Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2104744

Goodbye Buzz: A Report Confidence Ranking Model

Summary

In September 2019, the Asian giant hornet was discovered on Vancouver Island, British Columbia, Canada. Since that time, sightings have started to occur in Washington State as well.In response to the conflict between limited investigative resources and the huge number of unverified reports, we develop a mathematical model for ranking reports based on their confidence level to assist government agencies in making policy.

We develop the Pairwise Long-Distance Dispersal Model to predict the spread of the Asian giant hornet by K-Means clustering to divide the Asian giant hornet reporting locations in Washington State into different regions. We use the nonlinear method of least squares to optimize the parameters and logistic equation to simulate population growth to solve for the existence probability of the Asian giant hornet in each region to obtain the final spread model.The MSE of this model is 6.48074.

Confronted with a significant number of unverified reports and limited investigative resources, we must prioritize all reports in terms of confidence to develop a reasonable strategy for allocating investigative resources.Combining with existing population spread prediction models, we introduced TextRank-based text confidence and VGG-16-based image confidence. The weights of the variables are determined using the Fuzzy Analytic Hierarchy Process. Finally, the confidence prioritizing was performed using TOPSIS to obtain the probability of all unverified reports of being positive.

As time passes, our model would also be updated. After considering all factors, we choose to update the model every six months. Each update is able to update the model parameters, the keyword corpus and the VGG-16 network, to improve the accuracy and the robustness of the model. Particularly, we use our own constructed feature dictionary to further classify the negative reports, so that the accuracy of image detection has been effectively improved. Also when the new report reaches a certain scale, we will conduct regional redivide using semi-supervised K-means.In addition, we will also update our model if government departments take certain precautionary measures.In addition, we will also update our model if government agency take corresponding measures.

We define metric to reflect the eradication of the Asian giant hornet based on the Reporting Confidence Ranking Model.Government departments can use this metric for risk assessment and policy development.

Keywords: Semi-Supervised K-means; Logistic Equation; TextRank; VGG-16 Network

Contents

1 Introduction

1.1 Background

The Asian giant hornet (Vespa mandarinia) was detected in western British Columbia, Canada and Washington State, United States. V. mandarinia are an invasion species due to their ability to kill local honey bees and affect humans with their toxicity.

According to the relevant literature [8], the Asian giant hornet would be partial to survive in areas with warm to cool averaged annual temperatures, high average precipitation, and high human activity. Compared to native hornets, Asian giant hornet has a small ecological niche with environmental adaptability and dispersal potential. Therefore, without human intervention, it has a high tendency to spread to all suitable environments throughout the United States and to invade wild or captive colonies, killing or expelling native bees. In Europe, the invasion of Asian giant hornet once caused 18%-50% loss to beekeepers [4]. Accordingly, if the government and related agency do not take appropriate surveillance and management measures in a prompt manner, the economic and ecological environment of the United States will face critical challenges.

The existence of the Asian giant hornet has caused anxiety for a lot of individuals in Washington State. As a result, Washington State has established a helpline and a website to gather information and answer questions. In the face of numerous eyewitness reports, some of which have been identified, there are still a large number of unverified eyewitness reports. Therefore, the state must adopt a prioritization strategy to deploy its limited resources for follow-up investigations

1.2 Problem Summary

In this problem, our team is given data about images sent by witnesses, notes, date, locations, as well as identification results and comments. Our primary work is to address "how to interpret the data" and "how to develop a strategy for priority assignment with the limited resources".

To achieve our goals, specifically, we need to:

- Based on the available data and related information, develop a spread model of the Asian giant hornet.
- Combining images and text data, build a mathematical model of the likelihood that reports sent by the population is incorrectly identified.
- Use the above model to rank the likelihood of reports to develop a priority allocation strategy for limited resources.

- Determine the update frequency of the model and construct metric that reflect the eradication of the pest.

1.3 Data Cleaning

After an initial overview of the data, we find that there are a lot of issues with the given data. Therefore, we perform data cleaning to facilitate the subsequent work.

Although most of the report attachments are in the form of images, there are still data in the form of PDF, Word, video, and zip files. For PDF and Word forms, we extracted the images from them. For the video, we will intercept frame by frame and manually select the clearest and most representative one. For zip file, we will extract the images from them by decompression.

In addition, we also find that there is a small amount of unreasonable time for the report to be submitted. Therefore, we clean out the data other than 2019 and 2020.

1.4 Our Model

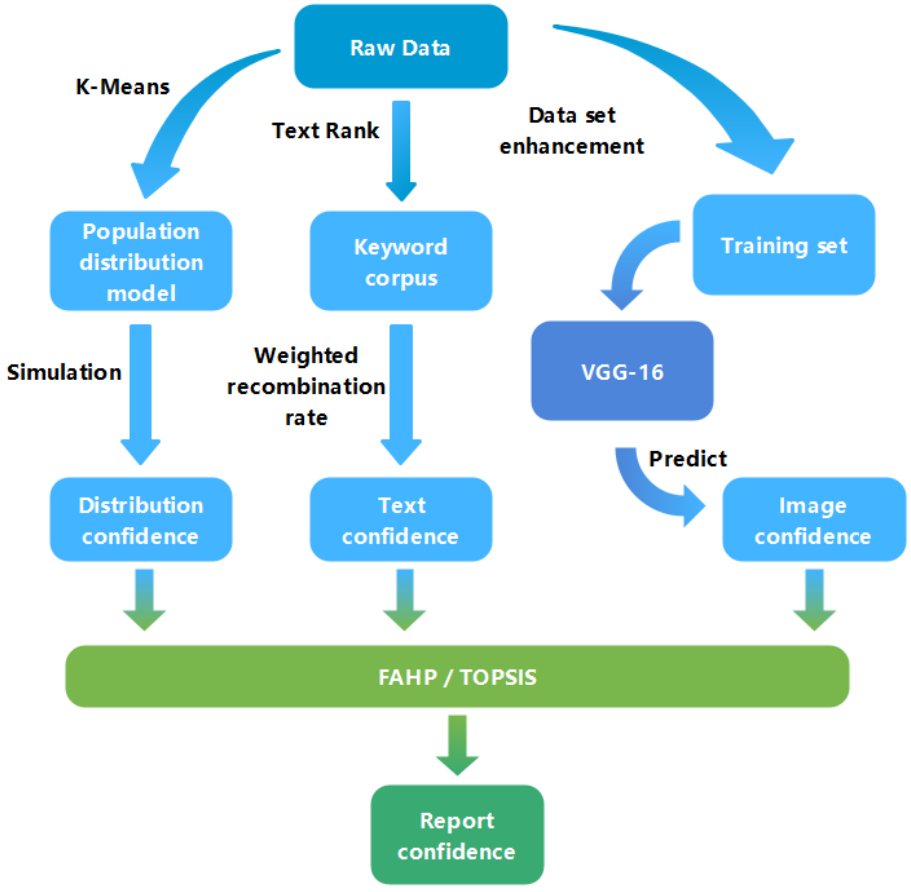

Figure 1: Model overview

In response to our goal to rank all unverified reports. We propose to construct three metrics that reflect the confidence of the reports, namely distribution confidence, text confidence and image

confidence.

For distribution confidence, we build a Pairwise Long-Distance Spread Model. We use a semi-supervised K-Means algorithm to divide Washington State in such a way that each area is as regular as possible and fits the population distribution.

We divide the cycle of the Asian giant hornet into a reproductive stage and a spread stage.

For the reproductive stage, we choose to fit the reproduction of the Asian giant hornet using the logistic equation, which approximates the reproductive potential of the population by the resource density of the region.

For the spread stage, for each region we simulate the probability of the existence of the Asian giant hornet and use a nonlinear method of least squares for parameter optimization.

For text confidence, we use the TextRank algorithm to filter the keyword corpus. The weighted Overlap rate of the report note and the keyword corpus is used as the text confidence.

For image confidence, we conduct image prediction using VGG-16 network and regard the prediction result as image confidence.

Finally, we assign weights to these three confidence levels with FAHP and use TOPSIS for the final ranking of the confidence levels.

2 Assumptions and Notations

2.1 Assumptions

To simplify our model, we make the following assumptions:

***Asm*.1** There is a positive correlation between the number of reports submitted and the population density and intensity of bees in the submitter's location.

***Asm*.2** The Asian giant hornet spread naturally in Washington State and is not affected by human activities.

***Asm*.3** The Asian giant hornet breeds and migrates according to its habitat, and the environment in Washington State is normal, with no serious anomalies.

***Asm*.4** No consideration of climate, environmental factors during the migration stage.The probability of the Asian giant hornet migrating from region to another region is only related to the distance.

2.2 Notations

Table 1: Notation Table

Symbol	Description
SSE	the sum of the squared error
C	the set of divided region
K	the number of divided region
v	the distance between two points
P	the possibility of existing AGH in region
$loss$	Loss function of nonlinear least squares method
T	the text base
S	the set of sentences of a text base
TC	Matching rate of the text base
E	Probability of complete pest eradication in Washington State
Pr	the confidence of report

3 Pairwise Long-Distance Dispersal Model

In this section, we develop the Pairwise Long-Distance Dispersal Model, which divides the Asian giant hornet submitter's location in Washington State into different regions by K-Means clustering. Then, drawing on the methods in the existing literature [3], we simulate the probability of the existence of the Asian giant hornet in each region, and use nonlinear method of least squares for parameter optimization to obtain the final spread model.

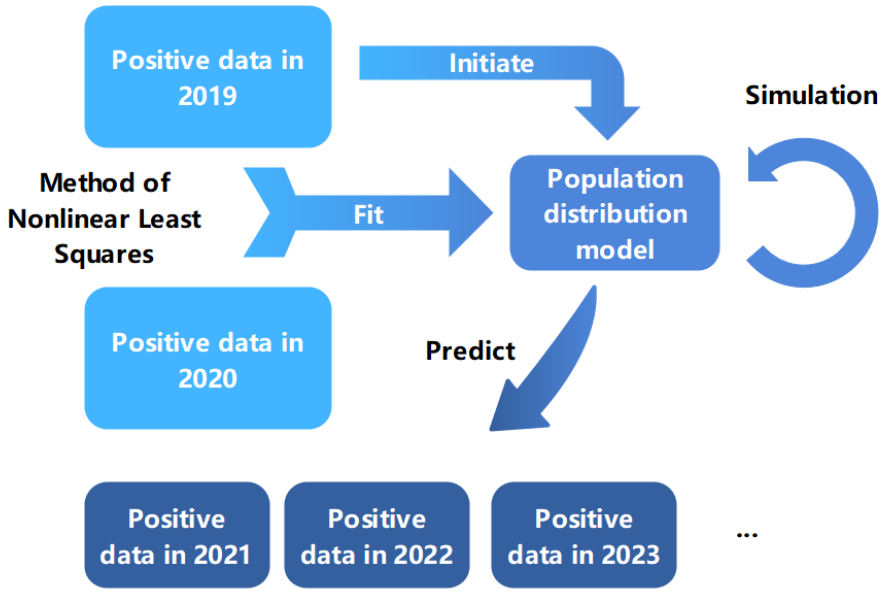

Figure 2: Model outlook

3.1 Improved K-Means clustering

Clustering belongs to unsupervised learning. K-Means algorithm is a very typical distance-based clustering algorithm that uses distance as an evaluation index of similarity, i.e. it is considered that the closer two objects are, the greater their similarity. The algorithm regards clusters as consisting of objects that are close in distance, and consequently takes getting compact and independent clusters as the ultimate target.

Consider data on Euclidean distances, using Sum of the Squared Error (SSE) as the objective function for clustering.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2 \tag{1}$$

Where K denotes a total of K clustering centers, c_i denotes the center i, $dist$ denotes the Euclidean distance,and C_i denotes the cluster of the center i. To minimize the SSE, we derive the SSE such that the derivative is equal to 0 and solve for c_k, as follows:

$$\begin{aligned} \frac{\partial}{\partial c_k} SSE &= \frac{\partial}{\partial c_k} \sum_{i=1}^{K} \sum_{x \in C_i} dist(c_i, x)^2 \\ &= \sum_{i=1}^{K} \sum_{x \in C_i} \frac{\partial}{\partial c_k} (c_i - x)^2 \\ &= \sum_{x \in C_k} 2(c_k - x_k) = 0 \end{aligned} \tag{2}$$

$$\sum_{x \in C_k} 2(c_k - x_k) = 0 \Rightarrow m_k c_k = \sum_{x \in C_i} \Rightarrow c_k = \frac{1}{m_k} \sum_{x \in C_i} x_i \tag{3}$$

Therefore, the optimal center of mass for the minimized SSE of the cluster is the mean value of each point in the cluster.

The pseudo-code of the algorithm is shown table[1].

For the value of parameter k, we decided to introduce a parameter σ to choose. Because if the value of k is not appropriate, it may make the divided region irregular with unreasonable size which is not conducive to the subsequent calculation.

The defination of σ is below:

$$v_i = \sqrt{(x_i - x_c)^2} \tag{4}$$

$$\sigma = \sqrt{\frac{\sum (v_i - \overline{v})^2}{n}} + (\overline{v} - r)^2 \tag{5}$$

where x is the coordinates of the latitude and longitude points, x_c is the coordinates of the center point, n is the number of points, and r denotes the activity radius of the Asian giant hornet.

We make the k-value iterate until σ is smaller than customized hyperparameter δ, at which point the k-value is the optimal solution.

Algorithm 1 Framework of K-Means Algorithm

Input: Latitude and longitude coordinates data of possible discovery sites of AGH D, the number of clusters after clustering k

Output: the data D divided into k regions

```
Create a set C of k clusters randomly
while C change do
    for all x ∈ D do
        for all c_k ∈ C do
            d_ik = dist(x_i, c_k)
            x_i ∈ C_k(k|d_ik = Min(d_ik))
        end for
    end for
    for all c_k ∈ C do
        c_k = E(C_k)
    end for
end while
```

The optimal solution for k is 12, and Figure [3] shows the results of dividing the region of Washington State using semi-supervised K-means.

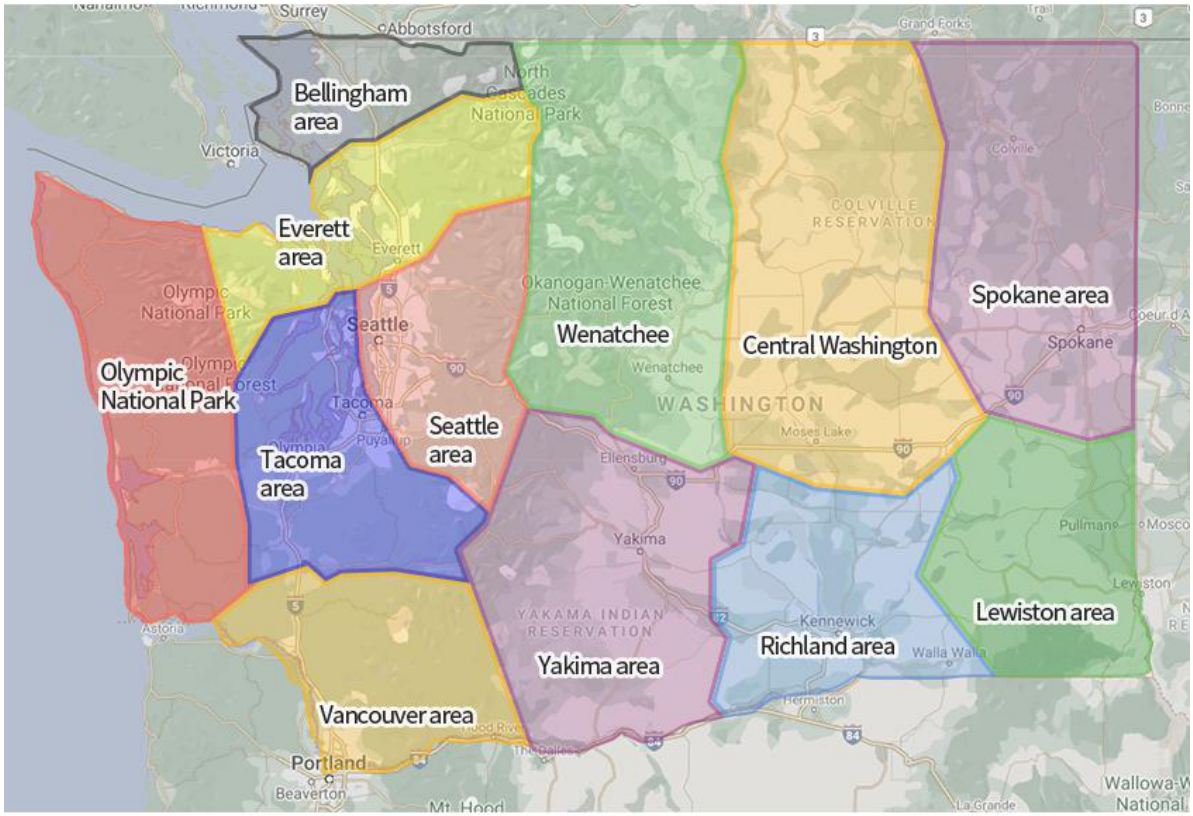

Figure 3: Divided Region Result

3.2 Migration Stage

Determining the range of spread of invasive species is difficult. First, we are unable to determine the specific details of the spread of the Asian giant hornet in different environments. Second, it is impossible to determine the true aggregation of populations simply by the location of positive reports.

Our model takes no account of climatic, environmental factors in the migration stage, but only the distance between the two regions.The pairwise distances between all locations are calculated using the Haversine formula [7].

$$d_{ij} = arccos(sin(lat_i) * sin(lat_j) + cos(lat_i) * cos(lat_j) * cos(lon_i - lon_j)) \tag{6}$$

where lat_i and lon_i are latitude and longitude of location i, and lat_j and lon_j are latitude and longitude of location j.

According to the literature [1], three distance models have been constructed by existing studies.E-model with the negative exponential kernal was applyed for the spread model. N-model with the normal kernal always applyed for description of stochastic. The C-model with the fat-tailed distribution is often used for long-distance desperal model. Our model is going to apply the fat-tailed distribution.

We start the initialization of P by initializing P to 1 for regions with positive reports, and to 0 otherwise.

Since we assume that the geographic range of our study is very large and the spatial resolution exceeds the species dispersal distance of biological means, we further make the simplifying assumption that P_{ij} values are independent of the likelihood of reaching adjacent sites.

$$P_{ij} = \frac{1}{1 + (\frac{d_{ij}}{\gamma})^2} \tag{7}$$

P_{ij} ranges from 0 to 1. γ is a tunable parameter.

The probablity that the Asian giant hornet would be not introduced from the region i to region j this year is $1 - P_{ij}$. The probablity that the Asian giant hornet would be not introduced to the region j from any other region is production of likelihoods of these independent events:$\Pi_i(1 - P_{ij}$.So the probability of the Asian giant hornet spreading from region i to region j is calculated as:

$$P_j = 1 - \Pi_i(1 - P_{ij}) \tag{8}$$

In this way, the distribution of pests in one year can be used to predict the distribution of pests in the next year. Also, the data of the next year can be taken to update the model parameters γ using nonlinear method of least squares. When there is a positive report in the region in the current year, P for that region is 1, otherwise it is 0. The parameters γ is selected so that the sum of squares of all differences between the calculated likelihoods and assigned values was minimal.

$$P_j^{'} = \begin{cases} 1, & \text{if region j have positive report ;} \\ 0, & \text{if region j dont't have positive report .} \end{cases} \tag{9}$$

Generally, we can assume that multiple measurements are independently and identically distributed among themselves, i.e., that one measurement does not affect the results of other measurements. On the other hand, assuming that the model we design does describe the real physical event well, the error of any one measurement should be a random error, which may be considered to satisfy the following Gaussian distribution

$$g_\sigma(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon^2}{2\sigma^2}} \tag{10}$$

We all know that the Gaussian distribution has the largest probability of the position of the axis of symmetry. It so happens that the axis of symmetry of $g_\sigma(\epsilon_j)$ is at the origin, and the direction of optimization we expect is exactly toward the origin. Thus, minimizing the residuals is equivalent to maximizing the joint probability of all measurements. Since the measurements are independently and identically distributed, it is straightforward to multiply $g_\sigma(\epsilon_j)$ for each measurement to obtain:

$$p(P^{'}, P, \sigma) = \Pi g_\sigma(P_i^{'} - P_i) \tag{11}$$

Combining the two equations, we can obtain:

$$p(P^{'}, P, \sigma) = (2\pi\sigma^2)^{-\frac{k}{2}} e^{-\frac{1}{2}\sum\limits_{i=1}^{k}\frac{(P_i^{'}-P_i)^2}{\sigma^2}} \tag{12}$$

Since k and σ are certain values, minimizing $p(P^{'}, P, \sigma)$ is equal to minimizing $\frac{1}{2}\sum\limits_{i=1}^{k}(P_i^{'} - P_i)^2$ We thus define the loss function as:

$$loss = \frac{1}{2}\sum_{i=1}^{k}(P_i^{'} - P_i)^2 \tag{13}$$

We re-write the loss function in the following form:

$$loss = \sum_{i=1}^{k} r_i^2(P) \tag{14}$$

where r_i is the residual i.Under this definition, the first and second order derivatives of the objective function can be expressed in terms of the Jacobi and Hessian matrices of the residuals.

$$\nabla loss = \sum_{i=1}^{k} r_i(P)\nabla r_i(P) = J(P)^T r(P) \tag{15}$$

$$\begin{aligned} \nabla^2 loss &= \sum_{i=1}^{k} \nabla r_i(P)\nabla r_i(P)^T + \sum_{i=1}^{k} r_i(P)\nabla^2 r_i(P)^T \\ &= J(P)^T J(P) + \sum_{i=1}^{k} r_i(P)\nabla^2 r_i(P) \end{aligned} \tag{16}$$

where $J(P)$ is the Jacobi matrix of the residuals and $\nabla^2 r(P)$ is the Hessian matrix of the residuals.

We use the Levenberg-Marquardt method to obtain the linear equation

$$(J^T + \lambda I)p = -J^T r \tag{17}$$

Finally, by solving the linear equation, we obtain the result that the loss function is minimized when γ takes the value of 3.9682e-05, which is 6.48074.

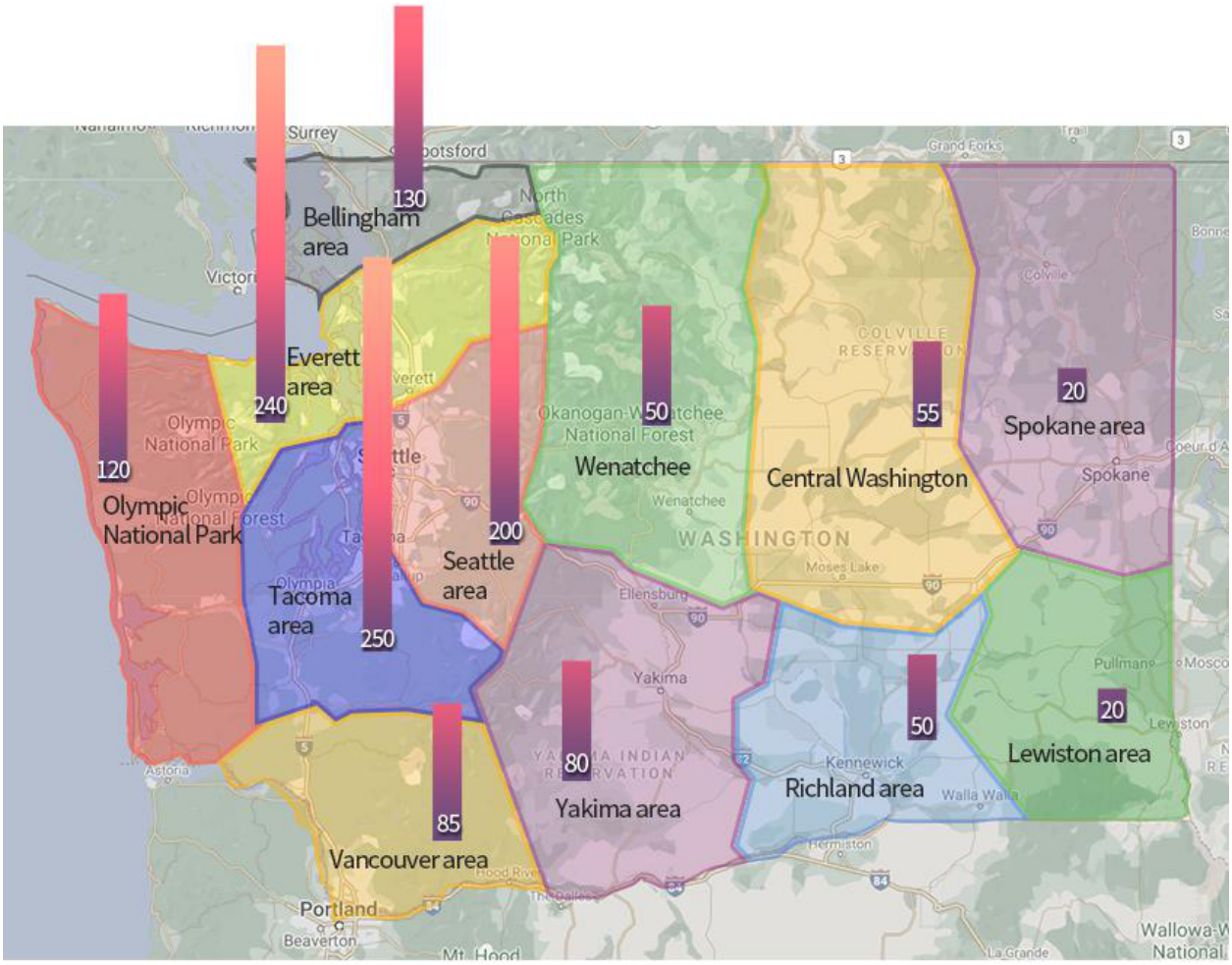

Figure 4: Model Result

We simulated the distribution of the Asian giant hornet in 2026 after a time interval of 5 years, and Figure[4] shows the distribution of the Asian giant hornet in 2026. This indicates that the Asian giant hornet will invade the entire state of Washington if government agencies do not take action.

3.3 Reproduction Stage

By viewing the relevant information [8] we learn that the Asian giant hornet prefers to live in areas with warm to cool average annual temperatures, high precipitation and active human activities.

The population peaks in August and shifts to hunting other bees in September and October. Males and females will leave the hive to mate in October and November. The entire population will die out in the winter, except for the queen [6].

For the reproductive stage of the population, we use the logistic equation to describe the population growth curve of the Asian giant hornet.

$$\frac{dx}{dt} = kx(a - x) \tag{18}$$

After solving this equation, the general solution is obtained as:

$$x(t) = \frac{a}{1 + Be^{-akt}} \tag{19}$$

Where $A = e^{ac}$,a denotes the environmental maximum capacity of the Asian giant hornet,and $B = \frac{1}{A} = e^{-x}$, k denotes the population potential coefficient of the Asian giant hornet,t denotes the time the population has been reproducing.

Based on assumption 5, we substitute the submitted frequency of the region for the population potential coefficient of the region.

$$k = \frac{n}{S} \tag{20}$$

n denotes the number of reports in the region in a year and S denotes the area of the region.

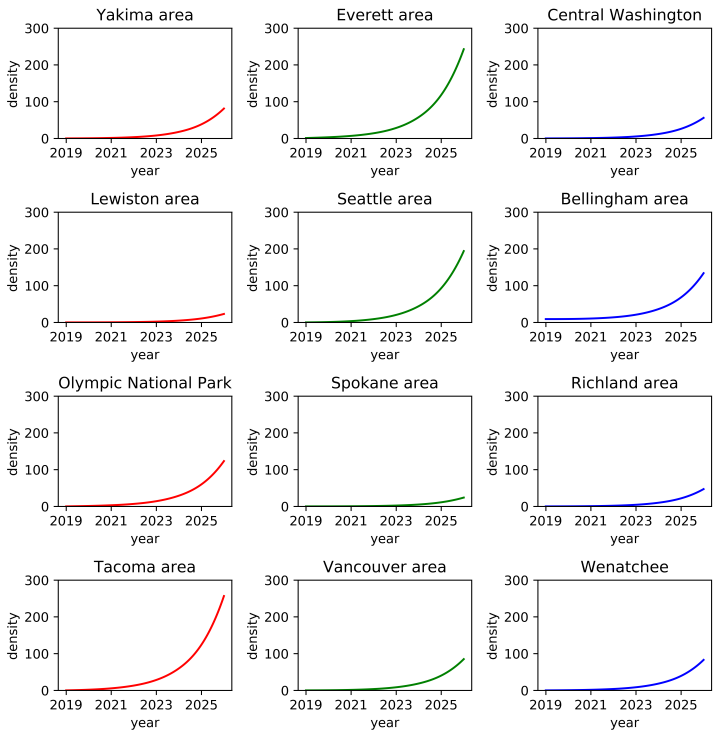

Figure 5: Model Result

We conducte simulations with a time horizon of 5 years and the results are shown in Figure[5]

4 Report Confidence Ranking Model

Faced with a massive number of unverified reports and limited investigative resources, we must rank all reports in terms of confidence to develop a reasonable prioritized resource allocation strategy. Combining with existing population spread prediction models, we introduce text confidence based on TextRank[2] Algorithm and image confidence based on VGG-16 Network.The weights of the variables are determined using the Fuzzy Analytic Hierarchy Process. Finally, confidence ranking is performed using TOPSIS.

4.0.1 Text Confidence

The TextRank Algorithm is a graph-based ranking algorithm for keyword extraction and document summarization, improved from Google's PageRank algorithm for ranking the importance of web pages, which uses co-occurrence information (semantics) between words within a document to extract keywords, which can extract keywords and key phrases from a given text, and extract key sentences from the text using an extractive automatic digest method.

We segment the complete sentence of the lab comment text T.

$$T = [S_1, S_2, ..., S_n]$$

For each sentence S_i belongs to T, the word division and lexical annotation are processed, and deactivated words are filtered out, and only words of the specified lexical nature, such as nouns, verbs and adjectives, are retained.

$$S_i = [t_{i,1}, t_{i,2}, ..., t_{i,n}]$$

where $t_{i,j}$ is the candidate keyword after retention. Construct the candidate keyword graph $G = (V, E)$, where V is the set of nodes, consisting of the generated candidate keywords, and then construct the edges between any two points using the co-occurrence relation, where edges exist between two nodes only when their corresponding words co-occur in a window of length K, with K denoting the window size.

$$WS(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j) w_{jk}} w_{jk}} WS(V_j) \tag{21}$$

According to the above formula, the weights of each node are propagated iteratively until convergence. The node weights are sorted in reverse order so as to obtain the most important m words, which are tagged as candidate keywords in the original text, and combined into multi-word keywords if adjacent phrases are formed.If a number of extracted keywords are adjacent to each other in the text, then they constitute a key phrase being extracted.

With the TextRank algorithm, we get two keyword bases, the positive keyword base $T_{positive}$ and the invisible keyword base $T_{negative}$.Their corresponding weights are $W_{positive}$ and $W_{negative}$ respectively.

Based on the two keyword libraries, we define two text confidence metrics:

$$TC^{positive} = \frac{\sum\limits_{w_i \in T} w_i}{\sum\limits_{w_j \in T_{positive}} w_j} \tag{22}$$

$$TC^{negative} = \frac{\sum\limits_{w_i \in T} w_i}{\sum\limits_{w_j \in T_{negative}} w_j} \tag{23}$$

Where T represents the keyword mentioned in the submitter's accompanying note, w represents the weight of the keyword, and TC ranges $[0, 1]$.

Table 2: Keyword Table

Word(negative)	Weight(negative)	Word(positive)	Weight(positive)
wasp	0.3035	specimen	0.3466
sawflies	0.2724	specimens	0.3466
colored	0.2724	wsda	0.2684
sawfly cimbex	0.1969	submitting	0.2286
yellowjacket	0.1920	usda	0.1466
...	...	...	...
stingers	0.0292	confirm	0.1885

4.0.2 Image Confidence

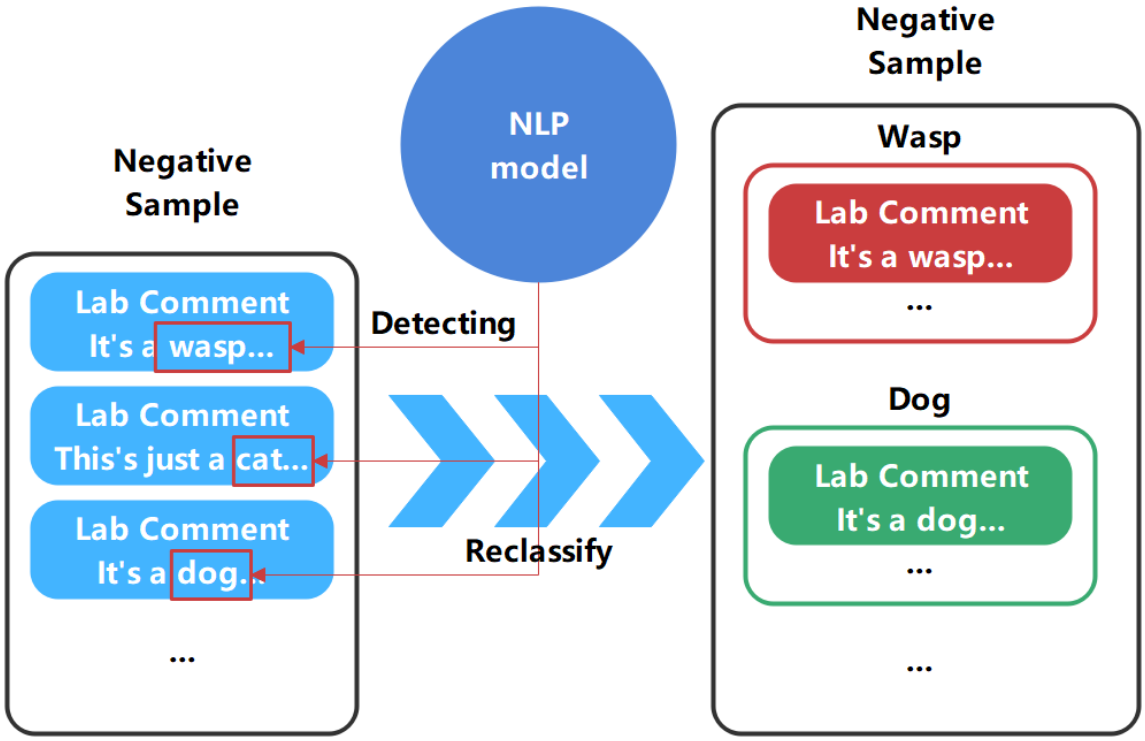

Figure 6: Negative Report Reclassification Model

The images accompanying the report also bear important information. The laboratory staff also identified the Asian giant hornet mainly through the images, so we cannot ignore the confidence level of the images. We plan to start by classifying the images identified as negative by keyword extraction technique, and the classification results are as follows

Table 3: Keyword Table

Category	Number	Category	Number
wasp	832	hornet	212
bee	319	cicada	57
yellojacket	72	horntail	29

Data imbalance problem caused by too few images provided in the positive report and reports of scarce species. We decided to increase the data by data augmentation, taking methods such as flipping, rotating, scaling, cropping, and adding Gaussian noise to increase the number of images with small sample size for training.

The results are shown in the figure[4]:

Table 4: Keyword Table

Category	Number	Category	Number
wasp	416	hornet	212
bee	319	cicada	228
yellojacket	288	horntail	290
positive	240		

VGG is a convolutional neural network model proposed by Simonyan and Zisserman [5], whose name comes from the acronym of the Visual Geometry Group at the University of Oxford, where the authors work.

The model participated in the 2014 ImageNet image classification and localization challenge and achieved excellent results: it ranked second in the classification task and first in the localization task.

The resolution of all images is 450×600. But the input requirement of VGG network is 224×224 RGB image, so we convert the resolution of all images to 224×224.

Table 5: VGG-16 Framework

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weeights layers	11 weeights layers	13 weeights layers	16 weeights layers	16 weeights layers	19 weeights layers
input(224 × 224 RGB image)					
conv3-64	con3-64 LRN	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	con3-128 LRN	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 conv1-256	conv3-256 conv3-256 conv3-256	conv3-256 conv3-256 conv3-256 conv3-256
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512	conv3-512 conv3-512 conv3-512 conv3-512
maxpool					
FC-4096					
FC-4096					
FC-1000					
softmaxl					

In the table5, each column corresponds to a structural configuration.

the table5 shows that VGG-16 contains 13 Convolutional Layers, 3 Fully connected Layers, and 5 Pool layers.

The convolutional layer can extract the features of the image and the weights of the convolutional kernel are learnable, thus it can be guessed that the convolutional operation can break the limitation of the traditional filter in the high-level neural network and extract the desired features according to the objective function. Formula for matrix convolution.

$$(f * g)(1,1) = \sum_{k=0}^{2} \sum_{h=0}^{2} f(h,k) g(1-h, 1-k) \tag{24}$$

The main role of the pooling layer is to reduce the matrix size and parameters, which serves to speed up the computation while preventing overfitting. In the whole convolutional neural network, the convolutional layer and pooling layer are for input feature extraction, then the fully connected layer acts as a "classifier".

The use of relu as the activation function in VGG-16 is mathematically represented as

$$relu = max(0, x) \tag{25}$$

Our loss function uses a cross-entropy loss function:

$$L = \frac{1}{N}\sum_{i} L_i = \frac{1}{N}\sum_{i} -\sum_{c=1}^{M} y_{ic} log(p_{ic}) \tag{26}$$

where M is the number of categories, y_{ic} indicates the variable (0 or 1),which is 1 if the category is the same as the category of sample i, and 0 otherwise; p_{ic} is the predicted probability for the observation that sample i belongs to category c.

The softmax regression converts the final output into a normalized probability distribution. We simply feed the state of the upper layer into a linear layer to make predictions on the input image.

$$softmax(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} ez_k} \tag{27}$$

We divide the training set, validation set and test set in the ratio of 6:2:2. We conducted 20 rounds of training with a learning rate of 0.001 and a batch size of 32.

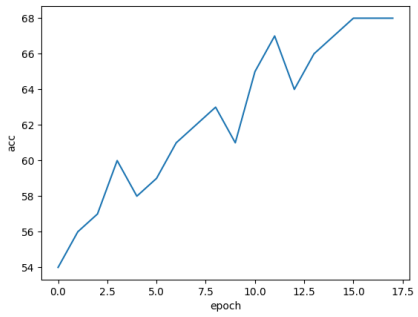

Figure 7: Accuracy Curve

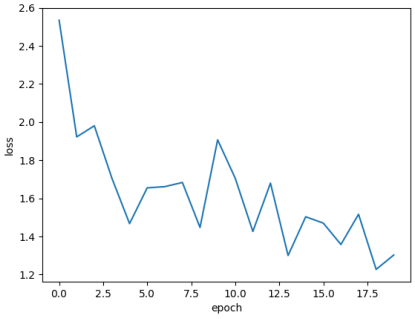

Figure 8: Loss Function Curve

The performance of our model on the test set is 68% accuracy. Because of the low prediction accuracy of VGG-16, we do not intend to use it as the main variable in the model.

4.0.3 Report Ranking

Fuzzy Analytic Hierarchy Process is a quantitative analysis method that combines Analytic Hierarchy Process with fuzzy comprehensive evaluation. In this paper, FAHP method is used to determine the weights, and its main algorithm steps are as follows. The data for the importance matrix judgment in the Fuzzy Analytic Hierarchy Process is obtained from the expert scoring, and the specific method is as follows.

Respondents are asked to score two-by-two comparisons according to the important factors affecting consumers' shopping decisions, and each pair of attribute comparison items was scaled by 0.1-0.9, and users are asked to fill out the importance relationship matrix by comparing the importance of different index factors. The importance matrix is filled out by comparing the importance of different index factors. The importance relationship is expressed as an examination function f (x, y), which represents the importance scale of factor x and factor y for the overall, and the importance matrix for f (x, y) is constructed by using the list comparison method. using a list comparison method to construct the priority relation matrix, and the scales are described as shown in Table.

Table 6: Inportance Comparison

defination	description	measure
Equally important	x and y are equally important	0.5
Slightly important	x and y are slightly important	0.6
Obviously important	x and y are obviously important	0.7
Special important	x and y are special important	0.8
Extreme important	x and y are extreme important	0.9

Based on the above definition, we construct the prioritized relationship matrix C using distribution confidence P, picture confidence I, text confidence $TC^{positive}$, $TC^{negative}$, and distribution confidence P.

where:

$$C = \begin{matrix} C_{11} & C_{12} & \cdots & C_{1n} \\ C_{21} & C_{22} & \cdots & C_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{nn} \end{matrix} = \begin{pmatrix} 0.5 & 0.4 & 0.4 & 0.2 \\ 0.6 & 0.5 & 0.5 & 0.3 \\ 0.6 & 0.5 & 0.5 & 0.3 \\ 0.8 & 0.7 & 0.7 & 0.5 \end{pmatrix} \tag{28}$$

Normalize the elements of the determination matrix.

where:

$$B = (b_{ij})_{m \times n} \tag{29}$$

$$b_{ij} = \frac{a_{ij}}{\sum_{i=1}^{n} a_{ij}}, (i, j = 1, 2, ..., n) \tag{30}$$

Then, the elements in matrix B are summed by rows to obtain vector $C = (c_1, c_2, .., c_n)^T$

where:

$$c_{ij} = \sum_{i=1}^{n} b_{ij}, (i, j = 1, 2, ..., n) \tag{31}$$

Finally, we would normalize the vector C to obtain the Eigenvector $W = \{w_i, w_2, ..., w_n\}$

After calculation, our weights are $W = [0.2245, 0.2451, 0.2451, 0.2851]$.

In this paper, the improved TOPSIS judging method is used as the online review usefulness ranking filtering model algorithm. The basic idea is: on the basis of determining the weights of each attribute index, normalizing the original data matrix, calculating the distance between each evaluation object and the optimal solution and the worst solution respectively, and obtaining the relative proximity of each evaluation object to the optimal solution as the basis for evaluating the merits. The specific algorithm steps are as follows.

(1)In order to eliminate the magnitude effect among different attributes and make each attribute feature equally expressive, the raw data are first normalized. Let the matrix of the multi-attribute decision problem matrix be $A = (a_{ij})_{m\times n}$

$$b_{ij} = \frac{a_{ij} - \overline{a_j}}{s_j}, i = 1, 2, ..., m; j = 1, 2, ...n \tag{32}$$

(2)Create a weighted specification matrix $C_W = (c_{ij}^w)_{m\times n}$

(3)Calculate the distance from each alternative to the positive ideal solution and the negative ideal solution. The distance of the alternative d_i to the positive ideal solution is:

$$s_i^+ = \sqrt{\sum (c_{ij} - c_j^+)^2}, i = 1, 2, ...m \tag{33}$$

The distance of the alternative d_i to the negative ideal solution is:

$$s_i^- = \sqrt{\sum (c_{ij} - c_j^+)^2}, i = 1, 2, ...m \tag{34}$$

(4)Calculate the queuing index value (i.e., the composite evaluation index) for each program:

$$Pr = \frac{s_i^-}{s_i^- + s_i^+} \tag{35}$$

Table[7] shows part of the results of the sorting

Table 7: Rank Result

Global Id	Score
{47C1EC30-0EC5-4067-9B78-D630DE5B293E}	0.824
{FF28F63A-4464-46CF-91B8-719EF7CB1A0F}	0.606
{8A974AA4-B09C-4D0E-98CC-AA3CFE105110}	0.556
{FA2E1639-6C6E-4841-A4B0-7C322F2B8398}	0.545
{A37637D5-7A74-4DDC-BF74-B8D8334738BB}	0.535
...	...
{4E587C5B-ADE5-4C46-879D-E3C8701DC462}	0.081
{ADECEBD1-248D-4ECF-A909-B17CBE0B17F3}	0.081
{A8F5AA22-3F29-4533-A0DD-21204DA91E70}	0.079
{940EEAF4-A404-485E-8A03-A7B5A792DAF0}	0.078
{098FDC08-ACCB-4D67-8052-59E167196DBC}	0.076

4.1 Model Update

After our analysis of the laboratory review dates, it is found to be agglomerative and uncertain. Also consider that the span of our spread model for the prediction of the Asian giant hornet is half a year. Consequently, we decided to update the model once every six months.

When August is reached, at this time the Asian giant hornet is ready to enter the reproduction stage. We change the value of P to 1 for the region where the new positive points appeared in the past six months. we also add the new comments to the text base and reuse the TextRank algorithm to update the keywords.We also add the new images to the image training set to optimize the training of the convolutional neural network.

When February comes, the Aisan giant will enter the breeding stage in the following six months. At this point, we will update the model with reports from the previous six months. The density of reported points in each region during these six months will be used as the potential coefficient of the population in that region. And we update the spread model coefficients γ using the nonlinear method of least squares withthe positive reports for these six months.

In both updates, we also add the new comments to the text base and reuse the TextRank algorithm to update the keywords.We also add the new images to the image training set to optimize the training of the convolutional neural network.

When there is a certain amount of added data, we can consider re-performing K-Means clustering for a more reasonable division of Washington State. Additionally, it is worth noting that when government agencies take action to exterminate the Asian giant hornet or reduce its resilience to the environment, we also should update the P and k for each region in the model update to make adjustments.

5 Sensitivity Analysis

Variance-based sensitivity analysis is a type of global sensitivity analysis. Working within a probabilistic framework, it decomposes the variance of the model or system output into fractions

attributable to the set of inputs or inputs. Variance-based sensitivity measures are attractive because they can measure sensitivity across the input space, can handle nonlinear responses, and can measure the effects of interactions in nonadditive systems.

We use the Sobol method to perform sensitivity analysis on the picture confidence, negative text confidence, positive text confidence, and distribution confidence of the model. As seen from Figure[9], the distribution confidence shows a gentle sensitivity in our model, while the picture confidence shows a very strong robustness. This is similar to the results of our previous importance assessment on them.

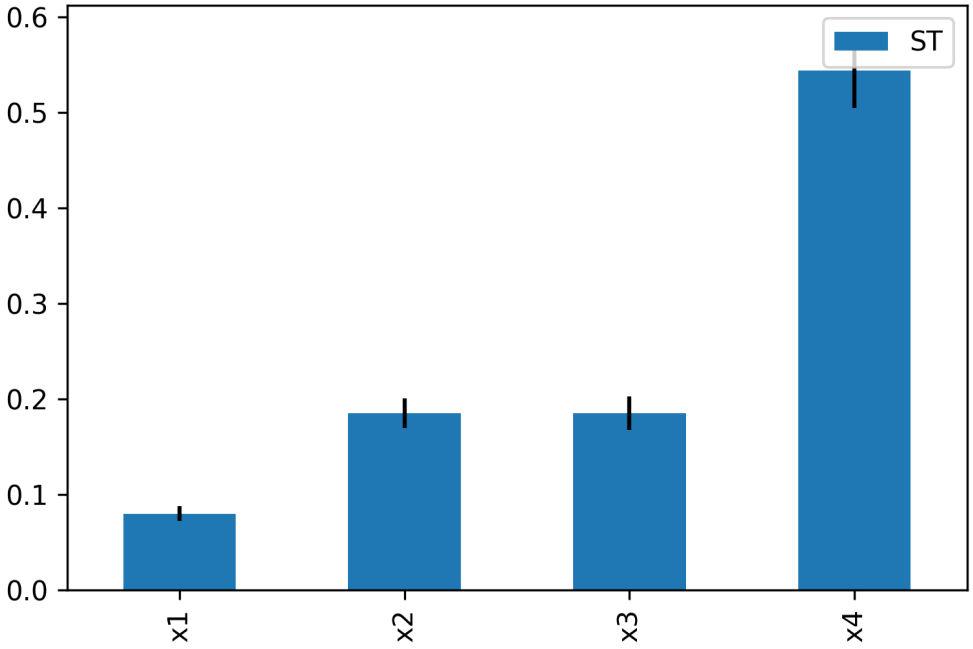

Figure 9: Sensitivity Analysis

6 Species Eradication Metrics

Quantifying whether the Asian giant hornet is extinct in Washington State is a difficult task. Because of the incompleteness and uncertainty of the data we have, we decided to transform the problem into the extent to which the Asian giant hornet was eradicated in Washington State. When it is eradicated, the probability of having the Asian giant hornet in all regions of Washington is 0. If it is not eradicated at all, the probability of having the Asian giant hornet in all regions of Washington is 1.

$$E = \frac{\sum_{i=1}^{K} P_i S_i}{\sum_{i=1}^{K} S_i} \tag{36}$$

where K denotes the number of regions and P denotes the probability of the Asian giant hornet occurring in each region. S is the area of each region.

7 Strengths and Weaknesses

7.1 Strengths

- Our model uses nonlinear method of least squares to update the data, which enhances the model store robustness and compatibility.
- All variables are fully integrated to assess the confidence level of the report in a multidimensional manner. Weights were assigned using FAHP and ranked using TOPSIS to enhance model store interpretability.
- We used the cutting-edge TextRank algorithm for text keyword extraction and constructed a feature dictionary.
- We use our own constructed feature dictionary to further classify the negative reports, so that the accuracy of image detection has been effectively improved.

7.2 Weaknesses

- Our dispersion model oversimplifies the objective environment in the absence of information.
- Due to the limitation of data, the feature dictionary we built is still not representative.
- VGG-16 neural network model has a low recognition rate for images.

8 Conclusion

The Asian giant hornet is an extremely aggressive species that poses a great threat to both the ecosystem and the agricultural economy. Using the results of our spread model development, we find that without immediate human intervention, the species could rapidly invade Washington State and cause incalculable damage.

Faced with a large number of unconfirmed reports, we combine both distribution confidence, text confidence and image confidence, using the TextRank algorithm, logistic Stix equation, convolutional neural network, nonlinear method of least squares optimization, and so on, to construct a report ranking model to rank the unconfirmed reports in order to help the relevant authorities develop a prioritized investigation strategy.

9 Our Memorandum

MEMORANDUM

DATE: February 9, 2021

TO: Washington government agency

FROM:MCM 2021 Team

SUBJECT:Strategies for dealing with AGH

In September 2019, the Asian giant hornet was spotted on Vancouver Island, British Columbia, Canada, and since then, sightings have been occurring in Washington State. Given the danger of the Asian giant hornet, it is imperative that government agencies monitor each location at all times. However, due to the conflict between limited human resources and the huge number of unconfirmed reports, we develop a mathematical model based on the confidence level of the reports to rank them and assist government departments in their decision making.

To predict the spread of the Asian giant hornet, we develop the Pairwise Long-Distance Dispersal Model, which uses K-Means clustering to divide the Asian giant hornet reporting sites in Washington State into different regions. We use nonlinear method of least squares to optimize the parameters and solve for the presence probability of the Asian giant hornet in each region to obtain the final dispersion model.

Through our future simulations, we find that without strong government intervention, the Asian giant hornet would rapidly spread throughout Washington State and multiply. This would be a huge blow to the ecological state and the economy of beekeeping in Washington State.

Confronted with a large number of unverified reports and limited investigative facilities, we have to rank all reports in confidence in order to develop a reasonable resource prioritization strategy. Combined with existing population propagation prediction models, we introduce TextRank-based text confidence and images confidence based on VGG-16. The weights of the variables are determined using the Fuzzy Analytic Hierarchy Process. Finally, the confidence ranking is performed using TOPSIS to obtain the likelihood of all unconfirmed reports of being positive reports.

Based on our ranking of all unveried reports, government agency can develop strategies to allocate resources for investigations. After considering other cost factors, investigators should prioritize the location of reports with high confidence to conduct investigations and avoid wasting resources.

As time goes by, new reports are bound to be added. We choose to update the model every six months. Each update is enabled to adjust the model parameters, update the keyword library and VGG-16 network, and improve the accuracy of the robustness of the model.

Since data is the basis of the entire model, it is advisable for government departments to develop strategies to encourage more people to upload reports to collect more valid data so that the model can be more accurate after each update.

We define the metric of the likelihood of the Asian giant hornet's existence in Washington State as a reflection of the probability of its eradication in Washington State.

This metric can be used as an indicator of the severity of the Asian giant hornet invasion in Washington State. When it is 1, it indicates in some degree that the state of Washington has been completely invaded, and when it is 0, it indicates in some degreethat it has been eradicated. Therefore, the government can establish a risk level based on this metric.

References

[1] Brian Leung, Oscar J Cacho, and Daniel Spring. Searching for non-indigenous species: rapidly delimiting the invasion boundary. *Diversity and Distributions*, 16(3):451–460, 2010.

[2] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

[3] Marina J Orlova-Bienkowskaja and Andrzej O Bieńkowski. Modeling long-distance dispersal of emerald ash borer in european russia and prognosis of spread of this pest to neighboring countries within next 5 years. *Ecology and evolution*, 8(18):9295–9304, 2018.

[4] Christelle Robinet, Christelle Suppo, and Eric Darrouzet. Rapid spread of the invasive yellow-legged hornet in f rance: the role of human-mediated dispersal and the effects of control measures. *Journal of Applied Ecology*, 54(1):205–215, 2017.

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

[6] Michael J. Skvarla. Asian giant hornets. `https://extension.psu.edu/asian-giant-hornets`.

[7] Kiev V.S. Mikhailov. Navigation and lotion. `https://deckofficer.ru/titul/study/item/navigatsiya-i-lotsiyas`.

[8] Gengping Zhu, Javier Gutierrez Illan, Chris Looney, and David W Crowder. Assessing the ecological niche and invasion potential of the asian giant hornet. *Proceedings of the National Academy of Sciences*, 117(40):24646–24648, 2020.

Appendices

Part of the code:

```
import os
import cv2


def save_image(num: int, image, folder: str):
    """Save the images.

    Args:
        num   : serial number
        image : image resource
        folder: folder name

    Returns:
        None
    """
```

```
    image_path = '../data/images_data/{}/{}.png'.format(folder, str(num))
    if image is not None:
        cv2.imwrite(image_path, image)


def extract_image(path: str, img_folder: str):
    """Extract images.

    Args:
        path      : video path
        img_folder: image folder name

    Returns:
        None
    """
    vc = cv2.VideoCapture('../data/raw_media_files/{}'.format(path))  # import video files

    # determine whether to open normally
    if vc.isOpened():
        ret, frame = vc.read()
    else:
        ret = False

    count = 0  # count the number of pictures
    frame_interval = 20  # video frame count interval frequency
    frame_interval_count = 0

    # loop read video frame
    while ret:
        ret, frame = vc.read()
        # store operation every time f frame
        if frame_interval_count % frame_interval == 0:
            save_image(count, frame, img_folder)
            count += 1
        frame_interval_count += 1
        cv2.waitKey(1)

    vc.release()


if __name__ == '__main__':
    video_file_list = [i for i in os.listdir('../data/raw_media_files') if
                       ('.MOV' in i) or ('.mov' in i) or ('.MP4' in i) or ('.mp4' in i)]
    for j in video_file_list:
        print(j)
        folder_name = j.split('_')[0]
        os.mkdir('../data/images_data/{}'.format(folder_name))
        extract_image(j, folder_name)
```