Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2123823

Trace the Buzz: A Comprehensive Study on the Reports About Hornets

Summary

With the fast rise of trade and tourism, invasive species have become a global concern. This paper aims to help the government make full use of the public reports to rationalize the deployment of limited resources to conduct investigations into the sparrow hornet.

For Problem (a), to predict the spread of the sparrow hornet, we construct a **Combined Forecast Model**. From the **spatial** perspective, **factor analysis** is used for variable selection (see Table 4.1) to measure the **environmental suitability** in a specified area. From the **temporal** perspective, we use **K-means clustering** on the three types of sightings (positive, unverified, unprocessed) in different months. The **Gaussian mixture model (GMM)** is applied to calculate the likelihood distribution of sparrow hornets in different months. The combination of the two models illustrates the spread trend of hornets. The results show that V_s is less than 0.1 throughout the period, indicating our model's high precision.

For Problem (b), we aim to predict the likelihood of misclassification. First, we construct a **Convolutional Neural Network (CNN)** for image recognition and then carry out the **text mining with an improved tf-idf model**. Finally, we combine them for better estimation. The CNN predicts very well while the text mining can tell us the relative likelihood. After we combine these two methods, AUC(a measure of accuracy) increases by 1.08%.

For Problem (c), to prioritize investigation of the positive reports, we establish a comprehensive model by combining the two models in Problem (a) and Problem (b), which is based on **Regularized Logistic Regression**. As a result of this model, we give more comprehensive judgments, with more than 99% accuracy.

For Problem (d), to update our model with new reports, **pseudo labelling** is introduced to utilize the unused (new) data. Compared with the result without pseudo labelling, AUC has increased to 0.96 by 2.13%. In terms of the appropriate update frequency, we use the **subjective evaluation method** to score each area based on five factors. Finally, we calculate the reporting frequency of a specific area based on its objective conditions and find that the highest number of reports occurs in July and August (5.2/month).

For Problem (e), to determine the evidence of eradication in Washington, we use the **regression model with longitudinal data** to modify the model in Problem (a). The improved model's advantage lies in its ability to predict changes in species populations with human intervention. We select local bee and sparrow hornet populations in 20 areas as samples. The state between two extreme points of the population curves indicates the eradication evidence.

At the very last, we analyze the strengths and weaknesses of our model as well as its sensitivity, whose results show that our model has high robustness, precision and accuracy. After that, a memo is attached.

Keywords: Environmental suitability; K-means clustering; CNN; Text mining; Pseudo labelling; Regression model with longitudinal data

Contents

1 Introduction

1.1 Background

Nowadays, biological invasions are a global concern in agriculture, food production and biodiversity. Invasive species have detrimental effects on the newly colonized areas, such as impoverishment of local species assemblages, as well as the decline of critical insects providing herbivore control and pollination services [1].

The recent finding of Vespa mandarinia (Asian giant hornet) in Canada and the USA has prompted concern that it could become an invasive species. Until the Entomological Society of America decides on the official common names for V. mandarinia, we are suggested to use '**sparrow hornet**' [2]. Sparrow hornet tends to nest in low mountain foothills, lowland forests or green space in urban landscapes [3]. Sparrow hornet is a quarantine pest for the USA for it will cause a significant loss to beekeepers [4]. Therefore, much care should be taken to predict the migration of hornets over time and identify manual reports to deploy government resources more efficiently to investigate the problem.

1.2 Problem Restatement

(a) Develop a model to predict the sparrow hornet's spread over time, and analyze the degree of precision of the model.

(b) Taking the look-alike species into account, study how to predict the likelihood of a mistaken classification using only the data set file and (possibly) the image files provided.

(c) Set a goal for taking priority in investigating the reports that are most likely to be positive sightings, and discuss the way to conduct the classification analysis.

(d) Address how to update the previous model using additional new reports over time, and determine the optimum frequency of new reports.

(e) Based on the model, find out what would constitute evidence that the pest has been eradicated in Washington State.

1.3 Problem Analysis

For Problem (a), we plan to predict the spread of sparrow hornets from both spatial and temporal perspectives. First, the spread of hornets may be constrained by many natural factors, so we consider using factor analysis to calculate the environmental suitability, as this method is suitable for cases where data is not available or with low quality [5]; second, the spread may be affected by seasonal and cyclical factors, so we can try to use clustering methods to determine the aggregation of hornets in different periods.

For Problem (b), the data we can utilize are the reporters' images and notes. We should make the most of them by statistical methods.

For Problem (c), to work out how to estimate the likelihood of reports being positive and prioritize investigation of them, we decide to consider all aspects of information involved above and establish a new evaluation model based on models we have applied in Problem (a) and Problem (b).

For Problem (d), pseudo labelling is an excellent way to improve the model accuracy when more unlabeled data are now available [6]; thus we want to use it to update our model based

on the additional new reports. As the frequency of model updates is closely related to the frequency of report submissions, we need to determine an appropriate number of reports in a given period.

For Problem (e), to determine whether the control measures have eradicated the pests, we can use the regression model with longitudinal data to modify the spread model by taking the changes of pest population with human intervention into account. With data on population changes after taking controlled measures, we can determine the point in time and evidence of pest eradication.

2 Assumptions and Justifications

To simplify our problems, we make the following basic assumptions, each of which is adequately justified.

(a) In each month, the number of clustering centers of sparrow hornets is positively correlated with their activity level, reaching a peak in July and August.

(b) The activity period of sparrow hornets was from April to October of year. During other times, the sparrow hornets were in hibernation. The events of sparrow hornets found by the public during these limited periods can be considered unreliable and ignored.

(c) The ethnic influence of sparrow hornets is normally distributed over distance.

(d) The environmental suitability of the same area is generally consistent from year to year.

3 Notation

Symbol	Description
$N(x, u_i, d_i)$	The probability density function of the i-th Gaussian model
$P(x)$	The likelihood of sparrow hornets appearing in the position x
π_i	weight of the i-th Gaussian model
β	Weight of Model 2
V_s	coefficient of variation
c	Term frequency
$P(Neg)$	Likelihood of negative things
$P(Pos)$	Likelihood of positive things
P	The actual number of positive reports
N	The acutal number of negative reports
θ	Regression parameters
λ	regularization coefficient
b_k	Fixed effect factor
η_{ki}	Random effect factor

4 Analysis and Modelling

4.1 Spread of sparrow hornet over time

After analyzing the provided data set, we can ensure that the spread of sparrow hornet is predictable.

Understanding the role of biotic (species ecology, predation and interspecific competition) and abiotic (anthropogenic role, climate change, geographical barriers) factors has become the major challenge in ecology and biogeography. In this problem, however, we rely on abiotic variables because little is known about other biotic factors, such as species evolution, predators and competitors, which may affect the location of sparrow hornet [7].

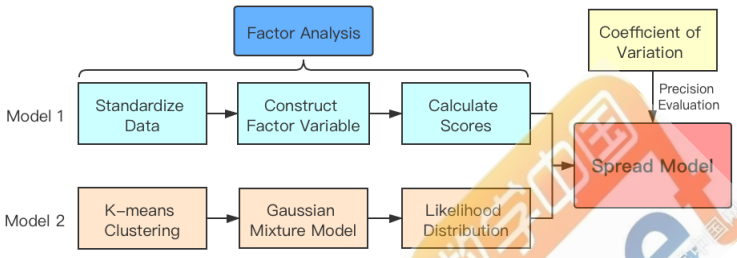


Figure 4.1: The procedures of establishing Combined Forecast Model

4.1.1 Data Preprocessing

We first weed out events where the date is incorrectly formatted though the number of these events is small. We screen out the data before September 2019. This is because sparrow hornets were first found in Sep. 2019 in Vancouver Island, and in Dec. 2019 in Washington State. So any report before Sep. 2019 is certain to be negative; For each year, we select the data from April to October. According to reference [8], sparrow hornets will go through hibernation outside this interval. So reports from Apr. to Oct. are much more meaningful and valuable.

4.1.2 Model 1: Factor Analysis Method

The basic idea of factor analysis is to group variables according to the correlation so that variables are more correlated within the same group and less correlated in different groups. Each group of variables represents a basic structure, which is referred to as a common factor. So it is possible to describe each component of the original observation in terms of a linear function of unmeasured common factors and the special factors.

The factor analysis can be expressed in the following model:

$$\begin{cases} x_1 = a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m \\ x_2 = a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m \\ \vdots \\ x_p = a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m \end{cases} \quad (m < p)$$

In which, x_1, x_2, ..., x_p mean p standardized variables with a mean of 0 and a standard deviation of 1, F_1, F_2, ..., F_m refer to m factor variables. Moreover, this model can be translated into the form of the matrix.

$$X = AF + a\varepsilon$$

in which, F is the common factor; A is the factor loading matrix, which is the loading of the i-th original variable on the j-th factor variable; a_{ij} is the projection of x_i onto the coordinate axis F_j if the variable x_i is viewed as a vector in an m-dimensional factor space; ε is the special factor, indicating the part of the original variable that cannot be explained by the factor variables (which can be thought of as the residuals in a multiple regression model).

Results of Model 1 Environmental suitability is defined as the conditional probability of occurrence of a species given the state of the environment at a location [9]. Many abiotic drivers determine sparrow hornet colonies distribution. For instance, the abundance of sparrow hornets is positively associated with green spaces in urban landscapes [3]. On the other hand, sparrow hornet has been described as a species highly sensitive to heat and extreme climate conditions, and negatively associated with high temperatures [10]. As a result, we consider as many documented or relevant impact variables [a] as possible when carrying out the factor analysis, in order to adequately estimate the environmental suitability of the Northwest United States. Finally, 16 variables that can explain sparrow hornet current distribution are selected.

Table 4.1: Variable used in the analysis

Classification	Factor	Abbreviation
Climate	Annual mean temperature (°C)	T_m
	...	
	Annual mean radiation (MJ/m^2)	R_m
Habitat structure	Forest stands (%)	H_for
	...	
	Scrubland (%)	H_scr
Vegetation productivity	Normalised difference vegetation index	HDVI
Hydrography	Average distance to rivers (m)	Dist_riv
Topography	Elevation (m)	Elv
Human activity	Average population	D_pop

At last, values of all the variables are converted into z-score.

It can be seen from the cumulative percentage of variance that the first four principal components already explain nearly 87.9% of the total variance, so these could be selected for analysis. The scree plot on the right illustrates that the fold's slope is steeper for the first four principal components. However, it tapers off later, which is another evidence that it is appropriate to take the first four principal components.

[a]Data Source: Climatic variables are retrieved from the Weather Atlas (`https://www.weather-us.com`). The website Elevations and Distances in the United States (`https://pubs.usgs.gov/gip/Elevations-Distances/elvadist.html`) is used to gather topographic data. The data of vegetation productivity and hydrography can be obtained from `http://glovis.usgs.gov/`. Anthropogenic drivers such as population are derived from the Census Bureau (`https://www.census.gov/`).

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.031	50.196	50.196	8.031	50.196	50.196	6.706	41.912	41.912
2	3.029	18.929	69.125	3.029	18.929	69.125	3.684	23.027	64.939
3	1.834	11.459	80.585	1.834	11.459	80.585	2.283	14.271	79.210
4	1.172	7.326	87.911	1.172	7.326	87.911	1.392	8.701	87.911
5	1.035	6.469	94.380						
6	.596	3.723	98.103						
7	.304	1.897	100.000						
8	1.976E-15	1.235E-14	100.000						
9	4.252E-16	2.657E-15	100.000						
10	3.064E-16	1.915E-15	100.000						
11	2.096E-16	1.310E-15	100.000						
12	9.555E-17	5.972E-16	100.000						
13	-4.054E-17	-2.534E-16	100.000						
14	-1.699E-16	-1.062E-15	100.000						
15	-2.519E-16	-1.574E-15	100.000						
16	-3.687E-16	-2.304E-15	100.000						

Extraction Method: Principal Component Analysis.

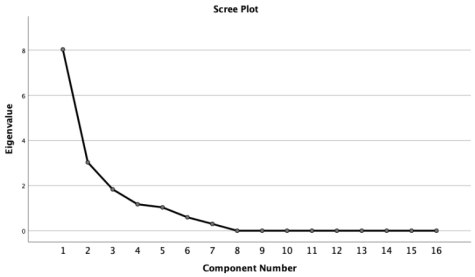


Figure 4.2: The results of total variance explained and scree plot

Next, we apply the formula $Y = Z_x \times t$, where Z_x is the normalized matrix and t is the normalized orthogonal eigenvector matrix to obtain the values of the four principal components. Based on the percentage of variance in Figure 4.2 for the weighting, we can calculate the overall score and the environmental suitability of 14,260 selected points.

$$Y_{total} = 0.50196Y_1 + 0.18929Y_2 + 0.11459Y_3 + 0.07326Y_4$$
$$ES = 500 + 100Y_{total}$$

From the two figures below, we can see that the USA and Canada's western coasts are ideal for the survival of hornets. A second peak occurs in and around the North Cascades National Park on the Rosario Strait's east coast. The Kootenay National Forest in the interior is the third best place for hornets - although not as suitable as the coastal areas, the environmental suitability has been primarily maintained at around 300. By comparison, we can see that 'high-scoring' areas all share some common and distinctive characteristics, such as high forest cover and NDVI, proximity to water sources, low human presence etc. These findings are consistent with the habits of sparrow hornets recorded in related research material [3] [10], so we consider the level of environmental suitability calculated by factor analysis is convincing to some extent.

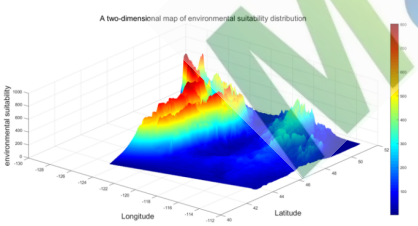


Figure 4.3: Two-dimensional map of environmental suitability

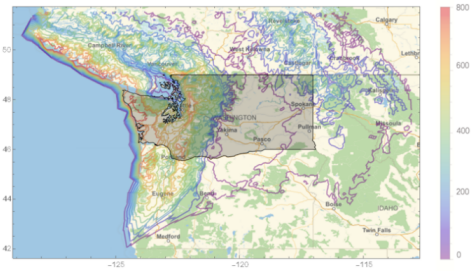


Figure 4.4: Contour map of environmental suitability

4.1.3 Model 2: K-means clustering algorithm and Gaussian Mixture Model

In Model 2, considering that the sparrow hornets live in groups and are distributed around the queen, we first thought of applying the K-means clustering algorithm to find the sparrow hornets' cluster centres. When applying the K-means clustering algorithm to find the clustering centers, we, in order to improve the reliability of the data, directly delete the data judged as "Negative ID" and use the remaining part.

According to the reference [8], sparrow hornets' growth can be divided into six stages, and

we correspond to months to determine the number of clustering centers in each month. We preset more clustering centers in the months when sparrow hornets are active. The specific settings have been shown in the figure below.

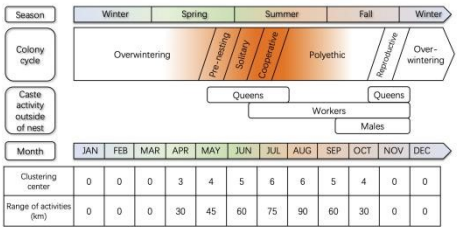


Figure 4.5: Variation of the number of clustering centers by month

Since the initial point selection of the K-means algorithm is random, this algorithm may fall into a local optimal solution instead of a global optimal solution. To avoid this situation, we calculate many times and find the situation with relatively good output as the global optimal solution. Based on the processed data, we finally used the K-means clustering algorithm to find the clustering centers' location in each month and calculated the categories of each event. The results are shown in Figure 4.6.

After getting each cluster center's position and the number of ethnic groups in each month, we began to apply the Gaussian mixture model to calculate the likelihood distribution of sparrow hornets in various places within the specific geographic area.

Let us briefly introduce the Gaussian mixture model. The Gaussian mixture model uses the Gaussian probability density function (normal distribution curve) to quantify things accurately. It is a model based on Gaussian probability density function (normal distribution curve) that decomposes things into several.

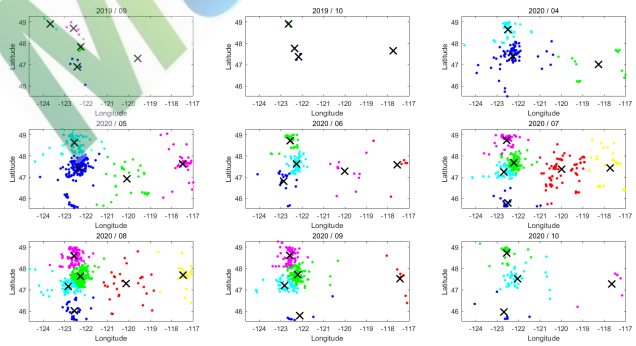


Figure 4.6: The distribution of clustering center

The mixed Gaussian model consists of K Gaussian models (that is, the data contains K clustering centers). The probability density function of GMM is as follows:

$$p(x) = \sum_{i=1}^{K} p(i)p(x|i) = \sum_{i=1}^{K} \pi_i N(x, u_i, d_i)$$

where x represents the position of the sample, represents the likelihood of sparrow hornets appearing in the position x, $p(x|i) = N(x, u_i, d_i)$ represents the probability density function of the i-th Gaussian model, and u_i, d_i, π_i represent the mean, standard deviation and weight of i-th Gaussian model respectively. π_i is proportional to the number of ethnic group members, and $\sum_{i=1}^{K} \pi_i = 1$. The greater the number of ethnic group members, the greater the influence, which is very reasonable.

In our model, the probability of sparrow hornets appearing in various places within a specific geographic area is affected by multiple cluster centers in this area. The change curve of this effect with distance is the probability density curve of the normal distribution.

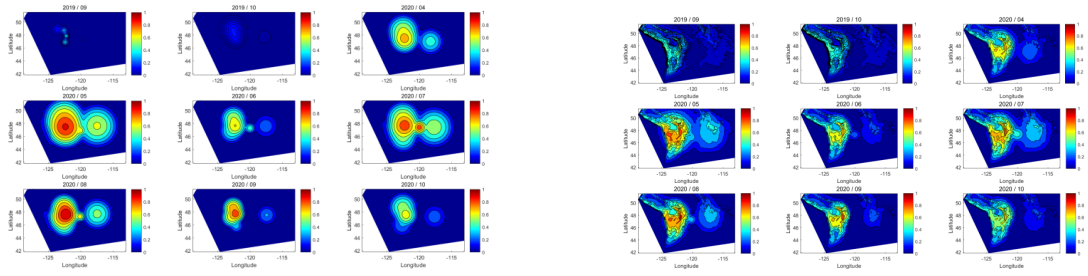
Results of Model 2 We calculate the likelihood of sparrow hornets appearing everywhere in a specific geographic area in each month, rescale it between 0 and 1, and finally draw a contour map in Figure 7(a).

4.1.4 Combined Forecast Model

Model 1 calculates the environmental suitability according to the region's natural conditions and scales it to between 0 and 1 to obtain the likelihood distribution of sparrow hornets. Model 2 predicts the likelihood distribution of sparrow hornets in different months based on the existing data of the distribution of sparrow hornets. And then we combine the results of these two models at a ratio of $1 : \beta$. In this part, let us set $\beta = 1$. Furthermore, we will study how to select its value in Sensitivity Analysis. Finally, we obtained the likelihood distribution of sparrow hornets in different months in recent years, as shown in Figure 7(b).

Since the K-means clustering algorithm results are not necessarily consistent each time it runs but are similar, this causes the results obtained by our final model to fluctuate within a specific range. Therefore, we need to calculate the precision of our model predictions. Our definition of precision is the degree of error in the prediction results of the comprehensive prediction model, which corresponds to the magnitude of the error. Therefore, we use the coefficient of variation V_s to reflect the level of precision. The smaller the coefficient of variation V_s, the smaller the variation range of the prediction results, which means the higher the precision level. The calculation of the coefficient of variation is as follows where σ and $\bar{X}$ represents standard deviation and average of data X.

$$V_s = \frac{\sigma}{\bar{X}} \tag{4.1.1}$$



(a) results of Model 2 (b) results of the combined model in Problem (a)

Figure 4.7: Likelihood distribution of sparrow hornets in different months

We calculated the coefficient of variation of outputs predicted by month. From Figure 4.8,

we can ensure that our model's precision is high enough to estimate the likelihood.

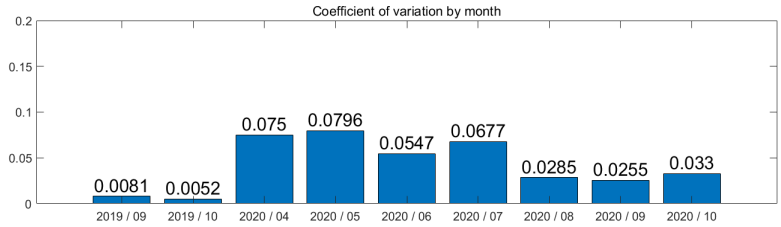


Figure 4.8: V_s of outputs by month

4.2 Prediction of Mistaken Classification

After inspection of the given spreadsheet, we find that most reported sightings are identified as mistaken ones. Hence, in this part, we are to establish a model to help predict the likelihood of a mistaken classification.

N.B. We may call something such as a report, a figure, a note or even a word, as 'positive' or 'negative' in the following. This means its corresponding 'Lab Status' is positive or negative.

4.2.1 Model 3: Image Recognition Based on CNN

Data Preprocessing While the given data was cleaned in the last part, it still shows a severe imbalance of the data distribution, in that, there are 2028 negative reports versus only 14 positive ones. This is not good for the following learning of sparrow hornets' features.[b] We come up with two approaches to solve this. For one thing, we increase the number of positive samples. In detail[c],

- we flipp the positive figures vertically and horizontally to triple its number;
- we download some figures from reliable websites.[d]

For another, we decreased the number of negative samples. In detail, we randomly selected some of them, making its number equal to positive ones.

Through preprocessing, now there are 258 positive samples and 258 negative ones. We mix them uniformly and split them by the ratio of 7 to 3 for training and test.

As to the targets, we label the positive and negative target as 1 and 0 respectively, denoting the likelihood of a positive target. Finally, there are 411 training samples & targets and 177 test samples & targets.

Convolutional Neural Network Convolutional neural networks(CNN) have played an important role in the history of deep learning, especially in object recognition. For instance, they

[b]Actually we had once attempted to use all of them to train our model. It was found that the model was inclined to predict all the samples as negative regardless of the real targets.

[c]Resizing and normalization are also incorporated in preprocessing. They are not mentioned in the main body since they made no contributions to the increase of positive samples

[d]Data Source: Asian Giant Hornet(`https://en.wikipedia.org/wiki/Asian_giant_hornet`), Get to know the Asian giant hornet, or 'murder hornet'(`https://agrilifetoday.tamu.edu/2020/05/11/get-to-know-the-asian-giant-hornet-or-murder-hornet/`)

were used to win many contests, including the ImageNet object recognition challenges [11]. So here we establish a CNN for preliminary prediction.

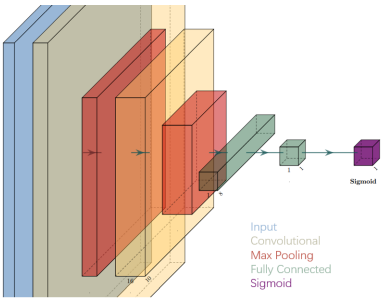


Figure 4.9: The architecture of our CNN

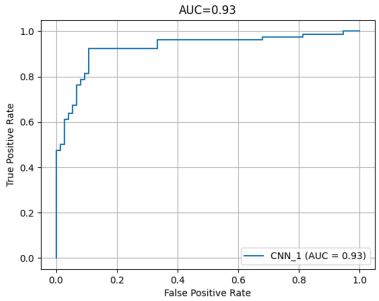


Figure 4.10: ROC curve of CNN's prediction

We construct our CNN with PyTorch. The architecture of the network is displayed in Figure 4.9(drawn with *TikZ*). There are two convolutional layers, two pooling layers and two fully connected layers. The shape of the input is $(b \times 3 \times 96 \times 96)$, where b denotes the batch size, 3 denotes the number of channels, and 96×96 are the height and width. The shape of the output is (b). Train for 200 epochs.

The test accuracy reaches its peak of 0.8968 at the 93rd epoch. The model at this epoch is saved as the final predictor. We plot the ROC Curve in Figure 4.4.10. The horizontal and vertical axes are the false positive rate (FPR) and true positive rate (TPR). AUC denotes the Area Under ROC Curve. A model with AUC closer to one shows better performance in prediction. Now our model's AUC equals to 0.93. With only 411 training samples, this result is quite acceptable.

4.2.2 Model 4: Text Mining in the Level of Words and Sentences

The given data set also includes some notes attached by the reporters, which are not to be ignored. We have just modelled the prediction of $P(Pos)$ and $P(Neg)$ based on images, while these notes are also thought to be helpful to our prediction.

Introduction of the improved tf-idf model In information retrieval, tf–idf, short for term frequency–inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus [12]. Such a model is made up of two parts literally.

Term frequency, i.e., c_w, is the number of times a term w occurs in a document, while inverse document frequency is a measure of how much information the word provides, i.e., if it's common or rare across all documents. In most cases, a term with a higher c_w tends to contribute more to the document. Nevertheless, a stopword like 'the' or 'I'm' can hardly give us any information.

This inspires us to estimate the 'positive' likelihood of each word, viz., $P_w(Pos)$ according to its c_w. What is different here is about 'Lab Status'. For instance, a 'negative' word will contribute inversely to the note. So we suggest an improved tf-idf model according to its c_w considering the Lab Status.

Data Preprocessing & Qualitative Analysis We continue using the data from Sep. 2019. For each note, eliminate the stopwords [e]. Then we group these notes by its Lab Status. Furthermore, we quantify Lab Status in the form of $P(Pos)$.

For 'Positive ID', $P_{pos}(Pos) = 1$. For 'Negative ID', $P_{neg}(Pos) = 0$. For 'Unverified', $P_{uv}(Pos) = 0.5$, because even official specialist cannot identify it. Also, a matter with the likelihood of 0.5 is hardest to predict. For 'Unprocessed', $P_{up}(Pos) = \dfrac{P}{P+N}$, where P and N denote the actual number of positive and negative reports separately. This is because it has not been classified rather than be unable to be classified. So we use the mathematical expectation as $P(Pos)$.

Let n, p, u_p, u_v represent the number of the four kinds of reports mentioned above in order. We will have these two equations.

$$N = n + u_p \times \frac{N}{N+P} + 0.5 \times u_v \tag{4.2.1}$$

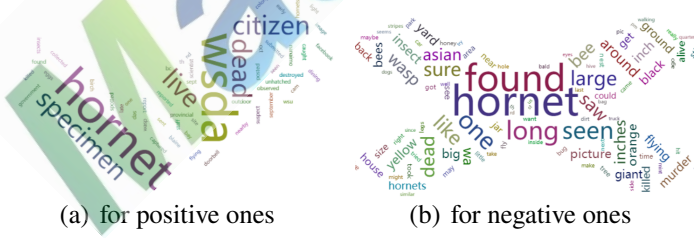
$$P = p + u_p \times \frac{P}{N+P} + 0.5 \times u_v \tag{4.2.2}$$

Solve the simultaneous equations and we can obtain the value of N and P.

$$T = N + P = n + p + u_v + u_p = 4440$$

$$N = \frac{n + u_v/2}{1 - u_p/T} = 3251, \quad P = T - N = 1189$$

After filling $P(Pos)$ for each note, we create the word clouds of positive notes and negative notes. This can reflect a word's 'positive' or 'negative' attribute qualitatively.



(a) for positive ones (b) for negative ones

Figure 4.11: Word clouds of notes

The word 'hornet' often appears both in positive and negative notes, which can be easily explained by the creature's name. Additionally, the word 'wsda' is conspicuous in the positive word cloud. This may be because the notes including 'wsda' are submitted by WSDA or its staff with more knowledge about the sparrow hornet.

Modelling of a Word For a 'Lab Status' i, $P_i(Pos)$ can be regarded as a sort of score. We can grade each word except stopwords.

Now let us set the grading rule. If a word w appears c_i times for status i, it will get a score of $c_i \times P_i(Pos)$. The final score is calculated by adding four scores together. This score have

[e]The stopwords are provided by the corpus of Python's Natural Language Toolkit

no upper limit. So we divide it by the total times it appears, i.e., $\sum_i c_i(w)$. We define this value as $P_w(Pos)$. Apparently, $P_w(Pos)$ is between 0 and 1.

$$P_w(Pos) \xlongequal{def} \frac{\sum\limits_i c_w(i) \times P_i(Pos)}{\sum\limits_i c_w(i)}, \qquad i \in \{pos, neg, uv, up\} \tag{4.2.3}$$

If we rewrite Eq. 4.2.3 as

$$P_w(Pos) = \sum_i P_i(Pos) \times \frac{c_w(i)}{\sum\limits_i c_w(i)}, \qquad i \in \{pos, neg, uv, up\} \tag{4.2.4}$$

then $P_w(Pos)$ is the weighted mean of $P_i(Pos)$, whose weight equals $\frac{c_w(i)}{\sum\limits_i c_w(i)}$.

Modelling of a Note A note is composed of several words. Thus the positive likelihood of a note, viz., $P_n(Pos)$ is also a composition of $P_w(Pos)$. It is found that some words like 'Manson' and 'incident' only appears once, rendering them much less reliable for the positive likelihood. In this sense, they should contribute less to $P_n(Pos)$. This means $P_n(Pos)$ is positively correlated with a word's term frequency, c_w. Based on the above, we define $P_n(Pos)$ as

$$P_n(Pos) \xlongequal{def} \begin{cases} \dfrac{\sum_{w \in \mathbb{S}_n} \ln c_w \times P_w(Pos)}{\sum_{w \in \mathbb{S}_n} \ln c_w} & , \mathbb{S}_n \neq \emptyset; \\ \dfrac{P}{P+N} & , \mathbb{S}_n = \emptyset. \end{cases} \tag{4.2.5}$$

where $\mathbb{S}_n$ is the set of all words in the note n. As mentioned in the last part, the data distribution show a severe imbalance between negative and positive ones. So if a note is not empty, we will take a logarithm of c_w to bridge the gap. However, there also exist some reports without notes. In this case, we choose the mathematical expectation as their value.

With this definition, let us recalculate the positive likelihood of each note. Notes are grouped by 'Lab Status'(marked as G1, G2, G3, G4). The distribution of each group's $P_n(Pos)$ is shown in Figure 4.12. The subtitles represent the means of $P_n(Pos)$.

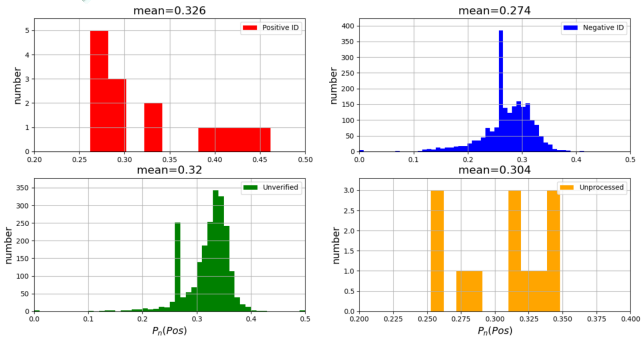


Figure 4.12: Distribution histograms of four group's $P_n(Pos)$

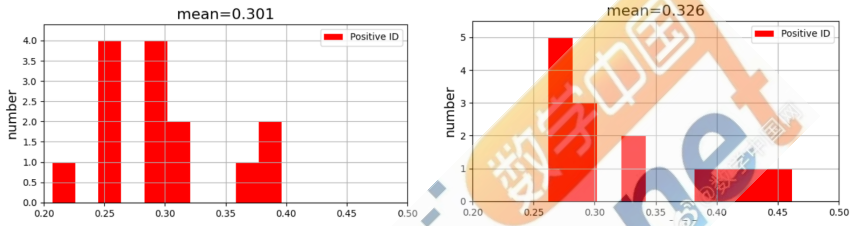
In general, G1 has the greatest mean $P_n(pos)$ and G2 has the least mean $P_n(Pos)$, which is consistent with reality. But the maximum of G1's $P_n(Pos)$ is less than 0.5. This is normal

for that many words appearing in positive notes also appear in negative ones. Moreover, the calculated value of $\frac{N}{N+P}$ is 0.262045. So it can be seen that both G2 and G3 have a spike when $P_n(Pos) \approx 0.26$, which can be attributed to their 'empty' notes.

To check the effect of the logarithm, we compare the distribution of G1's $P_n(Pos)$ considering logarithm and without it. The latter corresponding formula for a non-empty note is

$$P'_n(Pos) \xlongequal{def} \begin{cases} \dfrac{\sum_{w\in\mathbb{S}_n} c_w \times P_w(Pos)}{\sum_{w\in\mathbb{S}_n} c_w} & ,\mathbb{S}_n \neq \emptyset; \\ \dfrac{P}{P+N} & ,\mathbb{S}_n = \emptyset. \end{cases} \tag{4.2.6}$$

Figure 4.13 informs us that G1's $P_n(Pos)$ moves towards 1 in general, and the mean has increased from 0.301 to 0.326 by 8.31%. Thus Definition 4.2.5 proves to be more reliable than Definition 4.2.6.



(a) Using Definition 4.2.6 (b) Using Definition 4.2.5

Figure 4.13: Comparison of the distribution of G1's $P_n(Pos)$

4.2.3 Combination of Image Recognition and Text Mining

Although our improved tf-idf model predicts worse than the CNN, the relative likelihood is accurate. Furthermore, it measures a report from a different aspect. In this sense, we can combine these two models to increase the accuracy of prediction.

We plot the distribution of the prediction results in Figure 4.14. The two histograms differ significantly in shape, which is not suitable for linear combination. In this case, we turn to combine them in a nonlinear form.

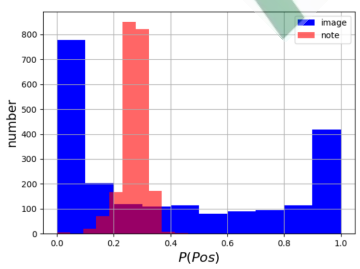


Figure 4.14: Distribution of the results

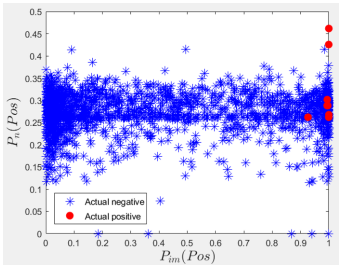


Figure 4.15: 2-D visualization of the result's distribution for training data

In order ro represent the contributions of the images and the notes, we introduce the coefficient C_1 and C_2 as their exponents. The general form is as below.

$$\ln \hat{P}(Pos) = \ln m \times \{k \times [P_{im}(Pos)]^{C_1} \times [P_n(Pos)]^{C_2} + b\} \tag{4.2.7}$$

By fitting the training data, the values of the coefficients are

$$m = 0.6489, k = 0.0956, C_1 = -39.3, C_2 = -4.4, b = -3.806$$

C_1 is nearly nine times larger than C_2, which tells us the result depends much more on the image than the textual part. This is also consistent with Figure 4.15, where actual positive points only appear when $P_m(Pos)$ is greater than 0.9.

We calculate the result of the test samples and display the distribution in Figure 4.16. The value of AUC is increased to 0.94 by 1.08%.

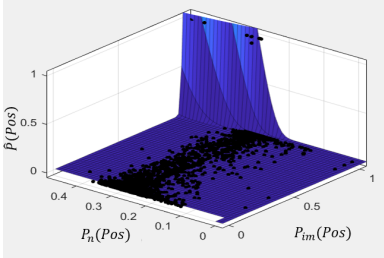


Figure 4.16: 3-D visualization of the result's distribution for test data

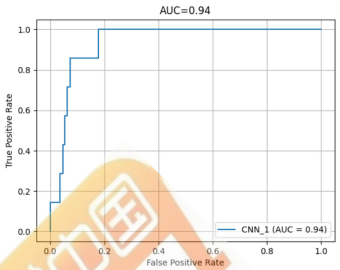


Figure 4.17: The ROC curve based on combined model in Problem (b)

4.3 Determine the likelihood of a Positve Report

Based on the work above, we intend to integrate these two models through Regularized Logistic Regression to improve prediction accuracy.

Regularized Logistic Regression We take the likelihood of sparrow hornets appearing obtained by the two models in Problem (a) and (b) as input values, respectively marked as p_{ab1} and p_{ab2}, and the actual likelihood is the output value, which is marked as y. The data set we select refers to the events that provide the picture and are judged as positive or negative, where Positive ID corresponds to 1, and Negative ID corresponds to 0. Then we randomly divide this data set into the training set and test set.

Before training, considering the input only has two features, we need to apply the feature mapping method to expand the two features into 28 features.

$$mapFeature(p_{ab}) = [1, p_{ab1}, p_{ab2}, p_{ab1}^2, p_{ab1}p_{ab2}, p_{ab2}^2, \cdots, p_{ab2}^6]^T$$

where p_{ab} and mapfeature(p_{ab}) respectively represents the 2-dimensional vector and 28-dimensional vector.

After that, we started to train the model on the training set. The cost function $J(\theta)$ is calculated as follows.

$$h_\theta(p_{ab}) = \frac{1}{1 + e^{-\theta^T p_{ab}}}$$

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[-y^{(i)}\ln(h_\theta(p_{ab}^{(i)})) - (1-y^{(i)})\ln(1-h_\theta(p_{ab}^{(i)}))] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2$$

where θ represents regression parameters, λ represents the regularization coefficient, which we set as 0.6. When the cost function $J(\theta)$ is the smallest, the regression effect is best, and the regression model can be considered the final model.

Results and Analysis To evaluate our model, we test this new model's accuracy on the test set. The final predicted result matches the real situation perfectly.

Analyzing the predicted results, we found that even though there were two misjudgments in the event, the overall accuracy rate has reached 0.99, which shows that our model is reliable. Therefore our model is helpful to determine the likelihood of positive reports positive and prioritize investigation of them.

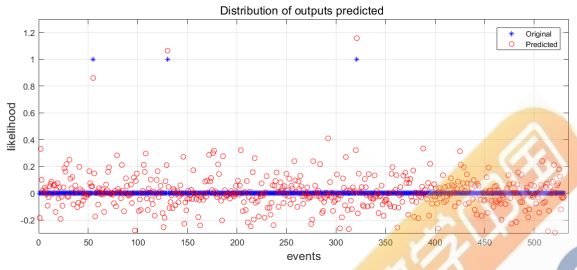


Figure 4.18: Distribution of predicted outputs

4.4 Update of the Model

Although the CNN in section 4.2.1 has a relatively high accuracy of 0.8969, it can be furtherly improved due to a great deal of new data, especially from the additional new reports. Through pseudo labelling, we are now able to make use of the unexploited data.

Pseudo labelling is the process of using the labelled data model to predict labels for unlabelled data. It is often used in semi-supervised learning, where most data are unlabelled. Nowadays, this technique is used for hyperspectral image classification and the identification of COVID-19 on chest X-rays. [11][12] Based on image recognition results of CNN, we will use pseudo labelling to improve our model. The procedure is shown in the flow chart below. (Figure 4.9)

4.4.1 Model 3*: Improved CNN With Pseudo Labelling

Pseudo labelling is the process of using the labelled data model to predict labels for unlabelled data. It is often used in semi-supervised learning, where most data are unlabelled. Nowadays, this technique is used for hyperspectral image classification and the identification of COVID-19 on chest X-rays [13][14].

Based on image recognition results of CNN, we will use pseudo labelling to improve our model. The procedure is shown in the flow chart below. (Figure 4.19)

Because additional data from Washington state on Asian hornets similar to what was provided in the spreadsheet is explicitly not allowed in Problem (c), we add the unlabeled data (unverified and unprocessed) that were not originally in the training set as "new report data" to this model, which can help us test whether additional new reports will improve our model. We reckon that the precision of the value of $P(Pos)$, which is farther away from 0.5 is more reliable and convincing. Therefore, we only label the predicted result which is either

larger than 0.9 or less than 0.1. To check this, we visualize some of the prediction results of unverified images in Figure 4.21.

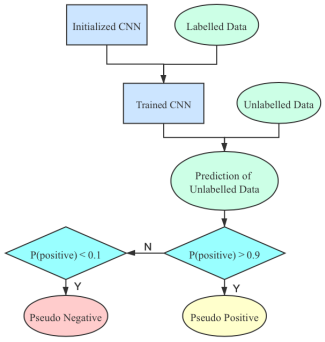


Figure 4.19: The procedures of pseudo labelling

Figure 4.20: A real sparrow hornet

The 1st and the 3rd figures' values of $P(Neg)$ are below 0.1, showing they are likely to be positive sightings. We can compare them with a real sparrow hornet in Figure 4.20. They are very similar to the three typical characteristics of a real hornet's appearance. Therefore, such prediction is relatively reliable and can be labelled as (pseudo) positive.

After pseudo labelling, we include them into the original labelled dataset and retrain our CNN obtained before. At the 7th epoch, test accuracy reaches a higher peak of 0.9096, which has increased by 1.43% compared to previous 0.8968.

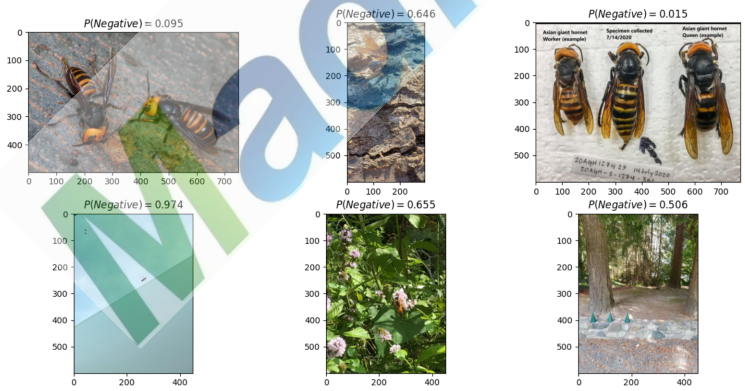


Figure 4.21: Prediction result of unverified images

Compared with the result before pseudo labelling, the value of AUC has increased to 0.96 by 1.08%. The result demonstrates the effectiveness of improving our model by setting the new report as a pseudo label added to the training model.

4.4.2 Subjective Evaluation Method

The key to solving an appropriate frequency of model updates lies in determining an appropriate number of reports in a given period. As a result, we use the **subjective evaluation method** to score each area based on different factors closely correlated with the emergence of potential reporting cases. We finally select five key evaluation indicators for analysis: **environmental**

suitability (1), seasonal factors (2), major external events (3), economic development (4), and **reporting stability (5)**. The ranking results of importance are as follows.

$$I_2 > I_3 > I_1 > I_4 > I_5$$

The specific scoring rules are: each area's score consists of two parts (initial score and additional score), where the initial score is dependent on the environmental suitability of this area, and other four factors determine the additional score. The area's score is updated every day until the threshold (e.g. 100) is reached, which means that the area should submit new reports at this point. Then, all the scores of this area are cleared, and the scoring starts again. According to the data we have, we estimate there will be a report submitted from this area in almost **six days**, which also indicates that **17 points** need to be added in average per day on a 100-point scale. After distributing the five factors' weights, the initial score (IS) formula is as follows.

$$IS_i = 17 \times Weight_3 \times \frac{ES_i}{Median_{ES}}, \qquad i = 1,\ 2,\ \ldots,\ 63$$

in which $Weight_3$ means the weight of the third factor (environment suitability) and 63 refers to the total number of the areas in Washington State.

There are also some rules regarding the determination of additional score:

(a) **Seasonal factors** require different additional scores in different months in the same area. We will consider the number of clustering centers (NCC) in Model 2 as a determinant of the additional score. Its formula is as follows.

$$AS_{aj} = 17 \times Weight_1 \times \frac{NCC_j}{Median_{NCC}}, \qquad j = 1,\ 2,\ \ldots,\ 12$$

(b) The impact of **major external events** is mainly reflected in the fact that when a positive sighting occurs in the region, the area's daily additional score can be set to 125% of the initial score for the next six months. (This effect can be stacked)

(c) The higher the level of **economic development** in the area, the higher the frequency of reporting. The following equation can express this relationship.

$$AS_{ci} = 17 \times Weight_4 \times \frac{GDP_i}{Median_{GDP}}, \qquad i = 1,\ 2,\ \ldots,\ 63$$

(d) **Reporting stability** can be positively correlated with the duration of consistently negative reports for the area. Specifically, suppose an area has more than six months of consecutive negative reports recorded. In that case, the additional value is -50% of its initial value from the sixth month until a positive record occurs in the area.

Let us consider a specific example. Suppose there is a circular area A with a radius of $30km$. In this area, $\frac{GDP_A}{Median_{GDP}} = 1.2$ and the expected environmental suitability is 84 (median is 115 in the whole state). Moreover, we assume that it is now at the beginning of May. It is also important to note that there was one positive case reported in the region at the end of April. The current estimate is that there will be no further positives in the area for six months after that.

By considering the various factors in our evaluation model, we conclude that the monthly reportable quantity in Region A varies over time in the next eight months, with the highest in July and August (5.2/month) and the lowest in November and December (1.3/month).

4.5 Evidence of Eradication in Washington State

Eradication is defined as eliminating every individual of a species from an area to which recolonization is unlikely to occur [15]. There are various ways to eradicate the sparrow hornet recommended by the United States Department of Agriculture (USDA). For instance, we can use lethal traps during the early (pre-nesting and solitary stages) and late seasons (reproductive stage) to kill queens, or take a proactive approach to find and eliminate nests during the middle seasons (cooperative and polyethic) [16]. As a result, judgements about the effectiveness of eradication will directly determine the extent to which different methods can be applied in different areas.

4.5.1 Regression Model With Longitudinal Data

Some relevant literature reveals that invasive hornets are particularly problematic due to the adverse impact on local bee assemblages [4]. So when constructing the model to evaluate eradication effectiveness, we mainly focus on two crucial factors - changes in the population of **local bees** and **sparrow hornets** in a given area over time. According to our hypothesis, the local bee population theoretically shows **an increase followed by a decrease** during the control process. In contrast, the invasive hornet population shows the opposite trend, with **a decrease followed by an increase**. Given this, we can model a quadratic regression of density on time:

$$y_{ij} = b_{0i} + b_{1i}t_{ij} + b_{2i}t_{ij}^2 + \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n_i \tag{4.5.1}$$

where y_{ij} means the density of local bees (or sparrow hornets) of the j-th observation at the i-th site, and t_{ij} refers to the time point of the j-th observation at the i-th site; n is the number of sites where the invasive species occur, and n_i is the number of observations at the i-th site during the whole period.

Based on the above, the quadratic regression model above is rewritten using a mixed model to highlight the area's overall density data. A mixed-effects model is a statistical model containing both fixed effects and random effects [6].

$$b_{ki} = b_k + \eta_{ki}, \quad k = 0,\ 1,\ 2 \tag{4.5.2}$$

in which b_k is the **fixed effect factor** for the whole area (not related to the observation site),

Area code		A
Time span		May 20xx-Dec 20xx
Initialized score (per day)		2.48
Additional score (per day)	AS_a	May: 5.83; Jun: 7.29; Jul: 8.74; Aug: 8.74; Sep: 7.29; Oct: 5.83; Nov: 0; Dec: 0
	AS_b	May-Oct: 3.10; Nov/Dec: 0
	AS_c	3.06
	AS_d	Nov/Dec: -1.24
Report frequency (per month)		**May: 3.5; Jun: 3.9; Jul: 4.4; Aug: 4.4; Sep: 3.9; Oct: 3.5; Nov: 1; Dec: 1**

Figure 4.22: Report Frequency Assessment Form

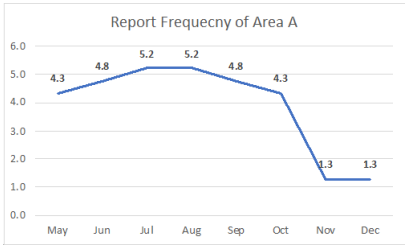


Figure 4.23: Report Frequency of Area A

while η_{ki} is the **random effect factor** varying with the observation site. The random effect factor is assumed to obey normal distribution with a mean of zero and constant variance of d_k^2, and be independent of each other.

Substituting Eq. 4.5.2 into Eq. 4.5.1 yields a regression model with longitudinal data.

$$y_{ij} = b_0 + b_1 t_{ij} + b_2 t_{ij}^2 + \eta_{0i} + \eta_{1i} t_{ij} + \eta_{2i} t_{ij}^2 + \varepsilon_{ij} \tag{4.5.3}$$

It can be rewritten in the form of the matrix.

$$Y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix}, \quad X_i = \begin{bmatrix} 1 & t_{i1} & t_{i1}^2 \\ 1 & t_{i2} & t_{i2}^2 \\ \vdots & \vdots & \vdots \\ 1 & t_{i3} & t_{i3}^2 \end{bmatrix}, \quad b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}, \quad D = \begin{bmatrix} d_1^2 & 0 & 0 \\ 0 & d_2^2 & 0 \\ 0 & 0 & d_3^2 \end{bmatrix}$$

$$V_i = X_i D X_i^T + \sigma^2 I$$

in which Y_i is assumed to obey normal distribution with a mean vector of $X_i b$ and variance matrix of V_i. The estimation of the coefficients requires the application of Maximum Likelihood Estimation (MLE), instead of Original Least Squares (OLS) for it cannot handle cases where the model is not linear. The likelihood function for Y_i is given as follows.

$$\mathcal{L}(b, D, \sigma^2) = \prod_{i=1}^{n} \{(2\pi)^{-\frac{n_i}{2}} |V_i|^{-\frac{1}{2}} \exp[-\frac{1}{2}(Y_i - X_i b)^T V_i^{-1} (Y_i - X_i b)]\}$$

Finally, by solving the maximum point of the likelihood function $\mathcal{L}$, the estimated values of b, D and σ^2 can be calculated.

The **model's significance** is that the fixed effect coefficient b_k, which reflects the overall effect in the area, can predict the time point of eradication in an average sense. In contrast, the random effect coefficient η_{ki} incorporates the variability between different sites (e.g. climate, topography, population), thus assessing the confidence interval of time point of eradication under a certain level.

Results and Analysis To determine the local bee population, we use data from the National Agricultural Statistics Services of the USA[f], which reports the number of colonies per county where the honey was harvested. To get the valid data on the spread of hornets under controlled conditions, we modify the spread model mentioned in section 4.1 using the regression model with longitudinal data. The period of the controlling process is chosen from April (queens first appear) to October (soon to be hibernating)

Next, we randomly select 20 areas and predict local bee and sparrow hornet populations' changes over time after some control measures are taken. Figure 4.24 shows the situation of local bees in three of these areas.

[f]Data Source: the National Agricultural Statistics Services of the USA(`https://quickstats.nass.usda.gov/results/98D6C754-2F7A-319C95FE-CB8F673A140E`)

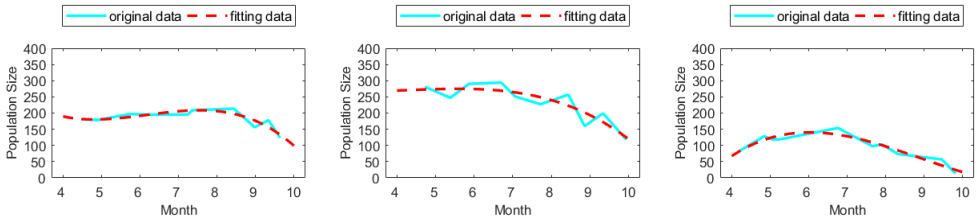


Figure 4.24: Population of local bees with human intervention

We can plot the model results (Figure 4.25) by testing the model with the hypothetical data. When $t = 6.83$, the sparrow hornet population in the 20 areas of Washington will reach the minimum point with roughly 1; when $t = 7.46$, the local bee population in these areas will reach the maximum point with approximately 241. Therefore, we can conclude that when the time range is from 6.83 to 7.46, **the number of local bees is not less than 240**. **The number of hornets is not higher than 1**, which indicates that the pest has been **eradicated in these 20 areas** of Washington State. The model can also be used for calculations for other areas of the state.

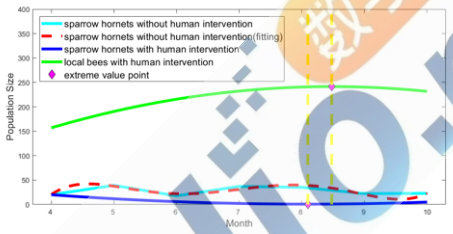


Figure 4.25: Population changes with and without human intervention

5 Sensitivity Analysis

In our model, the changes of several values may significantly impact the model's results, so we need to analyze them one by one.

5.1 Model's Sensitivity in Problem (a)

In order to analyze the sensitivity of the weights of Model 1 and Model 2, we select different values of β to calculate the precision of the Combined Forecast Model. We select each element in the set $\{0.1, 0.3, 0.6, 1, 3, 6, 10\}$ in sequence as the value of β. The boxplot is showed below.

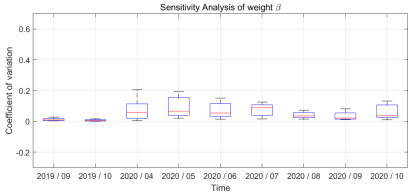


Figure 5.1: Sensitivity analysis of β

It can be concluded from the figure above that β has little effect on the prediction precision of our model, which shows that our model is robust and very reliable.

5.2 Model's Sensitivity in Problem (c)

To test the influence of the value change of the regularization coefficient λ, we pick different values of λ and predict the likelihood distribution of sparrow wasps appearing. The predicted result is shown in Figure 5.2.

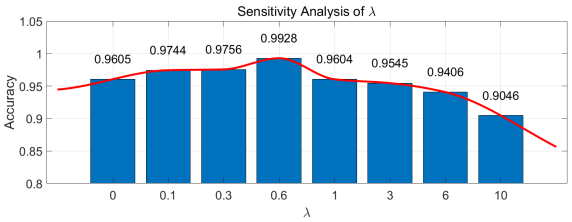


Figure 5.2: Sensitivity analysis of λ

When λ changes within a certain range, the accuracy of our model prediction is stable above 90%. When λ increases over 10, the prediction accuracy of the model drops rapidly. When $lambda$ is equal to 0.6, our model can achieve the best prediction.

6 Model Evaluation

Strengths:

- **Comprehensive consideration**: We have predicted the spread of the sparrow hornets from both temporal and spatial perspectives, and selected 16 factors to calculate the environmental suitability. Almost every useful factor has been included in Model 1. (see Section 4.1)
- **Full utilization of information**: When evaluating the likelihood distribution of sparrow hornets, we have applied the feature mapping method to expand the 2 features into 28 features, making the most of the features of limited data. (see Section 4.3)
- **Sophisticated model improvement method**: Through pseudo labelling, we are able to make use of unexploited data, which can help improve the accuracy of the original prediction model. (see Section 4.4)

Weaknesses:

- **Insufficient data volume**: The ratio of positive to negative reports is too extreme. Although we have taken some measures (e.g., flipping the images horizontally and vertically to increase the amount of data) to reduce its negative impact on the training set, we could not eliminate it completely;
- **Subjective error**: When calculating the update frequency, we apply a subjective evaluation method to assign weights to five factors , which is reasonable to some extent but not accurate enough. (see Section 4.4)

MEMORANDUM

To: The Washington State Department of Agriculture

From: Team # 2123823

Subject: Prediction of the Spread of Sparrow Hornets

Date: February 9, 2021

Dear Sir:

With the growth of trade, transportation and tourism under globalization, the introduction of sparrow hornets is now one of the greatest threats to the ecological system of Washington State. In our work, we have built a series of models in hope to help the government make full use of the data provided by the public reports to rationalize the deployment of limited resources to conduct investigations.

Analyzing the result predicted, we have many important discoveries and advice, which can help you significantly improve the efficiency of the investigation.

i) Focus on the events occurring in the areas with a high likelihood of sparrow hornet appearing, and prioritize investigation of these events. Comparing the environmental suitability between different areas, we find that there are three area centers 'friendly' to sparrow hornet: the first is the northwestern coastal region of the United States; the second is the North Cascades National Park and its vicinity on the east coast of the Rosario Strait; the third is the Kootenay National Forest in the interior. Over time, sparrow hornets will migrate to these three area centers. Then combine with the data of the current year, we can accurately predict the distribution of sparrow hornets in the next year. There are detailed data in the text above. You can determine where to deploy the elimination measures to achieve the best effects, given the limited resources of government agencies.

ii) Use image recognition and text mining methods to reduce your workload and effectively increase the investigation's accuracy. In addition, you can encourage people to take pictures or videos of sparrow hornets and upload them on the website to help you estimate the distribution of sparrow hornets. In this process, notes following the pictures or videos will improve the accuracy of your estimation. However, due to the low quality of the photos taken by the public, many of them are not suitable for image recognition, and most of the uploaded notes do not contain useful information. Therefore, you ought to appropriately promote relevant knowledge to the public to improve photos, videos and notes. For example, as to how to take photos of high quality, we suggest that the head, thorax and abdomen of sparrow hornet, which features are apparent, had better be taken in the photos. What's more, public notes should contain useful information, such as color, stripe and body size and so on.

iii) Pay more attention to events taking place in the months of summer, when sparrow hornets tend to be more active, and increase the frequency of investigation and updating data.

According to results predicted, we find that sparrow hornets are most active in July and August. During the period from November to March of the next year, sparrow hornets are almost in hibernation and appear a little, when you had better properly decrease the frequency of investigation and updating data, to avoid the waste of human resources and material resources.

iv) According to the predicted location of the migration of sparrow hornets, notify beekeepers and farmers and other stakeholders in time to avoid unnecessary economic losses. As an alien specie, sparrow hornet tends to hunt bee colonies closer to it on a large scale, causing economic losses.

As to how to eradicate the sparrow hornets, after looking up relevant information, we provide several approaches as follows.

i) Targeting Queens. The queen in an ethnic group tends to go out in April and October, and you'd better conduct more traps during this time to catch and kill the queen. Once the queen is dead, this ethnic group will disappear in a short time.

ii) Targeting Nests. You can track the sparrow hornets back to the nest, find their nest and destroy it.

iii) Public Outreach. In daily life, you need to strengthen the popularization of knowledge about pest control.

The suggestions above will help you better predict the distribution of sparrow hornets, update the data promptly and more effectively resist the invasion of alien species, protecting the local ecosystem. We sincerely hope that our study will benefit you and better protect the ecology of Washington State, USA.

Sincerely,

Team #2123823

References

[1] Adam M Baker and Daniel A Potter. Invasive paper wasp turns urban pollinator gardens into ecological traps for monarch butterfly larvae. *Scientific Reports*, 10(1):1–7, 2020.

[2] Unknown. Asian giant hornets. Retrieved from `https://extension.psu.edu/asian-giant-hornets` Accessed Unknown.

[3] Muna Maryam Azmy, Tetsuro Hosaka, and Shinya Numata. Responses of four hornet species to levels of urban greenness in nagoya city, japan: Implications for ecosystem disservices of urban green spaces. *Urban Forestry & Urban Greening*, 18:117–125, 2016.

[4] Alberto J Alaniz, Mario A Carvajal, and Pablo M Vergara. Giants are coming? Predicting the potential spread and impacts of the giant asian hornet (vespa mandarinia, hymenoptera: Vespidae) in the usa. *Pest Management Science*, 77(1):104–112, 2021.

[5] MP Robertson, N Caithness, and MH Villet. A PCA-based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and distributions*, 7(1-2):15–27, 2001.

[6] Fernanda L Schumacher, Clecio S Ferreira, Marcos O Prates, Alberto Lachos, and Victor H Lachos. A robust nonlinear mixed-effects model for covid-19 deaths data. *arXiv preprint arXiv:2007.00848*, 2020.

[7] Makoto Matsuura. Comparative biology of the five japanese species of the genus vespa (hymenoptera, vespidae). 1984.

[8] Makoto Matsuura and Shôichi F Sakagami. A bionomic sketch of the giant hornet, vespa mandarinia, a serious pest for japanese apiculture (with 12 text-figures and 5 tables). *Journal of the Faculty of Science, Hokkaido University*, 19(1):125–162, 1973.

[9] John M Drake and Robert L Richards. Estimating environmental suitability. *Ecosphere*, 9(9):e02373, 2018.

[10] Ana S Bessa, Joao Carvalho, Alberto Gomes, and Frederico Santarem. Climate and land-use drivers of invasion: predicting the expansion of vespa velutina nigrithorax into the iberian peninsula. *Insect Conservation and Diversity*, 9(1):27–37, 2016.

[11] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016.

[12] Anand Rajaraman and Jeffrey David Ullman. *Mining of massive datasets*. Cambridge University Press, 2011.

[13] Hao Wu and Saurabh Prasad. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Transactions on Image Processing*, 27(3):1259–1270, 2017.

[14] Angelica I Aviles-Rivero, Philip Sellars, Carola-Bibiane Schönlieb, and Nicolas Papadakis. Graphxcovid: Explainable deep graph diffusion pseudo-labelling for identifying covid-19 on chest x-rays. *arXiv preprint arXiv:2010.00378*, 2020.

[15] Judith H Myers, Anne Savoie, and Ed van Randen. Eradication and pest management. *Annual review of entomology*, 43(1):471–491, 1998.

[16] Small Banded Pine Weevil. New pest response guidelines. *Dec.ny.gov*, 2004.