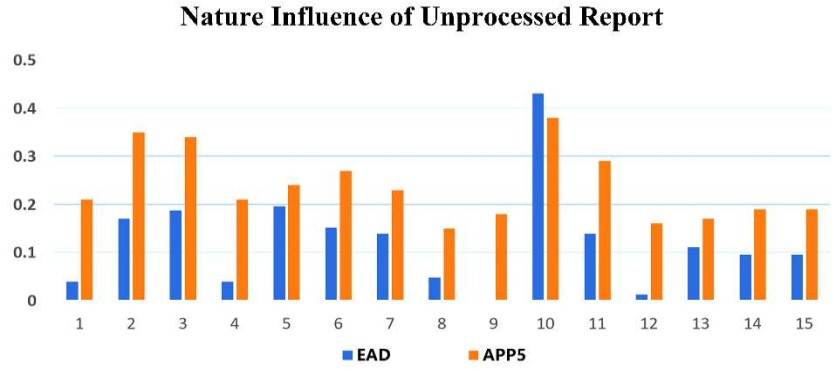

Figure 13

Then with the feature extraction and Bayesian Belief Network, report 5 and 7 have high priority, report 1, 2, 11, 15 have medium priority. Though report 1 and 4; 14 and 15 are describing the same insect, but there child nodes are different due to feature extraction. This phenomenon reflects that our model is flexible for different conditions and sensitive to features of insect.

5.4 Analysis and Solving of Question Four

Bayesian Belief Network is build based on frequency statistics, so the updating strategy is same as chapter 5.2.4:

- Credible feature & possible for Vespa mandarinia: Update frequency statistics with feature resolution as weight.
- Vague feature: Discard this feature in frequency statistics.

Update of maximum entropy model

The maximum entropy model predicts the applicability of the target species in different regions based on the known occurrence of the target species and related environmental variables. Under the premise that the environmental variables in various places will not change significantly in the short term, the update of this model will mainly depend on reliable reports of species distribution, that is, the model only absorbs reports marked as Positive, and considers reports marked as Negative to be useless.

In an ideal situation, to expand the limited number of samples, all new samples marked as Positive should be included in the prediction as much as possible. However, considering that too frequent updates may bring greater workload. We should reasonably evaluate the impact of the new sample points on the original model, that is, judge whether the newly added sample points are outliers under the assumption that the current model is correct

At first, we assumed that the existing model is correct. Here we extract the longitude and latitude of 14 samples, which are recorded as $la_1 \dots la_{14}$ and $lo_1 \dots lo_{14}$. The existing model shows that the regions with high suitability are mainly distributed in the western coastal areas, and the sample covariance coefficient$\mathrm{Cov}(la, lo) = 0.752$. We believe that the longitude and latitude of the sample points that meet the model have a certain correlation. Here we use PCA (Principal Components Analysis) to process the longitude and latitude. And we have:

Table 4

Principal Component	la	lo	Cumulative
F_1	0.7071	0.7071	0.8526

Subsequently, for each newly added point(la_i, lo_i), we calculate the main components of it as well as the existing points, then we use the form of $\mathrm{Y} = \mathrm{X}\beta + \epsilon$ to fit their suitability of the species with their principal components. To estimate the deviation degree of the added point, we use the Cook distance:

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}$$

Where r_i represent the Studentized residual of the $i\,th$ point, $h_{ii} = x_i'(X'X)^{-1}x_i$ denotes the leverage and p {\displaystyle p}is the number of predictors for each observation. (here $p = 2$)

Since Cook's distance is in the metric of an F distribution with p and n-p, if $D_i > F_{0.5}(p, n-p)$, the added point should be regarded as an outlier and we should update the maximum entropy model. Otherwise, we should add it to sample points, and test whether the next point is the outlier.

Update of Cellular Automata

For the updating of cellular automata, **we consider the hypothetical testing**. This method was proposed by K. Pearson. The basic idea of this method is if the result of an independent experiment against the hypothesis that we made then we should reject the original hypothesis otherwise we should accept it. But one significant rule is that we do not easily reject the hypothesis.

Now we look at the cellular automata. We have assumed that the movement of *Vespa mandarinia* follows Gaussian distribution $\mathrm{N}(\mu, \sigma)$ where μ is the position of nest. Then according to the lemma for any positive sighting position $\mathrm{X}_i, i = 1 \ldots n$, the center of X_i denoted by

$$\mathrm{C} = \frac{\sum_{i=1}^{n} X_i}{n}$$

is an unbiased estimate of the nest location μ, and C follows a distribution of $\mathrm{N}(\mu, \frac{\sigma}{\sqrt{n}})$.

Set **significant level be 0.05**, by numerical integration we can get a series possibility within a circle with radius r.

Table 5 Possibility within a circle with different radius

Radius/km	**6.3**	**6.8**	**7.3**	**7.8**	**8.3**
Possibility	0.8916	0.9248	0.9492	0.9667	0.9787

When the circle is centered at the nest with radius of 7.3 km, the distribution probability can reach 0.9492. Now we set the threshold t = 7.3 km.

It shows that if there is a series of samples of positive sighting $\mathrm{X}_i = \{\mathrm{X}_1, \ldots, \mathrm{X}_m\}$ near a possible nest such that the center of X_i is more than 7.3 km away from the nearest possible nest We have good reason to think the current model is not capable to explain this situation. That is, we should update our model.

The steps are as follows:

Step 1. Calculate the distance $\mathbf{d}$ between the X_i and N.

Step 2. Find the nearest possible nest N of the newest positive sighting n.

Step 3. Calculate the center of N's neighbor positive sightings $\mathrm{Neighbor} = \{\mathrm{n}_1, \mathrm{n}_2, \ldots, \mathrm{n}_m\}$ and n. The center C is defined by

$$\mathrm{C} = \frac{n_1 + \cdots + n_m + n}{m + 1}$$

Step 4. Calculate the distance between the C and the N denote as dist

Step 5. If dist is bigger than the threshold t, t=7.3 km. turn to Step 6, otherwise turn to Step 7

Step 6. Repeat the process in problem 1, determine new position of nest.

Step 7. End

Besides if there is no new report for a whole year, we should update our model since the nest could probably move to new locations.

- **Sensitive analysis**

Now we proposed relative update times $\mathrm{r} = \frac{\mathrm{N}_{new}}{N_{up}}$ as a measure of update complexity where N_{up}

is the number of update times, N_{new} is the total new positive sighting report. And bigger the r is, more effective the model is.

We use our model in problem 2 to generate some new samples from the unverified detections and feed our model data in a random order in different frequency, 10, 15, 20, 25. Then we documented the relative update time r after a certain period $times = 10$, we take the average as the result. The result is as follows

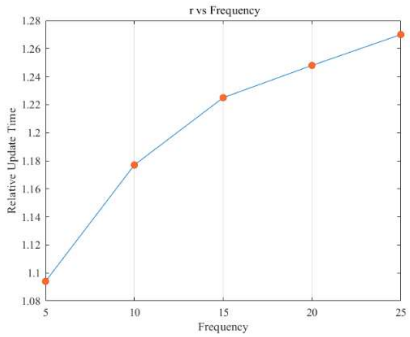

Figure 14.

It seems like with the report frequency increase, the efficiency of our model would go up with the updating.

And for a certain frequency, the r changes little after a certain time tends to be steady.

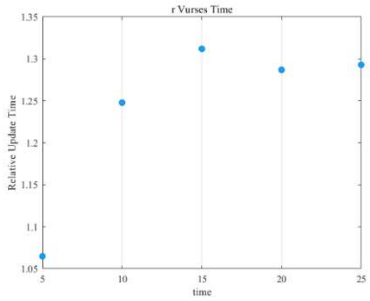

Figure 15.

Which shows after an enough period the efficiency of our model become steady.

5.5 Analysis and Solving of Question Five

Now we are going to talk about how we can make sure that this insect is eradicated. This key of solving this problem is to figure out the probability of *Vespa mandarinia*'s distribution in various regions under the premise of increased mortality. Then by comparing with the model of natural growth model, we can estimate the total group number of *Vespa mandarinia*. **If the model predicts the number of *Vespa mandarinia* is at a very low level, and there is no new report for a whole year. We can believe that this pest probably meets its end.**

Use cellular automata under high mortality. For the first half:

Now we also need to also need to apply the two rules for model in problem 1.The cellular movement possibility is $P = k1 \times k2 \times E^{k3}$. Where E is the comprehensive environmental suitability of C1, which is specifically determined by the suitability of the environment for the survival of Asian

hornet. The values of k1 k2 k3 are the same as in Problem 1

For the second half:

Due to the sudden increase in mortality,

1. If the cell C1 was originally 1, C1 has a probability of p2 to become 0 after iteration, and p2 represents the intensity of manual hunting.
2. In the experiment, p2=0.1~0.9, with an interval of 0.1

The result of algorithm is as follows:

Table 6: The number of counted cellular

Experiment Round	1	2	3	4
Totally Counted Cell	11	13	14	9

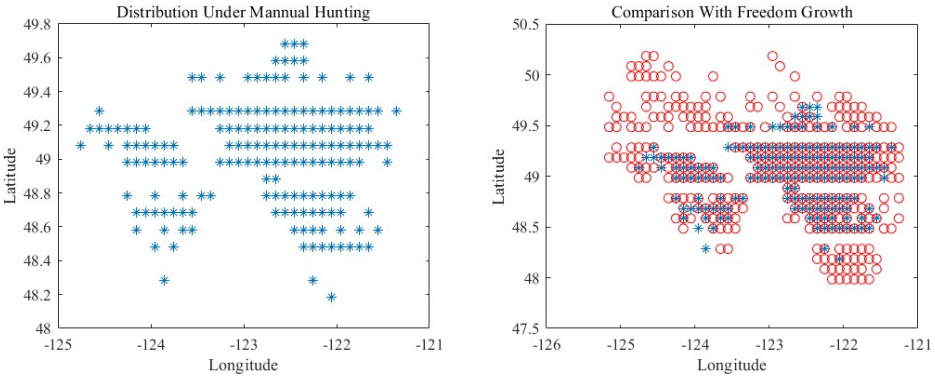

Figure 16:Comparison of manual hunting and freedom growth

From this model we can easily understand the total group number by detecting the number of reports in each area and compare it with the freedom growth model. In our model, the estimate number of the current Vespa mandarinia's number is now 75% of its heyday under the death rate of 30%.

VI. Evaluation and Promotion of Model

6.1 Strength and Weakness

6.1.1 Strength

1. The stimulation process is based on the living habits and the natural growth model to predict the distribution of Vespa mandarinia, that is we don't need much prior information about location.
2. We stimulate the movement of Vespa mandarinia which makes our conclusion more reasonable, and our model more explanatory.

6.1.2 Weakness:

1. Accuracy will become lower in large search range.
2. Picture processing is complex and time-consuming.

VII. Conclusions

7.1 Conclusions of the problem

- Based on the prediction of the Maxent niche model, the coast of Vancouver Island and the northern part of Washington State are suitable for Vespa mandarinia to survive.
- Prediction of spread direction of *Vespa mandarinia.*
  According to the conclusion of our model, the colony will have a high probability of spreading northeast or northwest. The forecast map shows that there are currently a large number of areas in Washington that may be invaded. Canada may also be affected by the *Vespa mandarinia* Some group of *Vespa mandarinia* may choose to stay in the U.S. and move to east.
- Estimate conditional probability of true report based on Bayesian Belief Network.

7.2 Methods used in our models

- The Maxent niche model
- Cellular Automata
- Hypothesis testing
- Bayesian Belief Network

VIII. References

[1] Phillips, S. J., Dudík, M. & Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling.. In C. E. Brodley (ed.), ICML, : ACM.

[2] WorldClim. worldclim21. https://www.worldclim.org/data/worldclim21.html Accessed 2/8/2021.

[3] Rother, C., Kolmogorov, V. & Blake, A. (2004). "GrabCut": interactive foreground extraction using iterated graph cuts.. *ACM Trans. Graph.*, 23, 309-314.

[4] F. Yi and I. Moon, "Image segmentation: A survey of graph-cut methods," 2012 International Conference on Systems and Informatics (ICSAI2012), Yantai, 2012, pp. 1936-1941, doi: 10.1109/ICSAI.2012.6223428.

[5] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann. ISBN: 1558604790

[6] Kojima, K., Perrier, E., Imoto, S. & Miyano, S. (2010). Optimal Search on Clustered Structural Constraint for Learning Bayesian Network Structure.. *J. Mach. Learn. Res.*, 11, 285-310.

[7Zhang, N. L. (2004). Hierarchical Latent Class Models for Cluster Analysis. *J. Mach. Learn. Res.*, 5, 697--723.

[8] Pennsylvania State University Extension. Asian Giant Hornets. https://extension.psu.edu/asian-giant-hornets Accessed 7/2/2021

[9] Duntsch, I. & Gediga, G. (1998). Uncertainty Measures of Rough Set Prediction.. *Artif. Intell.*, 106, 109-137.

[10] VarCon. Version 6 of the 12dicts word lists. http://wordlist.aspell.net/12dicts-readme/ Accessed 7/2/2021.

[11] Nguyen, T. P. H., Cai, Z., Nguyen, K., Keth, S., Shen, N. & Park, M. (2020). Pre-processing Image using Brightening, CLAHE and RETINEX.. *CoRR*, abs/2003.10822.

[12] Hatzikirou, H., Breier, G. & Deutsch, A. (2009). Cellular Automaton Modeling.. In R. A. Meyers (ed.), *Encyclopedia of Complexity and Systems Science* (pp. 913-922) . Springer . ISBN: 978-0-387-75888-6.

[13] Tomita, K., Kurokawa, H. & Murata, S. (2009). Graph-Rewriting Automata as a Natural Extension of Cellular Automata. In T. Gross & H. Sayama (ed.), Adaptive Networks: Theory, Models and Applications (pp. 291--309) . Springer .

[14] Griffeath, D. (2003). New constructions in cellular automata Oxford University Press Paperback .

[15] W.R.E. HoffmannNeumann, and E. SchmolzP. (2000). Technique for rearing the European hornet (Vespa crabro) through an entire colony life cycle in captivity. Insectes Sociaux, 351-353.