Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2109461

Confirming the Buzz about Hornets

Summary

Vespa mandarinia has been found on Vancouver Island in British Columbia, Canada since 2019 and has gradually spread to Washington State, posing a significant threat to other bees in the area and causing significant losses to beekeepers. The Washington State Department of Agriculture has received many reports of Vespa mandarinia, but most of them have been identified as not Vespa mandarinia. ~~and~~ A as more and more reports are submitted, the identification process is becoming difficult. With bad luck, the identification of the reports that are actually Vespa mandarinia will be delayed, with extremely serious consequences. In order to solve the above mentioned problems, we analyzed and proposed a model for report-assisted identification.

分析和提出一个模型？表达不是很好

For problem 1, we analyzed and predicted the spread of Vespa mandarinia in a certain period over of time ~~in the future~~. Due to the low number of Positive reports, we roughly predicted the spread of Vespa mandarinia by analyzing ~~and predicting~~ the migration paths of other species that could be consumed by Vespa mandarinia, based on the specificity that animals would migrate to food densities. The predicted results indicated that Vespa mandarinia tended to expand outward and that most of the Vespa mandarinia would migrate southeastward. After testing, our model has good accuracy.

用现在时态

For problem 2, we first perform feature extraction. ~~First,~~ W we extracted the key features of the negative samples contained in the "Lab comments" using the TF-IDF model, and calculated the similarity between the "Notes" of each report and the extracted key features. Second, we use ResNext50 to train the images to obtain the feature similarity between each image and all the Negative images; Third, we calculate the probability of Vespa mandarinia for each report in the corresponding month; Fourth, we use the geographic location information in the report to calculate the distance from the current Vespa mandarinia spread area. mandarinia spread area. ?

After that, we construct a logistic regression model and get the optimal result after repeatedly adjusting the parameters, at which the threshold value is 0.8 and the accuracy rate can reach 95.83%. We consider the logistic regression output as the probability that a report is Negative, from which we can conclude that when the probability of a report is higher than 0.8, the report will be marked as "Negative ID", otherwise it will be marked as "Positive ID".

For problem 3, we use the logistic regression model to calculate the probability that a report is Negative of all reports marked as "Unverified" and "Unprocessed", and then ~~sort them~~ t The reports are sorted from smallest to largest. The higher the position, the more likely it is that the report is a positive sighting report.

For problem question 4, we give a method to update the model. We consider that the model needs to be updated every time a new report is identified as Positive. The update is done by adding new data to the confusing feature keyword table and Negative images and then refitting the TF-IDF model and the ResNext50 model while adjusting the probability of possible Vespa mandarinia for each month and the current area of Vespa mandarinia spread. 不通

For problem question 5, we considered that all current Vespa mandarinia were completely eliminated if we could prove that all queens had been eliminated during the year, since the common Vespa mandarinia die in winter. Evidence that the queen has been completely eradicated could be the absence of Positive reports for a longer period of time.

不通

Keywords:TF-IDF, ResNext50, Logistic Regression

Contents

1 Introduction

1.1 Problem Background

Vespa mandarinia is the largest hoopoe in the world. It is extremely toxic and can be fatal in extreme cases, and it is extremely aggressive, attacking other bees and destroying a hive in a matter of hours. For countries where Vespa mandarinia has invaded, the effective eradication of all Vespa mandarinia is a very important task.

In October 2019, a nest of Vespa mandarinia was discovered on Vancouver Island, which was quickly destroyed, but since then, several more Vespa mandarinia have been found in Washington State, USA, which is adjacent to Vancouver Island. This fact proves that Vespa mandarinia has invaded Washington State, USA. To protect the local ecology and the lives of the people, the U.S. government must take action to eradicate Vespa mandarinia with limited resources.

To make the best use of resources, the U.S.government has established a hotline and a website where people can send reports to the hotline or website to report Vespa mandarinia found in their daily lives, but because people are not professionals, they often mistakenly identify other insects as Vespa mandarinia, which requires Professionals need to sift through these reports and find those that are Vespa mandarinia.

1.2 Restatement of the Problem

The Washington State Department of Agriculture wants us to further process and analyze the reports from the public in order to extract more effective information from the reports and increase the value density of the reported data. Our data mining process should involve the following aspects.

1. Predict the spread of Vespa mandarinia over time and judge the prediction accuracy.

2. Using the data and image files provided, a classification model was built to predict the likelihood of other Vespa mandarinia being mistaken for Asian giants.

3. Using the model developed, the most likely positive sighting reports were screened for investigation.

4. Over time, new reports are added to address updates to the model and how often they are updated.

5. Using the constructed model, the evidence that constitutes the eradication of pests in Washington is analyzed.

After completing these requirements, the Washington State Department of Agriculture wanted us to make a memorandum for reporting our results to them.

2 Model Assumptions

To simplify our model, we make some key assumptions.

1. Negative samples are all predatory objects of Vespa mandarinia.

2. The estimation of the physical characteristics (size, head and body color, etc.) of the insects was accurate in each case reported by the personnel .

3. The time of report submission is evenly distributed.

4. After confirming the existence of Vespa mandarinia in a certain area, the relevant personnel will go to the area to trap the Vespa mandarinia, and a very small number of Vespa mandarinia may escape from the trapping, but it is impossible for the queen to escape.

3 Data descriptive statistics

The department provided us with report data, including an Excel spreadsheet with basic information about the mass report, image data accompanying each report, and an Excel spreadsheet with the image data corresponding to the report. We were also provided with an introductory document on Vespa mandarinia, describing the habits of Vespa mandarinia and some species that are easily confused with Vespa mandarinia. Next sections are based on these data for analysis and processing. In this section, we will present a simple descriptive statistic of these data.

3.1 Summary of feature variables

The Crowd Report Basic Information dataset contains 8 feature variables, namely:" GlobalID"," Detection Date"," Notes "," Lab Status"," Lab Comments"," Submission Date ", "Latitude", "Longitude". The "GlobalID" is a unique identifier for each report data; "Detection Date" is the time when the population found the possible Vespa mandarinia species; "Notes" is a description of the species found or submitted by the population; "Lab Status" is a description of the species found or submitted by the population. If it is indeed Vespa mandarinia, it will be marked with a Positive ID; if it is confirmed that it is not Vespa mandarinia, it will be marked with a Positive ID. If it is indeed Vespa mandarinia, it is marked as "Positive ID"; if it is confirmed that it is not Vespa mandarinia, it is marked as "Negative ID"; if it is uncertain whether it is Vespa mandarinia, it is marked as "Unverified"; If the report is being processed, it is marked as "Unprocessed". "Lab Comments" is a comment from the government agency on the report uploaded by the person, which may include the government's thanks to the person who uploaded the report and the reasons for the person's misidentification. Latitude" and "Longitude" refer to the longitude at which the report was found to be Vespa mandarinia, respectively.

The image data and report correspondence table includes 3 feature variables, which are: "FileName", "GlobalID", "FileType ". Where "FileName" is the name of each image file. GlobalID" is the corresponding report ID of each image file. "FileType" is the type of each image file, including image, zip, and video.

3.2 Analysis of "Lab Status"

There are four values of "Lab Status", and now we are counting the number of occurrences of each of these four values in the total number of times. Our data shows that there are 14 reports with the value of "Positive ID", 2069 reports with the value of "Negative ID", 2342 reports with the value of "Unverified There are 2342 reports with the value of "Unverified" and 15 reports with the value of "Unprocessed".

And reports with the value of "Positive ID" account for only 2.85% of the total reports, reports with the value of "Negative ID" account for 44.20% of the total reports, and reports with the value of "Unverified" accounts for 44.20% of the total reports. From this result, we can find that half of the reports submitted by the public cannot be effectively identified because of the information, and nearly half of the reports have been confirmed,considering the other samples as Vespa mandarinia, because the general public lacks the relevant knowledge compared to the professionals in government departments, so they cannot correctly identify Vespa mandarinia and It is not possible to provide accurate information to professionals. Therefore, there are only a very small number of reports that provide correct information about Vespa mandarinia for the relevant authorities and help them to eliminate Vespa mandarinia, but there are a large number of useless reports ("Unverified" and "Negative ID") , the government wastes a lot of resources to process these reports, so this requires us to provide a model that helps government departments to identify and process the correct reports quickly and efficiently, thus improving resource utilization.

3.3 Missing Information Statistics

One of the major reasons why the department could not determine that the sample was Vespa mandarinia was the missing information, mainly the "Notes" information and the accompanying image file information, which are very important for our subsequent model construction, so here we make a count of its missing data. The "Lab Comments" are also included in the statistics. Meanwhile, "Lab Comments" is also an important factor for the subsequent model, so we put it together and show the statistics as follows.

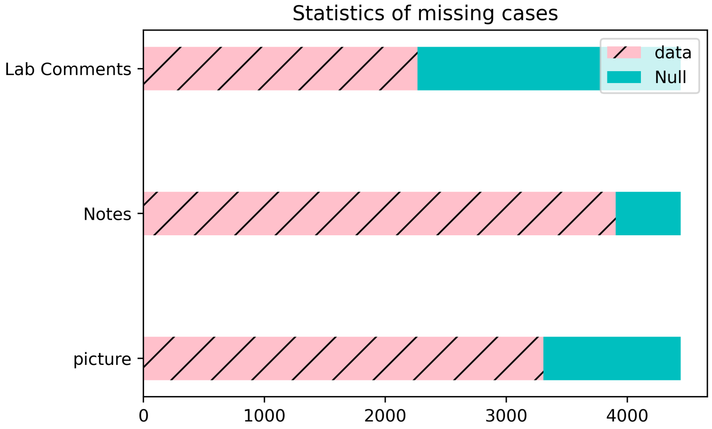


Figure 1: Missing data

The above bar chart depicts the number of valid data and vacancy values for Picture, Notes and Lab Comments, where Picture has 3306 valid data and 1135 vacancy values, Notes has 3904 valid data and 537 vacancy values, and Lab Comments has 2266 valid data and 2175 vacancy values.

3.4 Time Distribution Statistics

According to the literature, the majority of Vespa mandarinia die off during the winter and the queen of the colony goes dormant in the soil until the following spring when the fertilized queen becomes active again and gives birth to Vespa mandarinia. The literature indicates that the number of Vespa mandarinia in a colony reaches its peak in August. These documents are a good indication that the number of Vespa mandarinia is seasonally dependent. If the number of Vespa mandarinia is higher in a given month, the more likely it is that Vespa mandarinia will be found in that month, so we can conclude that Vespa mandarinia is more likely to be observed in summer and autumn (or after July). To verify our hypothesis, we counted the 14 reports identified as Vespa mandarinia by the month of discovery, and the results are shown below.

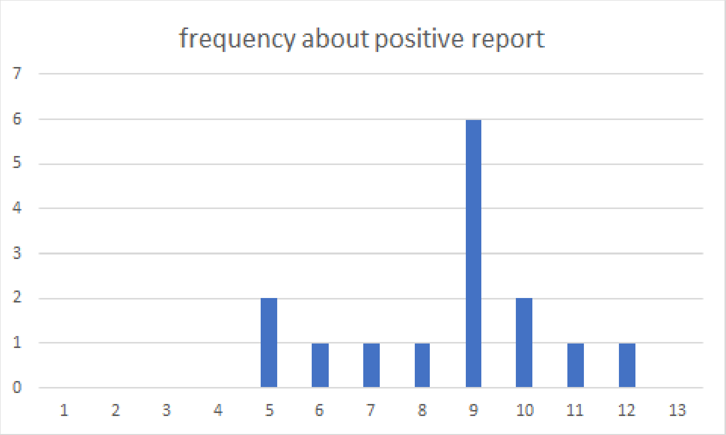


Figure 2: Mouth frequency

From the above graph we can see that no Vespa mandarinia were found in the first four months, the most Vespa mandarinia were found in September, and most of the Vespa mandarinia were found in summer and autumn (June 22 to December 21), which echoes our speculation.

4 Forecast of spread

Problem 1 asks us to predict the spread of Vespa mandarinia over time and give the accuracy of the prediction.

To analyze the trend of the spread of Vespa mandarinia, we plotted the following spread map based on the time of the 14 reports marked as "Positive ID".

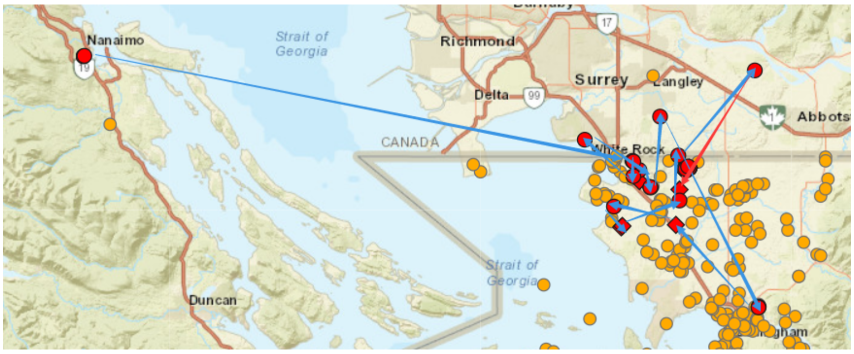


Figure 3: Map

The blue arrows in this figure indicate the spreading trend of the 14 Positive reports. From this figure, we can find that the distribution of Vespa mandarinia tends to spread from inside to outside and from the waters to the inland.

To validate our predicted trends, we wanted to be able to accurately estimate the spread of Vespa mandarinia using a time series model.

We apply reverse thinking to consider this problem, Vespa mandarinia will prey on other bee species, and according to animal nature, Vespa mandarinia will tend to migrate to food densities, i.e., the migration of other bee species will cause a linkage response of Vespa mandarinia population migration, and the migration of Vespa mandarinia has a certain lag.

Since other bee species are one of the important food for Vespa mandarinia and species belonging to the same subject are more likely to be confused, those species in the Negative report are likely to be food for Vespa mandarinia and Vespa mandarinia will migrate with them. Therefore, by predicting the migration of the location information in Negative report we can get the approximate migration path of Vespa mandarinia, as follows.

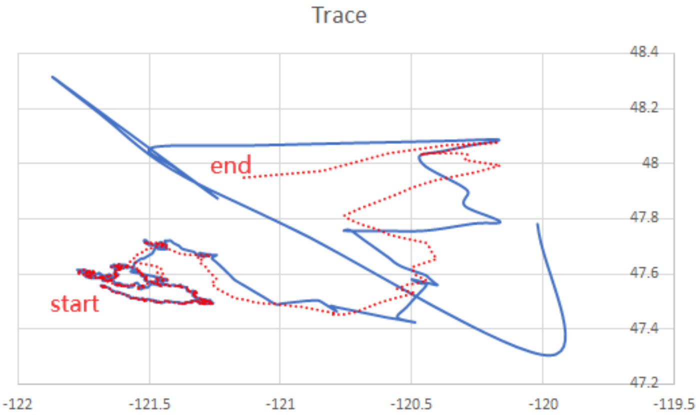


Figure 4: Trace

The blue line is the migration trajectory of the species in the Negative report, and the red line is the sliding average of the blue line. We can take the red line as the approximate migration path of Vespa mandarinia, whose migration is from the start point to the endpoint. From this figure, we can find that Vespa mandarinia gradually becomes dispersed from dense, which is consistent with our prediction, indicating that our prediction is more accurate, and according to the transformation of latitude and longitude coordinates in the figure, we can also find that Vespa mandarinia has a trend of spreading to the southeast.

In summary, we believe that Vespa mandarinia has a tendency to spread from the inside out and mainly to the southeast, and after our test, this prediction has a high accuracy rate.

5 problem two analysis

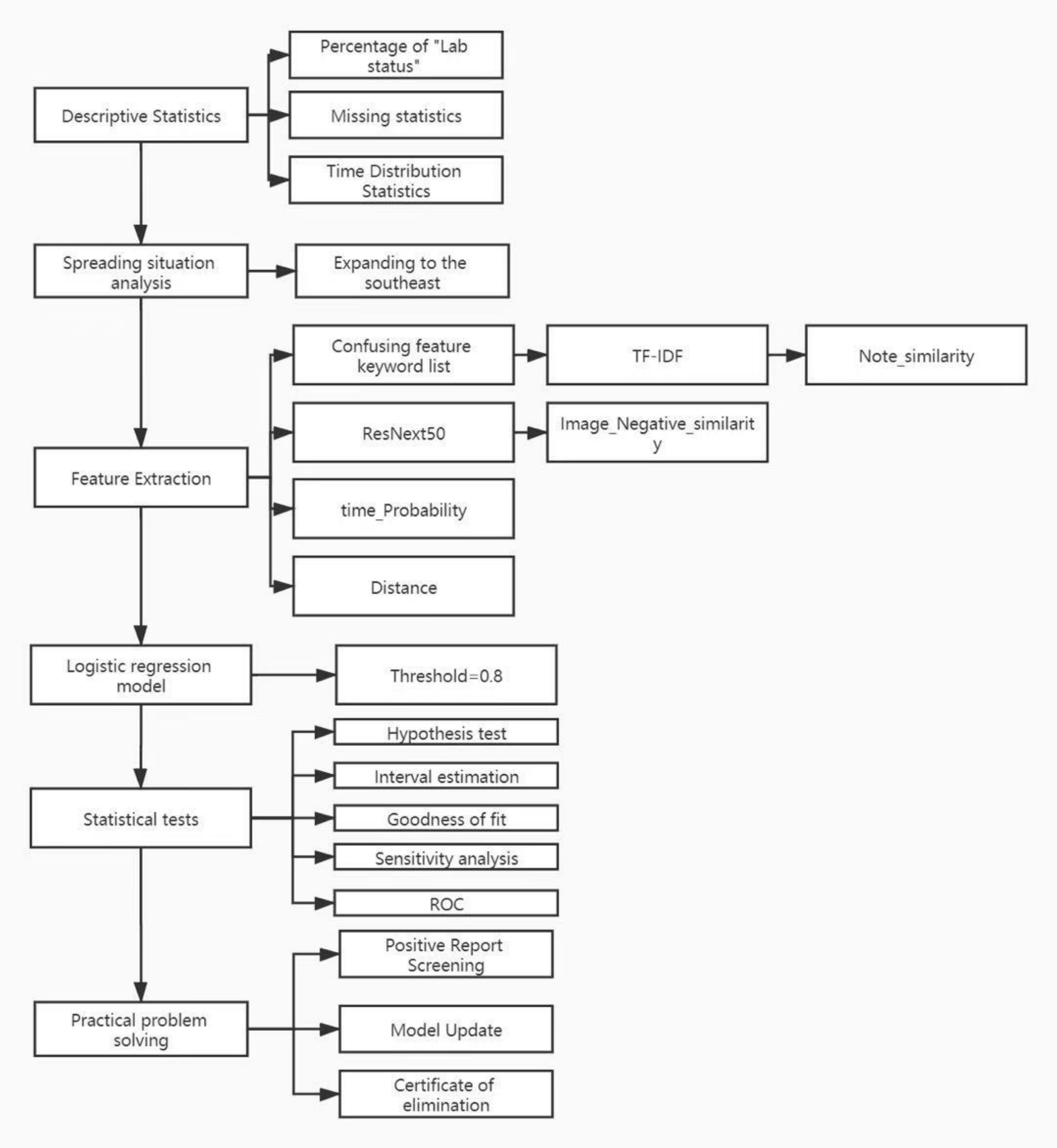
Problem 2 asked us to use the given mass report data and image files to build a mathematical model for predicting the likelihood of other insects being mistaken for Vespa mandarinia.

The misidentification in this question refers to the act of the public mistakenly identifying other insects as Vespa mandarinia and sending reports to government agencies about the insects. The misidentification is indicated by the data given, which are those reports that are judged to be "Negative ID" by professional bodies. The probability in this question is the probability that other insects are misidentified as Vespa mandarinia. When this probability is higher than a threshold, the government agency can determine that the report is misidentified and give it a "Negative ID" status; When it is below this threshold, we assume that the report is not misidentified and give the report a "Positive ID" statu.

In summary, we believe that this question is about constructing a mathematical model that is capable of deriving the probability of misidentification for each report. Then, by fitting the model, we can derive a probability threshold and characterize reports above this threshold as "Negative ID" and reports below this probability threshold as "Positive ID".

Considering that in the end we need a probability value, we choose logistic regression as our model

The flow of our logistic regression model construction is shown below.



Descriptive Statistics
Percentage of "Lab status"
Missing statistics
Time Distribution Statistics
Spreading situation analysis
Expanding to the southeast
Confusing feature keyword list
TF-IDF
Note_similarity
ResNext50
Image_Negative_similarit y
Feature Extraction
time_Probability
Distance
Logistic regression model
Threshold=0.8
Hypothesis test
Interval estimation
Statistical tests
Goodness of fit
Sensitivity analysis
ROC
Positive Report Screening
Practical problem solving
Model Update
Certificate of elimination

Figure 5: process

6 Feature Extraction

In this problem, we need to determine the probability of each report being misidentified and classify the reports as "Negative ID" or "Positive ID" according to the set threshold, so we do not care about the "Unverified" and "Unprocessed" reports. Based on this, we extracted the data marked as "Negative ID" and "Positive ID" from the whole mass report basic information table separately. There are only 14 data items marked as "Positive ID", but 2069 data items marked as "Negative ID", This data is clearly unbalanced and will have an impact on our final model.

According to the above filtering process, we obtain two new datasets, one of which contains only 14 data marked as "Positive ID" and the other one contains 2069 data marked as "Negative ID ". Our subsequent data pre-processing process processed only these 2083 data.

6.1 Feature keyword extraction for confusable species

After browsing the data, we found that several reports marked with "Negative ID" had the characteristic value "Lab Comments". A "Lab Comment" is a comment by a professional on a report submitted by the public, in which the professional gives the difference between the misidentified species and Vespa mandarinia or the reason for the misidentification.

It is well known that mass misidentification is generally due to two species being too similar in some aspects, such as size, shape, and color, and Vespa crabro is often mistaken for Vespa mandarinia. similarly, our model may also misidentify if no special distinction is made. So we need a reference set, and if the Note or image data in the mass report contains features from the reference set, then the report has a higher probability of being misidentified. Fortunately, by looking at the data, we found that the "Lab Comments" contains the features given by professionals that distinguish confusable species from Vespa mandarinia, so by extracting the keywords of these features, we can construct a table of feature keywords of confusable species, which can be used for subsequent data processing and model building. The table can be used for subsequent data processing and model building, and the detailed use of the table can be found in the next two subsections.

The process of constructing the confusable species keyword list is shown below

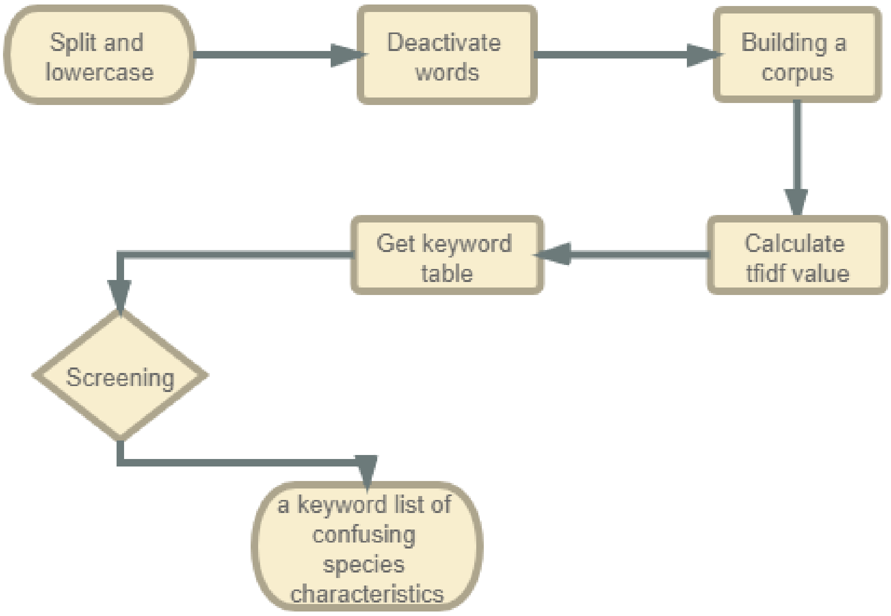


Figure 6: process2

6.1.1 Splitting words and removing stop words

First, we extracted the " Lab Comments" from the Negative report dataset by each data unit, and then divided each sentence with spaces between words to get a word-based dataset, after that, we converted all the words. After that, we store all words in lowercase and remove the deactivated words to get a new dataset.

6.1.2 Building a corpus

Since the dataset obtained after deactivating words fits the requirements perfectly, we treat the dataset obtained after deactivating words directly as a corpus.

6.1.3 Calculate TF-IDF value

The TF-IDF value is calculated for each word in each sentence. The specific calculation formula is shown below.

$$\mathrm{tf_{i,j}} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

$$\mathrm{idf_i} = \log \frac{|D|}{\left|\left\{j : t_i \in d_j\right\}\right|} \tag{2}$$

In the above equation $n_{i,j}$ is the number of occurrences of the word in document d_j, while the denominator is the sum of the occurrences of all words in document d_j. |D|: the total number of documents in the corpus.

$\left|\left\{j : t_i \in d_j\right\}\right|$:the number of files containing the word t_i and it will have a divisor of zero if the word is not in the corpus, so we use $1 + \left|\left\{j : t_i \in d_j\right\}\right|$

then we can get:

$$\mathrm{tfidf_{i,j}} = \mathrm{tf_{i,j}} \times \mathrm{idf_i} \tag{3}$$

We use the formula to calculate the TF-IDF values of all words in the corpus obtained in the third step.

First, for each sentence, we count the frequency of each word in it; after that, we count the frequency of each word in how many sentences in that corpus and the total number of sentences. Once these three data are obtained, the TF-IDF value of each word can be calculated according to equation (1)(2) above. Note that since the same word may exist in each sentence, there will be multiple TF-IDF values for the same word, and the TF-IDF value of the word is not the same in different sentences, for this category of words, we use the average of all TF-IDFs as their TF-IDF values.

After the above operations, we have obtained a dataset containing 786 words and the TF-IDF values corresponding to each word. Browsing reveals that some words do not describe well the difference between confusable species and Vespa mandarinia and should not appear in our feature keyword list. We also found that the TF-IDF values of this category of words were generally less than 0.5, so we used 0.5 as the threshold value and removed all words with TF-IDF values less than 0.5 from the obtained dataset, leaving 298 words in the end. Besides, we considered that there are six forms of verbs (verb prototype, third-person singular, past tense, present progressive, past participle) and two forms of nouns (prototype, plural) in the English writing process, so we added other forms of verbs and nouns that appeared in the original 298 words to this dataset, and the TF-IDF values of these newly added words were compared with the corresponding ones in the original dataset The TF-IDF values of these newly added words are the same as the corresponding TF-IDF values of that word in the original dataset. Finally, a feature keyword table containing 352 words is perfectly obtained, and some of the keywords and their TF-IDF values are shown below :

key	tfidf	key	tfidf
abdomen	0.8127	queen	2.8444
bees	1.4463	syrphidae	1.1569
carrion	2.4979	trichiosoma	1.0893
twigs	0.9916	white-ish	0.9917
vosnesenskii	1.3072	keyyellowjacket	3.6836

We visualize this result and get the word cloud shown below:

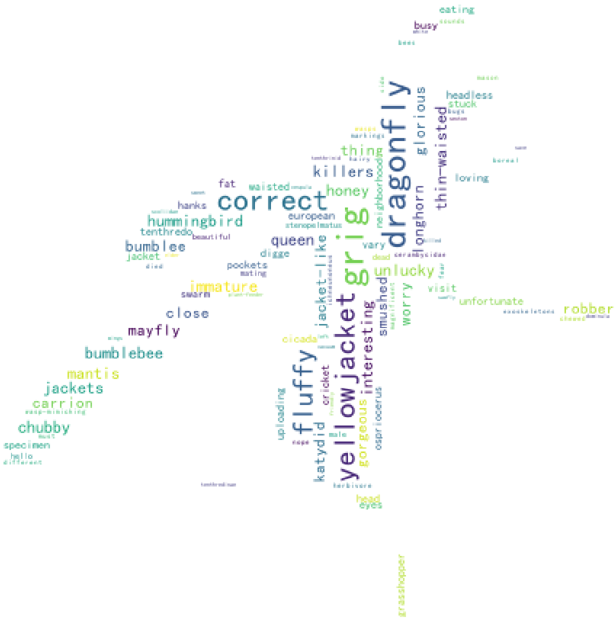


Figure 7: word cloud

6.2 Relevance of notes to feature keywords

In the previous subsection, we obtained a keyword list of confusing species characteristics, which contains 352 words and the TF-IDF value for each word. By browsing the data, we found that most of the reports have a note, which is a textual description of the report by the public. When the notes of a report have a high similarity with our keyword list of confusing species characteristics, the probability that the report is misidentified will be relatively higher. So we derived a new characteristic variable by calculating the correlation between each report and the confusable species characteristic keyword table, which we call "Notes_similarity".

We need to query the similarity of each note with confusable species feature keywords, which is a typical text similarity measure, and the most common method is to combine TF-IDF values with cosine similarity.

The cosine similarity can be measured by the cosine of the angle between two vectors, and according to the property of the cosine (1 for 0-degree angle and -1 for 180-degree angle), it is obvious that the larger the cosine value is, the more similar the two vectors are. In practice, we need to map the text data into two multidimensional vectors by some specific method, and then obtain the cosine value using the following cosine formula:

$$\text{similarity } = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}.$$

In this equation(3), A and B are two vectors, A_i and B_i are the individual components of A and B, respectively

In our practical procedure, we need to calculate the similarity of each "Notes" with the table of confusable species feature keywords, because in the previous subsection we have obtained the confusable species feature keywords and the TF-IDF value corresponding to each word, and this text data has been mapped into the vector space, so we also use the TF-IDF value to map the Notes into the vector space.

6.2.1 Notes Data Processing

First, we split each Note and convert each word to lowercase form and deactivate the word, a step consistent with subsection 6.1.1

Second, the TF-IDF value is calculated for each word in each "Notes", and the steps to calculate the TF-IDF value are also consistent with subsection 6.1.1. However, to facilitate the calculation of cosine values, we need to make the dimensions of the vectors mapped by the TF-IDF values all equal. So we construct a new corpus that contains all the confusable species feature keywords and all the words in the processed Notes, noting that each word appears only once in this corpus, and only one word is kept if there is word duplication. In our practice, this new corpus contains 429 words.

Once we have this corpus, we compare the words in each processed "Notes" with that corpus, and once we find that a word in a particular corpus does not appear in the processed results of Notes, we add that word in. It is important to note that the order of distribution of words in each vector must be the same as the order of distribution of words in the corpus so that each component of all vectors has the same fixed meaning. Since the words we add-in do not appear in Notes, we make its word frequency 0. Then the final TF-IDF value of these words is also 0, which does not affect the final cosine value. Similarly, we also checked the keyword list of confusable species features, adding the missing words and making the order of the words consistent with the order of the words in the corpus.

After this final step, we get 2082 vectors of 429 dimensions

6.2.2 Handling special data

In the descriptive statistics, we found that not all the reports have a "Notes", and for those data that do not submit Notes, the "similarity" we get is 0 according to the second step of the calculation process, which is not realistic. So we take the average of the "similarity" of all the data tagged as the same category with him as his value, for example, when a data tagged as "Positive ID" does not have the value of Notes, then the data set The average of the "similarity" of all data marked as "Positive ID" and containing "Notes" is used as the "similarity" of this data. "Similarity" of this data.

The final "similarity" of all the data was obtained perfectly, and the "similarity" of some of the data is shown in the following table.

Notes	Status	similarity
Hornet specimen sent to WSU	Positive ID	0.0637
This insect was large enough to trigger home security camera. An established bee/wasp? best that was located on the branch of a dead pine tree fell to the ground this summer as if There's a small area on my property that has overgr	Negative ID	0.2043
We caught and killed 2, 2nd on September 25th	Positive ID	0.0529
I found it in my windowsill and sucked it into my vacuum. I am worried there is a colony in my yard or on my property. I have two small kids so am very concerned.	Negative ID	0.1261
as big as a spider and exceptionally large body and head	Negative ID	0.1668

6.3 Image Correlation

Based on the given data, we can also divide the image data into two categories, one for images labeled with "Positive ID" and one for images labeled with "Negative ID". Using convolutional neural networks, we can perform image recognition to identify species features in the images labeled with "Positive ID" and species features in the images labeled with "Negative ID", respectively. With this model, whenever a new image data is submitted by the crowd, we can calculate the similarity between this image and the specific features of the image labeled as "Negative ID". The higher the similarity, the higher the probability that the report will be misclassified. Therefore, we can obtain a new feature that can be used to determine whether a report is misclassified, which we named "image similarity".

To compute "Image_similarity", we use a convolutional neural network model for image recognition.

6.3.1 Convolutional neural network model construction

Combining more complex network architectures on top of convolutional neural networks to improve accuracy, we used the ResNeXt network architecture for the Bumblebee dataset. Because the image data submitted by the crowd contains both video and images, we intercept the clearer scene of each video data, so that the image file contains only image data. After that, we adjust the aspect ratio of all the images to 3:1. Finally, according to the reported markers, the images are divided into two classes, namely Positive and Negative, and input into the volume and neural network model.

Input: Image

Output: Positive or Negative

After that, we divide 70% of the data into the training set and 30% into the test set, and then train the model and adjust the parameters to make the model have a better fit, i.e., have a smaller loss rate. During the fitting process, the variation of the loss rate of the model on the training and test sets is shown in the following figure

In the end, our model was able to achieve a loss rate of 0.638 on the test set. From the figure, we can see that the loss rate of the model is reduced a lot during the fitting process, but the final model still has a significant loss on the validation set.

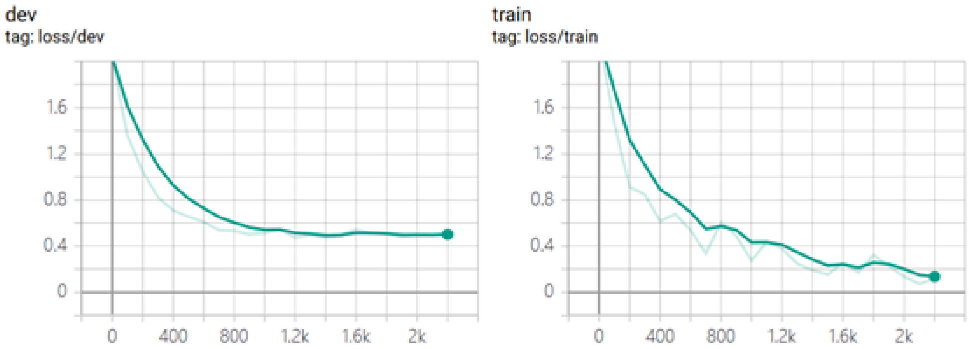


Figure 8: loss

6.3.2 Calculating similarity using convolutional neural networks

After training the model, we can feed the convolutional neural network with images. The convolutional neural network will output a two-dimensional vector (x, y), where x represents the probability that the image is labeled Positive and y represents the probability that the image is labeled Negative. Obviously the higher the value of y, the smaller the value of x, and the more likely it is that the report with this image is misidentified. Although we can directly determine whether a report is misidentified by the value of y, because of the high loss rate of our model, there is a large error if we use it directly, so we only choose to use the probability of an image being marked as Negative calculated by the model as a feature to determine whether a report is misidentified. We named this feature "Image_Negative_similarity".

6.4 Date Probability

In the descriptive statistics, we have found that Vespa mandarinia is more active from June to December, so we believe that the "Detection Date" in the mass submission report can also be used as a basis to determine whether this report is Negative. We extracted the month in the "Detection date" of each report and calculated the probability of finding a true Vespa mandarinia in that month. We use this probability as a new feature, named "time_Probability".

We processed the "Detection date" special certificate, keeping only the month information, to obtain a new column "month", and calculated the probability for each month according to the graph. For the month in which Vespa mandarinia has been detected, we use the proportion of the number of Vespa mandarinia detected in that month to the number of all Vespa mandarinia is the probability for that month. For the months in which Vespa mandarinia was not found for the time being, i.e., January, February, March, and April, so we took the mean value by season and assigned probabilities to January, February, March, and April, the probability of January and February with December, and the probability of March and April with May. At this point, all months have a probability value of finding a true Vespa mandarinia, as shown in the table below.

The characteristic variable "time_probability" is obtained by matching the probability value with the month.

month	1	2	3	4	5	6	7	8	9	10	11	12
p	0.071	0.071	0.142	0.142	0.142	0.071	0.071	0.071	0.428	0.142	0.071	0.071

6.5 Distance

We assume that for each Vespa mandarinia identified government authorities will go to the location of that Vespa mandarinia to carry out the eradication work, which is not only to get rid of that one Vespa mandarinia but also hopefully to destroy the whole colony. However, considering that Vespa mandarinia is mobile, each worker bee can have a range of no more than 8 km for feeding and other activities . Therefore, we believe that a few Vespa mandarinia in the local colony may have escaped capture. Also, Vespa mandarinia is an annual species, there is one queen in each Vespa mandarinia colony, and only the fertilized queen can give birth to other Vespa mandarinia. every winter all Vespa mandarinia except the queen die. Therefore, to simplify the calculation, we assume that the escaped Vespa mandarinia is not the queen of the colony and that the escaped Vespa mandarinia dies in that winter, i.e., after the government of the place where the presence of Vespa mandarinia is confirmed is observed to have completed the deworming, although, there may be several Although there may be a few escaped Vespa mandarinia, this Vespa mandarinia can only survive for that year, and they will die once they pass the winter and enter the spring of the following year.

When a new report is submitted, it records the location and time when the species was found by the public, and if this location is very close to the location where Vespa mandarinia was confirmed to be present during the year, then the higher the probability that the species is Vespa mandarinia and the lower the probability that the report is Negative. A new characteristic variable can be extracted, which is the distance between the location recorded in each report and the location where Vespa mandarinia was confirmed to be present in that year, and we call it "Distance". It is calculated as shown below.

We set the "Distance" of the first Positive report to 0. According to the data provided, this report was found on 2019/9/19 with a "Latitude" of 49.149394 and a "Longitude" of -123.943134. Based on this point we have the following calculation.

For reports submitted before this point in time, since there are no Positive reports, we assume that there is a high probability that these reports are Negative, so we assign a value M to the "Distance" of these reports, which is a number infinitely close to positive infinity M is a number that tends to infinity.

For reports submitted after this time, we first extracted all the locations where Vespa mandarinia was confirmed to exist before this time and within the same year according to the "Detection Date" in the report, and then calculated the distance from the location in the report to each location according to the formula for converting latitude and longitude to kilometers The distance to each location in the submitted report was calculated and the average of these distances was taken as the "Distance" of the report.

7 Logistic regression model

we construct a logistic regression model and test it statistically. First, because of the serious imbalance in the data provided by the government, we use a resampling method to make the data labeled as "Positive ID" and the data labeled as "Negative ID" closer in number. We finally obtain a balanced dataset, which contains 2069 data marked as "Negative ID" and 2058 data marked as "Positive ID" data.

"Note_similarity", "Image_Negative_similarity", "time_probability", and "distance" obtained after data preprocessing as the input of the logistic regression model, and "Lab status" of each report as the output of the model. "Distance" was used as the input of the logistic regression model, and "Lab status" of each report was used as the output of the model, thus a binary logistic regression model was constructed. Then, 70% of the data were divided into a training set and 30% of the data were divided into test set, and then the model was trained and fitted on the training set and tested on the test set.

Finally, we obtained the logistic regression model shown in the following equation

$$t\hat{a}g = \frac{1}{1+e^{-\beta}} \tag{4}$$

$$\beta = -18.152 + 1.948 \text{ time } + 20.98 \text{ 1text } + 1.67 \text{ 2image } + 71.48 \text{ lplace} \tag{5}$$

Using the above formula, we can calculate the probability of each report being misidentified, since the output value of logistic regression is a probability, between 0 and 1. We can set a threshold value that serves as a boundary for the dichotomous division. For the specific problem of Vespa mandarinia, after our repeated tuning of the references, we finally determined that the threshold value was set to 0.8, at which point the logistic regression fit this dichotomous problem best with an accuracy of 0.9583. The probability of misclassification for each report is shown in the following figure.

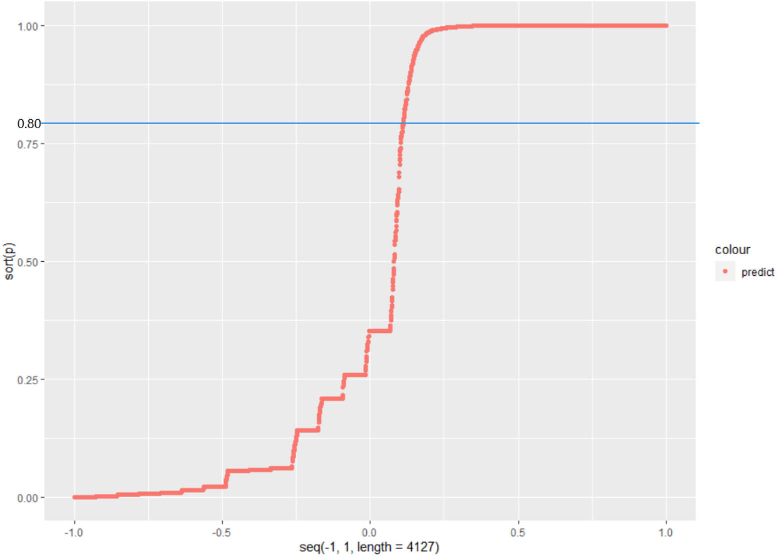


Figure 9: Probability

By looking at the graph, we can see that the data are more evenly distributed between the upper and lower parts of the horizontal line of 0.8, which is consistent with the results of our model.

At this point, we can use the model to efficiently determine whether a report is misidentified or not. If the probability of the report is higher than the threshold of 0.8, then the report is misidentified and will be marked as "Negative ID"; if the probability of the report is less than the threshold of 0.8, then the report is correct and will be marked as "Positive ID". Once a report is flagged as "Positive ID", the government authorities must take it seriously.

8 Statistical tests of logistic regression models

Because logistic regression is a statistical model, we need to perform a series of statistical tests on it, the results of which will be shown in this section.

8.1 Hypothesis testing

Hypothesis testing is the process of constructing an original hypothesis H_0, assuming that the original hypothesis is true, and then seeing if the data provide evidence against H_0. In logistic regression problems, academics generally construct z-statistics to test H_0.

In this section we want to test whether each characteristic variable has an effect on the explanatory variable "Lab status" and if so, then the coefficient of the characteristic variable from the logistic regression is not 0. So we make the following assumptions.

$$H_0 : \beta_i = 0$$

As long as we can prove that the original hypothesis does not hold, then it is influential and can be retained in our model.

Under the original hypothesis, the value of the z-statistic is the regression coefficient divided by its standard error, and the formula is shown below.

$$Z = \frac{\hat{\beta} - 0}{\hat{\sigma}_{\hat{\beta}}} = \hat{\beta} / \hat{\sigma}_{\hat{\beta}} \tag{6}$$

When setting the significance level at 0.05, empirically, if the absolute value of Z is greater than 2.0, the characteristic variable has a significant effect on the explanatory variables.

Using the formula, we can obtain the z-values for each of the characteristic variables, and the results are shown in the following table.

	z
time_probability	10.96
Notes_similarity	10.88
Image_similarity	8.16
distinction	20.42
_cons	-22.52

Each z-value is well over 2.0 in absolute value, so these four variables have a significant impact on the prediction results of Lab Status, and each of them needs to be retained in the model.

8.2 Interval estimation

Interval estimation is based on point estimation and gives an interval range for the overall parameter estimate of the coefficients, which is usually obtained by adding or subtracting the estimation error from the sample statistic. In statistics, this interval is also called a confidence interval.In our model, the 95% confidence intervals for the coefficients of each variable are shown in the following table.

	[Confidence interval of 0.95]
time_probability	[1.599367,2.296145]
Notes_similarity	[17.2001,24.7609]
Image_similarity	[1.27031,2.073816]
distinction	[64.61973,78.34186]
_cons	[-19.73215,-16.57282]

8.3 Goodness of fit

In a general linear regression model, there is an indicator R^2, this indicator is known as Pseudo R^2, this indicator can be defined in various ways, the most common one is McFadden's R^2 (likelihood ratio test), whose formula is shown below. In a general linear regression model, there is an indicator This indicator is known as, this indicator can be defined in various ways, the most common one is McFadden's R^2 (likelihood ratio test), whose formula is shown below.

$$\text{McFadden's R}^2 = 1 - \frac{\ln \mathrm{L}}{\ln \mathrm{L}_0} \tag{7}$$

where L_0 refers to the likelihood value of the model that contains only the constant term, and L refers to the likelihood value of the model containing all explanatory variables as well as the constant term.

Usually, the model is considered to have a good fit if the value of the McFadden's R^2 is greater than 0.5.

In addition to Pseudo R^2 ,logistic regression models can also use the information criterion to assess the fit of the model, and the commonly used information criteria are *AIC* and *BIC*

AIC guidelines are defined as

$$AIC = \frac{-2\ln \hat{L}(M_k) + 2P}{N} \tag{8}$$

$\hat{L}(M_k)$ is the likelihood value of the model, and P is the number of parameters in the model (including the constant term). The smaller the value of *AIC*, the better the model fit, all other things being equal.

The *BIC* guidelines are defined as follows

$$BIC = -2\ln L(M_{Full}) - (N-P)\ln N \tag{9}$$

Again, a smaller *BIC* value indicates a better fit of the corresponding model.

The results of these three definitions in the outside test are calculated as shown in the following table.

McFadden's R2	0.781
AIC	0.306
BIC	-33064.595

We can see that the model fits well.

8.4 ROC curve analysis

The *ROC* curve graphically combines sensitivity and specificity. The more convex the *ROC* curve is,the closer to the upper left corner, the more valuable the model is and the better the model is. Our *ROC* curve is shown in the following figure.

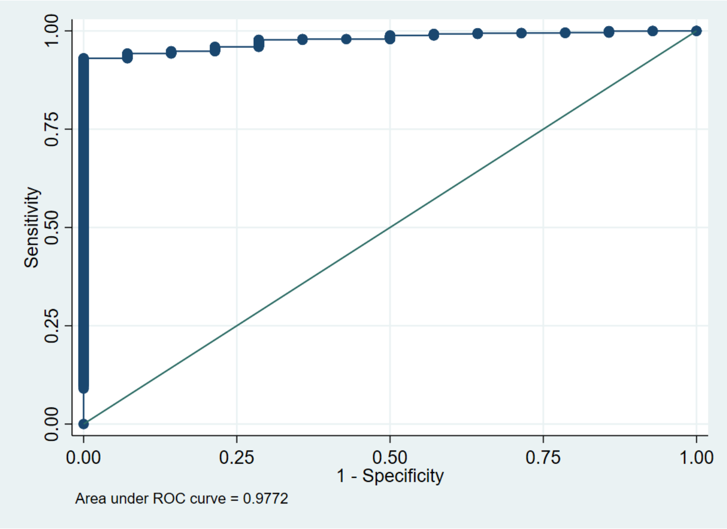


Figure 10: ROC

It can be seen that the *ROC* curve is very convex, indicating that our model works well.

9 Positive eyewitness report screening

Question 3 asked us to use the model developed to screen the most likely positive sighting reports for investigation.

The logistic regression model we built in Problem 2 can derive the probability that a report is misidentified based on the four input features "Note_similarity", "Image_Negative_similarity", "time_probability" and "Distance". probability" and "Distance", when the probability is higher than 0.8, the report is misidentified and will be marked with "Negative ID "Otherwise, the report is correct and will be marked with 'Positive ID'. Obviously, the higher the output probability of the logistic regression model, the more likely the report is misidentified; the lower the output probability, the higher the correctness of the report, which means the more likely the report is about the presence of Vespa mandarinia, and the local government needs to carry out deworming work.

So, to solve problem 3, we can calculate the probability of misidentification for each submitted report, and then proceed to arrange these reports in order from smallest to largest, with the more advanced data being more likely to be positive sighting reports.

To verify the effectiveness of our screening method, we calculated the probability of misclassification for the 2083 reports that had been marked by professionals and found that the 14 reports marked with "Positive ID" were all at the top, and some of the reports marked with "Positive The probabilities of the reports marked with "Positive ID" are shown in the following table.

Global ID	p
{5AC8034E-5B46-4294-85F0-5B13117EBEFE}	0.0021
{5EAD3364-2CA7-4A39-9A53-7F9DCF5D2041}	0.5151
{124B9BFA-7F7B-4B8E-8A56-42E067F0F72E}	0.6188
{AD56E8D0-CC43-45B5-B042-94D1712322B9}	0.6828
{F1864CC3-508C-4E60-9098-B158AB413B03}	0.7124

This result strongly validates our filtering method, which we then applied to the reports marked as "Unverified" and "Unprocessed".

Global ID	p
{BBBA5BA0-CAFB-43D3-8F1D-FB2D9CF777E0}	0.4586
{EAC16248-52B7-4B86-8BA3-0913E2EE3A3A}	0.5876
{A0161ABC-0636-445B-A877-BBB2CA55EC8F}	0.5976
{22E3A08D-494C-4539-8894-FDC32F2C9855}	0.6769
{6B29A36C-EE9D-4188-849D-79E6A5A3D161}	0.7764

10 Model Updates

Question 4 asks us to provide a way to update our model over time and to give the frequency of updating the model.

In constructing the logistic regression model we give four input features to the model: "Notes_similarity", "Image_Negative_similarity ", "time_probability", "Distance". and enables the model to output the probability of a report that is Negative. Since there are constantly people submitting their reports, our model will need to be updated based on the reports submitted by the people. Since the ultimate goal of our modeling is to identify Positive reports quickly and efficiently, and since the vast majority of reports submitted by people are Negative and only a very small number of reports are Positive, our model updates should be determined by the new reports detected as "Positive ID " reports. After our analysis, we conclude that the model will be updated whenever a new report marked as "Positive ID" is detected. The model is updated in the following way.

It is optional whether to include image files or to give "Notes" in the new reports submitted by the crowd, but according to the results of descriptive statistics, most of the reports will contain both. Therefore, as the number of detected Negative reports increases, the list of confusing feature keywords based on the "Lab comments" feature variable in the Negative reports is expanded and the number of Negative images increases. Therefore, the sub-models for computing "Notessimilarity" and "Imagesimilarity" - TF-IDF model and convolutional Neural network models are updated.

In addition, the reports submitted by the crowd generally contain the "Detection Date", "Latitude" and "Longitude". three characteristic variables. These three variables are used to generate the other two variables "time_probability" and "distance" that are input to the logistic regression model. However, the calculation of these two models does not change with the addition of the Negative report. "Timeprobability" requires calculating the probability of Vespa mandarinia for each month, and "Distance" requires finding the location of Positive reports in the same year. The location of the report, so the calculation of these two characteristic variables will change when a report marked as "Positive ID" is added.

Because from the results of sensitivity analysis, "time_probability" and "distance" are not stable compared with the other two characteristics, and small changes in the data cause large changes in these two characteristics in the logistic regression model. The small changes in the data will cause large changes in these two characteristics in the logistic regression model, but will not cause large changes in "Notessimilarity" and "Imagesimilarity". Also, considering that the masses will submit Negative reports frequently, each update will be a serious waste of resources, so we believe that only when PositiveID is detected, we must adjust the model parameters and update the model for the accuracy of subsequent predictions. The "time_probability" and "distance" are calculated by adding the information of the new Positive report into the calculation. At the same time, the confusing keyword table and the Negative image data are updated, and the "Lab comments" data of the Negative report and the Negative image data are updated during the time interval between the two model updates. The regression coefficients of the logistic regression are updated.

11 Extinction Evidence

Question 5 asks us to use our own model to give a proof that Vespa mandarinia has been completely eradicated.

Based on the literature, we know that only fertilized queens can give birth to other Vespa mandarinia and that by winter all Vespa mandarinia except the queen dies. Therefore, we believe that the ideal evidence for the eradication of Vespa mandarinia is that no more queens will give birth in the following year or that all queens have been eliminated in the current year.

Our model can predict the probability that the report is misidentified, and thus efficiently extract each Positive report, each Positive report corresponds to exactly a one-time point of model update. Assuming that the last update time is T_N, the Vespa mandarinia in the corresponding Positive report is the last observed Vespa mandarinia until the current time, which may be a queen bee or a common Vespa mandarinia. The interval between T_i is $\triangle T_1$, $\triangle T_2$, $\triangle T_3$,.... ....,$\triangle T_{N-1}$.

When $\lim_{T_N \to T_i}(1/\triangle T) = 0$,the current time node $T_N$tends to a particular time node T_m when the convergence to 0 can be considered as the Vespa mandarinia have all been eliminated. Another explanation is that at a certain time node If the time interval for finding a new Vespa mandarinia is positive infinity, then the time interval for finding a new queen must also be positive infinity, which means that all queens in the current year have been eliminated.

12 Model advantages and disadvantages

Model advantages:

- Our model makes integrated use of temporal, spatial, textual and graphical data, and the multivoting mechanism increases the robustness of the model.
- Our model is scalable, easy to implement on new datasets, and easy to update.

Model disadvantages:

- Our model may have ignored some other variables that have important effects on the distribution of Vespa mandarinia, such as temperature.

13 Memorandum

In response to the five issues raised by the relevant units, we make the following summary.

Address and discuss whether or not the spread of this pest over time can be predicted, and with what level of precision

Our proposed model is based on the submitted report information and combines it with very small information about the actual spread path to make predictions. The distribution of bumblebees can be seen on a map using known data drawn on all tagged reports from 2019 to date. Limited by the abundance of data, we simulated the actual bumblebee population using the few known positive samples, thus compensating for the lack of data, and combined with the location time data to derive the general trend of Vespa mandarinia population spread. The model indicates that Vespa mandarinia is gradually spreading into southeastern Washington and that residents need to be warned.

Most reported sightings mistake other hornets for the Vespa mandarinia. Use only the data set file provided, and (possibly) the image files provided, to create, analyze, and discuss a model that predicts the likelihood of a mistaken classification. image files provided, to create, analyze, and discuss a model that predicts the likelihood of a mistaken classification.

Most of the reports are judged to be negativeID. To solve this annoyance, the goal of our model is to prioritize the most likely PositiveID from the many reports submitted for expert verification, saving most of the time lost to negativeID reports. Therefore, in order to improve the prediction accuracy, we analyzed the provided data specifically, including text data, image data, time data, and geographic location data. Since the data has a serious imbalance problem, these four features are not directly utilized, but sub-models are added to extract features before the final model. The text data features were processed using the tf-idf model, the image data used the resNext50 model, and the temporal and geographic location data were processed as frequency data, respectively, for spatio-temporal similarity to the negativeID sample. The output of these four sub-models is considered as a copy of the number of votes belonging to the negativeID and the positiveID, and the predicted labels are obtained by using this vote as the input of logistic regression

Use your model to discuss how your classification analyses leads to prioritizing investigation of the reports most likely to be positive sightings.

Use your model to discuss how your classification analyses leads to prioritizing investigation of the reports most likely to be positive sightings.The logistic regression model will output the probability of belonging to negativeID. After some adjustment of the parameters, the model can predict all the samples belonging to positiveID in the dataset, and on this basis make the same prediction for the unprocessed and unverfied samples, the model gives the following two samples that are also likely to be Vespa mandarinia.

$9A5CB940-8951-4FE7-8619-DAF4A8FE1850$

$41EB9554-6676-4CD1-BE5B-5133CB3EE4B0$

Address how you could update your model given additional new reports over time, and how often the updates should occur.

Features include

timeProbability: the month in which the report is located, and the percentage of months in which the negativeID appears in the historical data.

Distance: the maximum value of the distance between the geographic location of the report and all the current positiveIDs, and then processed for probability.

Notesimilarity: the similarity of the lexicon between report notes and all the current negativeIDs.

ImageNegativesimilarity: the similarity between the report image and all the current negativeID images.

Since most of the submitted reports are "negative ID", the frequency of updating the model is determined by the detected "positive ID", i.e., the model is updated when a new "positive ID", is detected. The method of updating the model is as follows.

The new report text and image are optional, but most will include both. Time and space are necessarily included and are the data fields that are optimal for the model in scenarios where future conditions are predicted based on historical data. As the number of detected "negative ID" increases, the dictionary for text similarity calculation based on "negative ID" is expanded and the negative material of the image is increased. However, we set the time point of model update at the next time when PositiveID is detected, because the detection of PositiveID causes changes in temporal features and spatial features, and the parameters of these two features, from the sensitivity analysis, are not very stable, then the model parameters need to be adjusted for the accuracy of subsequent prediction when PositiveID is detected. In other words, the time-frequency and spatial features need to add the new PositiveID's relevant time and geographic location to the frequency calculation and expand the data of the other two features: the text material and image material of the "negative ID" accumulated during the time interval between two model updates. The regression coefficients of the logistic regression are then updated.

Using your model, what would constitute evidence that the pest has been eradicated in Washington State?

The ideal situation for the pest to be eradicated is that no queen will lay eggs in the following year, or that all queens have been eradicated in the current year since other hornets will not survive more than one year and are not capable of reproducing. The combination of temporal plus spatial features can locate the queen bee's presence, which exactly corresponds to the time point of the model update.

Judgments can be made by establishing test statistics based on the distribution of the actual model update time node sequence.

Finally, for the model to work better, the model assumptions have to be kept in a state that is as satisfying as possible, otherwise, the model may fail.

References

[1] *Lian, Yujun. Logit model STATA [EB/OL].2018-10-06.*

[2] *Archer, M E . 1995. Taxonomy, distribution and nesting biology of the Vespa mandarinia group (Hym., Vespinae). Entomol. Mon. Mag. 131: 47–53.*

[3] *Yanagawa, Y, KMorita, TSugiura, and YOkada. 2007. Cutaneous hemorrhage or necrosis findings after Vespa mandarinia (wasp) stings may predict the occurrence of multiple organ injury: a case report and review of literature. Clin. Toxicol. (Phila). 45: 803–807.*