Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2112243

Denoising the Buzz: A Report Analysis Model for Vespa mandarinia Reinforced by Multimodal Data

Summary

Since the first discovery of Vespa mandarinia in September 2019, Washington State has witnessed several confirmed sightings along with a multitude of mistaken reports from the public, which has exerted a huge burden on government agencies of Washington State. Therefore, it is high time an effective solution should be developed to process the data from public detection reports and reduce the consumption of resources for additional investigation.

To tackle the challenge mentioned above, we propose a multimodal modeling method with full utilization of the data given, including temporal-spatial features as well as the textual and visual data from public reports. Its output scores demonstrate the likelihood of mistaken reports or credible ones, which can help researchers better deal with the tough task.

First, 1. As for temporal-spatial features, we combine the use of grey system model, cellular automaton, and temporal-spatial clustering, to analyze and predict the spread properties and trajectories of Vespa mandarinia; 2. In terms of textual data, the LDA Topic model and naive Bayes algorithm are applied to notes and lab comments, to extract the words of strong correlation with high confidence, detailed description or mistaken possibility; 3. As to visual data, We build a CNN network for image classification and train it on images re-annotated by labels extracted from lab comments. The network can infer the confidence score of submitted photos or videos.

Then, we integrate our multimodal model after normalizing and summarizing the output scores of each component. Evaluation is conducted on the original data set, which eventually implies the strong correlation between the results of out model and the report priorities, and thus the accuracy and helpfulness of the model proposed is proved.

Next, we discuss the situation for updating the model by introducing the concept of model entropy, to represent the influencing factors of the unverified recently-inputted data. The conclusion is that data updates and network retraining are required once the entropy is above the designated threshold,

Finally, we decide the active threshold of Vespa mandarinia sighting reports based on the former analysis. and any report below is considered negative. With the living properties and data analysis taken into consideration, once no more recent report exceeds the threshold over the designated period, it can be regarded that Vespa mandarinia has been eradicated in Washington state.

Keywords: Multimodal modeling; Temporal-Spatial Clustering; the LDA Topic Model; Naive Bayes Method; Image Classification

Contents

1 Introduction

1.1 Background

Vespa mandarinia, also known as Asian giant hornet, is an invasive species of hornet as well as a recently invaded alien species to Washington state. They are harmful to local beneficial insects, *e.g.* European honeybees, breaking local ecosystem and thus is regarded as agricultural pests.

Due to its potentially severe impact on the local ecological environment, the government agencies have adopted a series of measures as the responses to the challenge. However, problems have been aroused that the massive amount of reports mistaking another type of insects for Vespa mandarinia consumed lots of resources for investigation. Therefore, it's essential for government agencies to adopt measures to address the challenge.

1.2 Problem Restatement

The primary questions for this problem are **how to interpret the data from public reports** and **how to prioritize public reports for additional investigation**. More specifically, we need to solve the following problems:

- **Problem 1**: Disscuss the spread of Vespa mandarinia over time, and the precision of prediction.
- **Problem 2**: Propose and analyze a multimodal model combining temporal-spatial, textual and visual features from public reports, to predict the likelihood of a mistaken classification.
- **Problem 3**: Prioritize reports with higher likelihood to be positive sightings for additional investigation based on the model proposed.
- **Problem 4**: Discuss the conditions and methods for updating the model proposed, especially from the perspective of time and frequency.
- **Problem 5**: Discuss the conditions for confirming the eradication of Vespa mandariniain Washington State.

1.3 Our Work

Our work follows the following workflow framework, as shown in the Fig. 1. The multimodal modeling method utilizes the raw data from different perspectives, including temporal-spatial clustering as well as textual and visual modeling. It then aggregates the features and scores from all three aspects and outputs the final confidence score indicating the likelihood of mistaken reports or credible ones. We call our model START for its four essential parts.

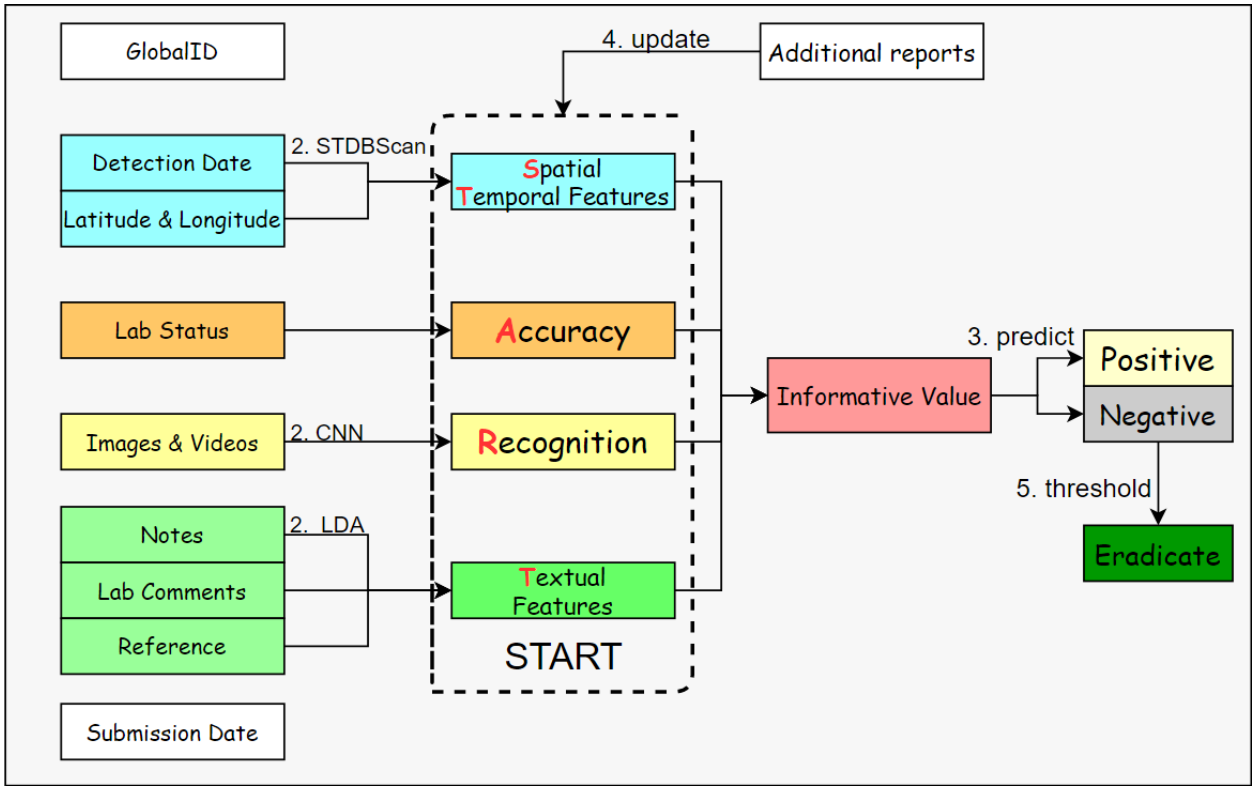


Figure 1. Workflow framework.

- In problem 1, we first conduct a preliminary analysis to portrait the overall characteristics of the raw data set. Then based on positive reports, we discuss how to predict the spread of Vespa mandarinia in both number growth and geographical distribution points, using grey system model and cellular automaton respectively.
- In problem 2, we propose a multimodal model composed of three parts to predict the likelihood of mistaken reports based on the quantitative scores from each part.

    1. From the perspective of temporal-spatial data, ST-DBScan, a temporal-spatial clustering algorithm is applied to model the data of detection date and sighting location, which helps us to analyze the spread trajectories of Vespa mandarinia, and then score the other reports according to their deviation with the estimated spread trajectory.
    2. In terms of textual data, We apply the LDA Topic model and naive Bayes algorithm to notes and lab comments, to extract the words of strong correlation with high confidence, detailed description or mistaken possibility, and thus scoring the report notes according to their confidence and helpfulness.
       Besides, we also extract the ground truth label of negative reports for the following image classification tasks, since the number of each class in the original data set is of severe imbalance.
    3. As for visual data, with the help of lab comments, we re-split the images into 10 categories including Vespa mandarinia and other species of high possibility to be

mistaken for. We then build a CNN network and train it on the re-annotated images, which can infer the confidence score of submitted photos or videos.

- In problem 3, we integrate our multimodal model and apply it to the original data set. We normalize and then sum up the scores so that reports with higher scores are more likely to be positive ones. Visualization of results turns out that all positive labeled reports are of high ranks and proves the accuracy and reliability of our model.
- In problem 4, we discuss this issue by the situation of lab status. We introduce the concept of model entropy to represent the influencing factors of the unverified new data. If the entropy is above the threshold, data updates and network retraining are required.
- In problem 5, based on the ranking result from problem 3, we regard the score of the last positive report with a margin as the active threshold, and any report below is considered negative. If there is no recent report that exceeds the threshold over the period, we considered Vespa mandarinia eradicated in Washington state. The life cycle of Vespa mandarinia is also taken into consideration.

2 Problem 1: Prediction and Precision

In this section, we preprocess the data set and analyze the spread of Vespa mandarinia over time.

2.1 Data Cleaning

We first remove those records whose detection time is earlier than 2019 because the Vespa mandarinia was discovered in September 2019. The submission date is mainly decided by the faculty in WSDA and useless for our subsequent analysis so we just simply delete this column. In the end, we eliminate 72 unqualified items and leave 14, 2055, 2286 and 2055 items for positive ID, negative ID, unverified and unprocessed data, respectively.

2.2 Overall Data Characteristics

We have plotted the number of public reports per month in Fig. 2 and find that it peak around September and October, which matches the life cycle of Vespa mandarinia[1] and also shows that they do exist in the wild. And since Vespa mandarinia hibernates in winter, positive report number drops in November and December, then decreasing to and remaining zero in the first few months of a year. In terms of the number of negative public reports, its highest value, unlike that of positive reports, appears in August, as Fig. 2 has shown. Therefore, the number of other insects is generally higher in August, that is, there is a certain difference between the life cycle of these native other insects and Vespa mandarinia.

From a geographical point of view shown in Fig. 3, Vespa mandarinias do not spread far. By examining longitude and latitude, positives reports form several obvious clusters and distances within each cluster are no greater than $\pm 0.5^\circ$ in longitude(around 31 miles) and $\pm 0.2^\circ$ in latitude(around 12 miles).

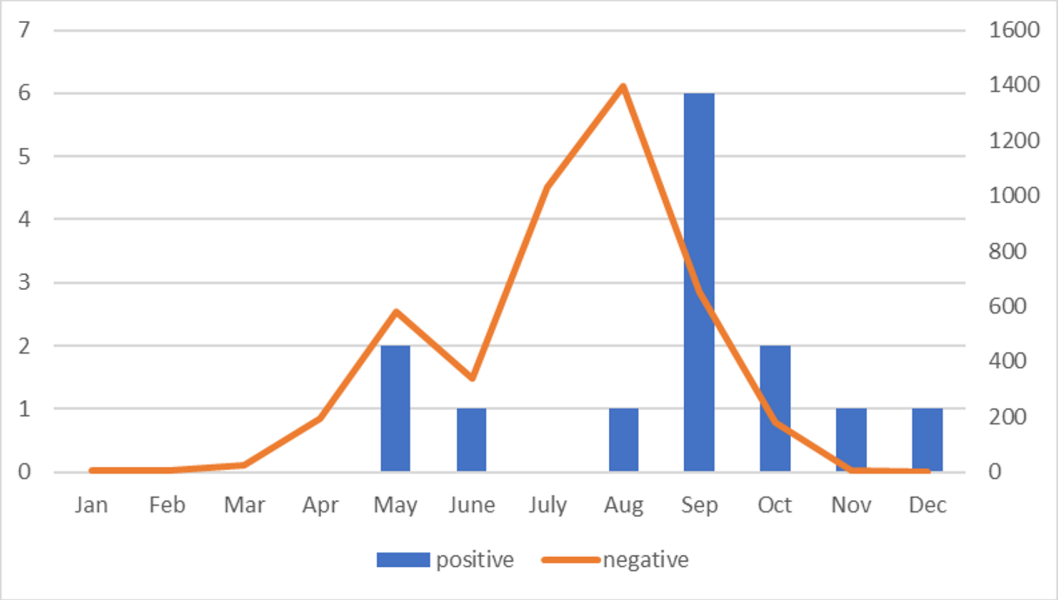
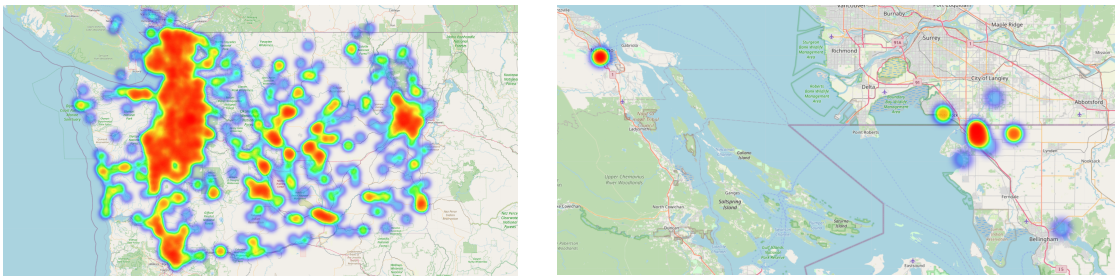


Figure 2. Time distribution (per month) of positive and negative public reports, positive reports in blue and corresponding to the left axis, negative ones in orange and corresponding to the right axis.



(a) (b)

Figure 3. (a) Geographical distribution of all public reports, warmer color meaning higher report density and (b) geographical distribution of positive public reports, warmer color meaning report appears later in the year.

2.3 Prediction and Precision

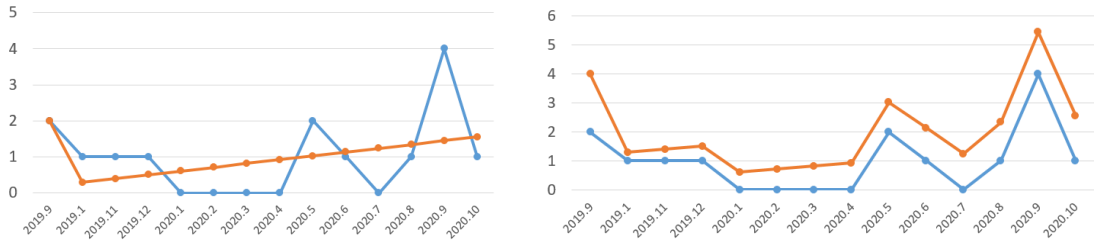
According to the insights we have gained from the sections above, we believe that the spread of Vespa mandarinia can be predicted in the following two aspects:

- **Number growth**: How the number of this pest in a given area would change over time.
- **Geographical distribution**: Which regions are potentially invadable and most likely to succumb.

2.3.1 Prediction on number growth

Although we have already picked out every positive report from the data set, only 14 Vespa mandarinia occurrences can be used to develop our prediction model. Furthermore, occurrence

number per month, shown as the blue lines in Fig. 4, appears to be in a relatively complex pattern that we find it hard to achieve a good performance with traditional algorithms like regression. Therefore, we adopted grey series forecasting methods from the grey system theory to better deal with problems with small amount of data and complicated internal laws[2]. More specifically, we have applied both GM(1,1) and GM(2,1) grey differential model on these 14 Vespa mandarinia occurrences and trained the two models to predict occurrence number every month. The results are shown as orange lines in Fig. 4.



(a) GM(1,1) model prediction results.

(b) GM(2,1) model prediction results.

Figure 4. Comparing GM(1,1) and GM(2,1) model predictions on the number of Vespa mandarinia occurrences per month, blue lines are from the original data set and orange lines are model predictions.

In Fig. 4a, our forecast results deviate greatly from the real data. That is because the actual occurrence series constantly floats up and down, while GM(1,1) as a first order model is suitable for sequences with strong exponential law and can only describe monotonous changes[3]. On the other hand, the second order model GM(2,1) yields predictions with much more acceptable errors, but it is more complicated and harder to implement.

2.3.2 Prediction on geographical distribution

Results in Fig. 3b showed that all Vespa mandarinia occurrences indicated by positive reports lay very close to each other, not even one degree away in longitude and latitude. For those within the same cluster, calculated straight distances between Vespa mandarinia occurrences are no greater than 34 miles. At further examination, it turns out that the 14 occurrences are centralized not only spatially but also temporally: they all occurred within only one year, where it's Vespa mandarinia's annually biological life cycle[1] that dominates their activities and would introduce errors. To address this problem, we searched for some related work and the following is what we have found:

In 2020, Alaniz *et al*. [4] predicted that in the western coast of the USA, an environmentally suitable zone, among many others, for Vespa mandarinia should appear in the states of Washington, Oregon, and in northern California. They applied ecological niche modeling together with a maximum entropy algorithm, combining 25 environmental metrics and spatial occurrences of the target species to estimate the native habitat suitability. After that, native habitat suitability is projected to the USA to find regions with similar environmental characteristics, which are then marked as potential invadable areas. Ecological niche modeling predictions on the spread of invasive species after their recent arrival usually produce a good level of accuracy[5, 6], and the results do match figure Fig. 3a quite well. Nuñez-Penichet *et al*. [7] obtained similar results through ecological niche modeling as well, but they further conducted several dispersal simulations using cellular automaton algorithm to verify those results and obtained good consistency.

Therefore, we believe that although Vespa mandarinia's short-term geographical distribution is hard to predict, it can still be predicted with an acceptable accuracy in the long term.

3 Problem 2: Multimodal Data Processing

3.1 Spatial Temporal Features

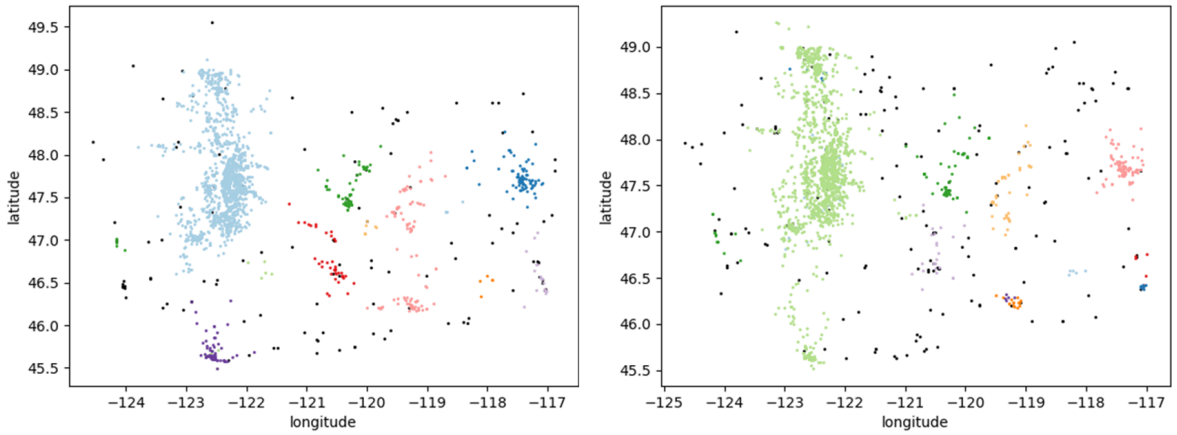
ST-DBScan[8] is a efficient density-based clustering algorithm, which has the ability of discovering clusters according to spatial and temporal values of the objects.

3.1.1 Setup of spatial temporal features

We apply the ST-DBScan model to positive, negative, unverified, and unprocessed reports, respectively. The temporal spatial data is constructed as <time, longitude, latitude>, and is fed into the model. The model needs three additional factors.

1. Eps1: The spatial density threshold (maximum spatial distance) between two points to be considered related.
2. Eps2: The temporal threshold (maximum temporal distance) between two points to be considered related.
3. Min samples: The number of samples required for a core point.

According to the reference[9] document, we conclude that Vespa mandarinia queens have a range estimated at 30km for establishing her new nest. Thus, we assign 30km to the largest radius of the activity of Vespa mandarinia. Besides, bees are social animals so the spotting cases within one month are most likely to be the same group of Vespa mandarinia. Finally, we set Eps1, Eps2, and Min samples to 0.3(after unit conversion),30 and 5, respectively.



(a)Negative reports distrbution (b)Unverified reports distrbution

Figure 5. Comparison of the distribution of negative category and unverified category.

We unexpectedly found that the distribution of negative and unverified data is very similar in Fig. 5, and a large amount of data are concentrated in plain areas with relatively low altitudes. It indicates that many records are false sightings.

3.1.2 Analysis on spatial temporal features

The components in ST-DBScan code can be used to value the connections between each other on the maps.

- positive/negative category

$$W = \begin{bmatrix} A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & A_7 & A_8 \\ B_1 & B_2 & B_3 & B_4 & B_5 & B_6 & B_7 & B_8 \\ C_1 & C_2 & C_3 & C_4 & C_5 & C_6 & C_7 & C_8 \end{bmatrix}$$

We assume that there are eight points, three of which form a cluster and the other five points do not form any cluster. This matrix $W^{3\times 8}$ represents the distance between these three core clustering points and all other data. For example, A_4 represents the distance between the fourth point and the first point of the three points that have formed a cluster. The smaller the distance, the closer the relationship between the fourth point and the cluster. The distance from the core clustering point itself is 0. We only need to add up all the columns in the matrix and divide by eight to get the aggregation degree of the eight points. Finally, the reciprocal of the value is defined as ST_{score}. For example, the influential factor of the i-th point is calculated as:

$$Score_{ST_i} = \begin{cases} \frac{1}{\frac{A_i+B_i+C_i+...}{N}}, \ i \in S_{positive} \\ -\frac{1}{\frac{A_i+B_i+C_i+...}{N}}, \ i \in S_{negative} \end{cases} \tag{1}$$

where N is the total number of points.

- unverified/unprocessed category

The next step is to determine the influence of positive and negative reports on neighboring areas. We begin with the idea of gravity. In physics, every particle attracts every other particle in the universe with a force that is directly proportional to the product of their masses and inversely proportional to the square of the distance between their centers:

$$F_g \propto \frac{m_1 m_2}{r^2} \tag{2}$$

Therefore, we define the influence of positive and negative categories on the temporal spatial data of unverified and unprocessed categories as the following formula:

$$Score_{ST_i} = \sum_{j \in positive} \gamma^{t_{i,j}} \frac{Score_{ST_j}}{r_{i,j}^2} + \sum_{k \in negative} \gamma^{t_{i,k}} \frac{Score_{ST_k}}{r_{i,k}^2}, \ i \in S_{unverified \ or \ unprocessed} \tag{3}$$

where γ is a discounted factor of time, and we set it to be 0.999. $r_{i,j}$ is the geographical distance between the i-th point and the j-th point. $t_{i,j}$ is time intervals(Unit: day) between the two reports. Since $Score_{ST}$ of negative report is negative, the latter one is also a negative number.

3.2 Textual Features

We utilize text data from three perspectives, extracting the words of strong correlation with three types including high confidence, detailed description, or mistaken possibility, which are denoted as CF, DT, MP respectively. The scores are calculated in the manner of the naive Bayes method, which represents the confidence and helpfulness report notes.

Besides, we match the ground truth labels of wasp species likely to be mistaken from lab comments, for the sake of the following image classification task.

3.2.1 Data processing

We first merge the data with images and mark each report by whether it is with an image, which implies that the lab comments may not only rely on the report notes. We then generally select features including report notes, lab status, lab comments, and the mark of whether it is with images for text mining. To further utilize the information conveyed by text data, we conduct the following processing methods to texts:

1. Texts are split into words, and misspellings are corrected.
2. Verbs are lemmatized to a first-person single form, nouns are lemmatized to a single form, and comparative degree and superlative degree of adjectives and adverbs are converted to their basic form.
3. Stopwords and punctuation are removed from sentences.
4. TF-IDF[10] (term frequency-inverse document frequency) is calculated and the LDA[11] (Latent Dirichlet Allocation) topic model is applied to the pre-processed data for text mining.

3.2.2 Insights of text data

After calculating the TF-IDF score of all the words in the pre-processed report notes, we find out the most frequently used words in report notes, and a wordcloud is generated as shown in Fig. 6.

The TF-IDF score for token t in document d from document set D is calculated as follows:

$$TF\text{-}IDF(t, d, D) = TF(t, d) \times IDF(t, D) \tag{4}$$

Where:

$$TF(t, d) = log(1 + freq(t, d)) \tag{5}$$

$$IDF(t, D) = log(\frac{N}{count(d \in D : t \in d)}) \tag{6}$$

Firstly, we filter out the most frequent words only in positive report notes, which turn out to be words of high professionalism and confidence, *e.g.* "WSDA" may imply that the detection was reported by professional staff of WSDA and thus more reliable, while "specimen" suggests that

Figure 6. Word cloud generated from report notes.

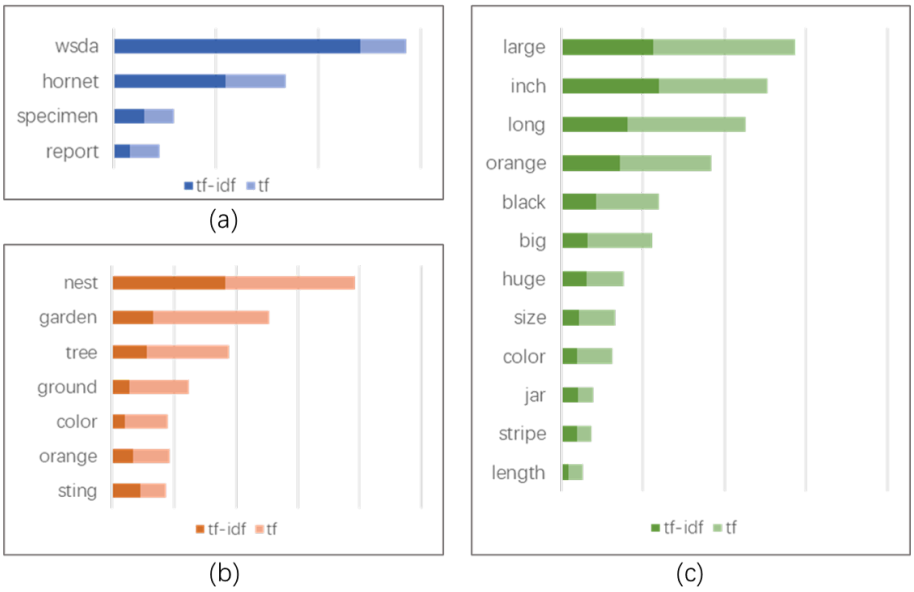


Figure 7. Terms of the highest TF-IDF scores.

scientific research has been conducted for the detection. Detailed results Detailed results are shown in Fig. 7 (a).

Then, We subjectively selected the detailed description words from the top 100 most frequently used words in all report notes, *e.g.* "large", "inch", "orange", "black", *etc.*, which means the report is more likely to be along with detailed descriptions and thus more helpful for lab researchers to decide its status. Detailed results are shown in Fig. 7 (b).

Next, when it comes to the lab comments of negative or unverified detection, they usually convey two kinds of information, including the actual species of the wasp detected and the reason for the negative judgment. Therefore, statistics are also conducted to lab comments of negative detection reports. After pattern matching with other species of high possibility to be mistaken (mentioned later), the rest of words turn out to be the features most likely to be misjudged for Vespa

mandarinia, *e.g.* "garden", "tree" are less likely to be the active area of Vespa mandarinia, while "orange" is the correction of the reported colors of "yellow" in notes (and thus we later replace "orange" with "yellow" in MP corpus). Detailed results are shown in Fig. 7 (c).

Based on the statistics above, we form the corpus of CF, DT, and MP, storing the most frequent terms under the three conditions mentioned above respectively.

term	frequency
WSDA	0.0441
specimen	0.0294
report	0.0294
staff	0.0147

(a) CF (High Confidence)

term	frequency
large	0.0172
inch	0.0133
orange	0.0129
stripe	0.0037

(b) DT (Detailed Description)

term	frequency
nest	0.0168
garden	0.0131
tree	0.102
yellow	0.0199

(c) MP (mistaken posibility)

Table 1. Excerpts from the corpus of different contexts.

Besides, we conduct pattern matching with common misjudged wasp species to lab comments for negative reports, from which to separate the real species of the wasp sighted. This provides ground truth labels for the following image classification task since the number of each class in the original data set is of severe imbalance. The numbers of each species are depicted in Fig. 8.

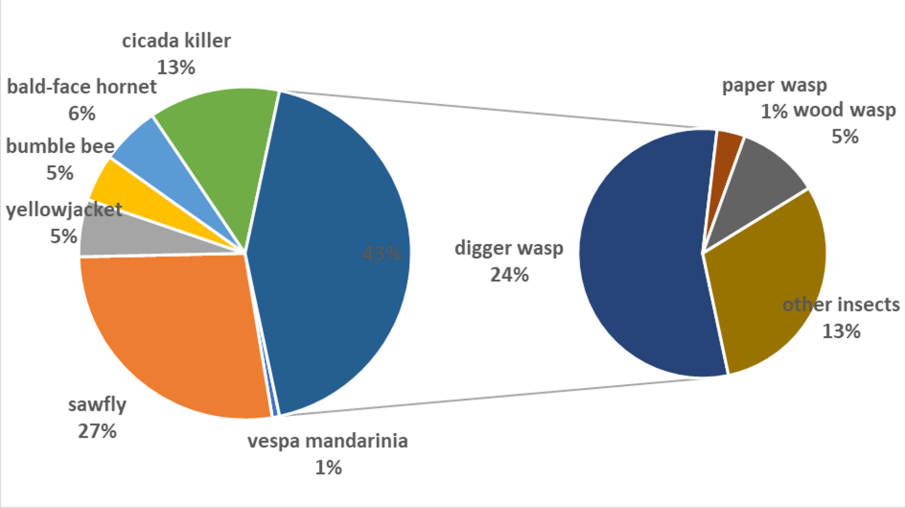


Figure 8. Percentage of insect type in all positive and negative reports.

3.2.3 Bayes scoring strategy

We apply the naive Bayes[12] method to report notes based on the CF, DT, MP corpus mentioned above. The Naive Bayes method is a conditional probability model, which can be

decomposed as:

$$\begin{aligned} P(c \mid d) &= \frac{P(d \mid c) \cdot P(c)}{P(d)} \\ &\propto P(d \mid c) \cdot P(c) \\ &= P(c) \prod_{k=0}^{n} P(t_k \mid c) \end{aligned} \tag{7}$$

where $P(c \mid d)$ is the probability of document d belonging to class c, and $P(c)$ is the prior probability of class c.

To simplify the model, we assign the weight of each class with 2, 0.5 and -1. Therefore, the textual scoring strategy can be expressed as:

$$Score_{textual} = \sum_{k=0}^{n} weight_{t_k \in c} \cdot TF(t_k) \tag{8}$$

where $TF(t_k)$ is the term frequency of term t_k in document d.

3.3 Visual Recognition

In 2013, AlexNet[13] used the CNN structure on the ImageNet[14] data set to defeat all traditional algorithms. Since then, CNN has become the mainstream method in image processing tasks. Compared with traditional algorithms that require feature engineering processing of pictures, CNN can learn the features of the classified objects by the neural network itself and iteratively update the network parameters.

We use Resnet18[15] as our CNN backbone and introduce the idea of transfer learning[16]. After loading the pre-trained weights of Resnet18 on ImageNet, we finetune the whole net on given data set and finally achieve a considerable effect.

Before we show the analysis results, there are several training details to be stated:

3.3.1 Loss function

During the training, we use Cross Entropy Loss (CE) as the loss function:

$$loss(x, class) = -log(\frac{exp(x[class])}{\sum_{j}^{C-1} exp(x[j])}) = -x[class] + log(\sum_{j}^{C-1} exp(x[j])) \tag{9}$$

where C is the total number of categories.

3.3.2 Normalization

Because the images provided by the public are very different, such as different light intensity and color difference, we perform Gaussian normalization processing on all images, and the processed data obeys Gaussian distribution.

$$x = \frac{x - \mu}{\sigma} \tag{10}$$

where μ and σ are the mean and standard deviation of each picture.

3.3.3 Training data

We divide the data set, and the ratio is the N(training set): N(validation set)=6:4. The results of the division are shown in the Table 2. Our model is trained for 20 epoches(64 batches in each batch) with learning rate α =5e-4 and Adam optimizer.

Category	bald face hornet	bumble bee	cicada killer	digger wasp	other insect
Train	95	84	245	422	247
Validation	63	55	162	280	163
Total	158	139	407	702	410

Category	paper wasp	sawfly	Vespa mandarinia	wood wasp	yellowjacket
Train	22	504	9	68	93
Validation	13	335	5	45	61
Total	35	839	14	113	154

Table 2. The number of images in data set used in Vespa mandarinia recognition.

Shown in Table 3, our model achieves 63% accuracy on the validation set and recognize 80% of Vespa mandarinia, which demonstrates that our model has a strong ability of finding Vespa mandarinia.

Category	bald face hornet	bumble bee	cicada killer	digger wasp	other insect
Accuracy	33.33%	29.09%	58.02%	80.36%	52.76%

Category	paper wasp	sawfly	Vespa mandarinia	wood wasp	yellowjacket
Accuracy	30.77%	80.00%	80.00%	8.89%	39.34%

Table 3. The accuracy of each category on validation set.

When it comes to the quantitative score, we use the convolutional neural network to run inference on each picture, and keep the value of the Vespa mandarinia category as the recognition score of the visual features of the picture.

$$Score_{Recognition_i} = Value(Recognition\ Result_i(Vespa\ mandarinia)) \tag{11}$$

4 Problem 3: Score Ranking

4.1 Aggregation Method

Based on the model of the second question, we score all the reported temporal-spatial data, textual data, and image data. We believe that the higher the temporal spatial correlation between a

certain report and the positive reports, the more text keywords appear, and the more accurate image recognition, the more likely the report is positive sighting. Therefore, we sum the three scores and then sort all the scores to find the 25 reports with the highest score. It has to be mentioned that we normalize the three scores for the sake of fairness.

The formula for aggregation is as follows:

$$Score = Score_{ST} + Score_{Textual} + Score_{Recognition} \tag{12}$$

Not surprisingly, all positive category scores are the highest, and 14 positive reports occupy the top 15 of the list. In addition to the positive category, there is also an unverified report ranked ninth, which shows that this report is very likely to be positive sighting. The top-25 highest score report is shown below in Fig. 9:

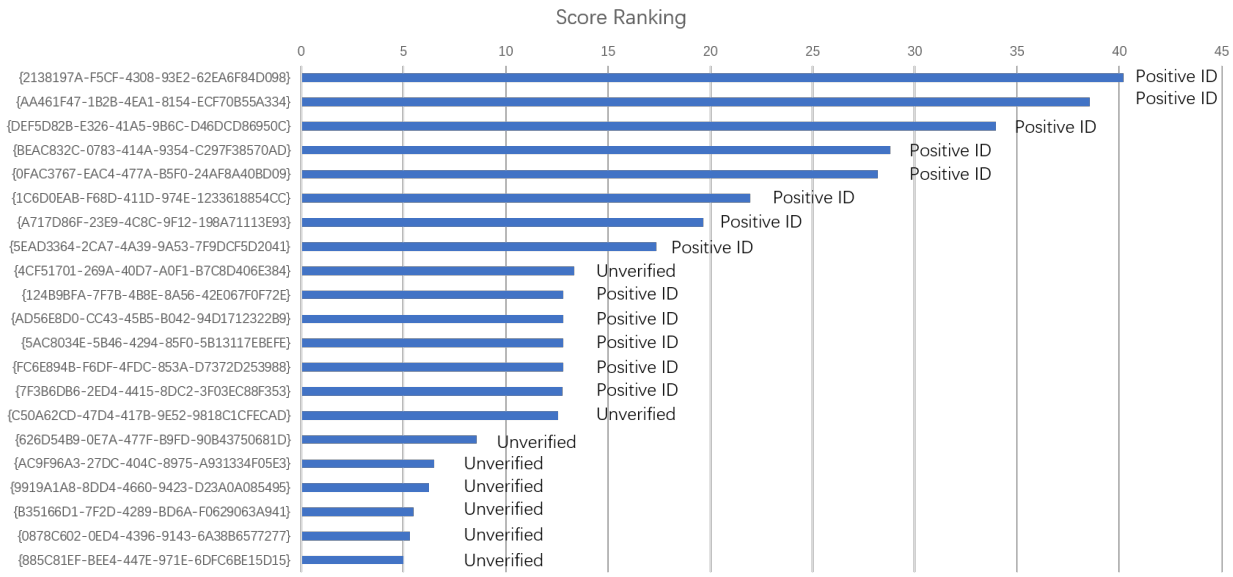


Figure 9. Top 25 reports with highest scores.

4.2 Geographical Visualization

Then we also visualize these points on a geographical map as shown in Fig. 10. The results show a high correlation and directly prove that our model works well. We recommend that WSDA pay more attention to these 25 reports first, because they indicate that Vespa mandarinia may appear nearby.

5 Problem 4: Model Entropy

In this section, we will discuss the stability of START model. Obviously, it is an important question to update our model given additional new reports over time and decide how often the updates should occur.

We decided to discuss this matter in two aspects, and divide the report into two categories: high confidence and low confidence type. Specifically, the report clearly marked as positive or

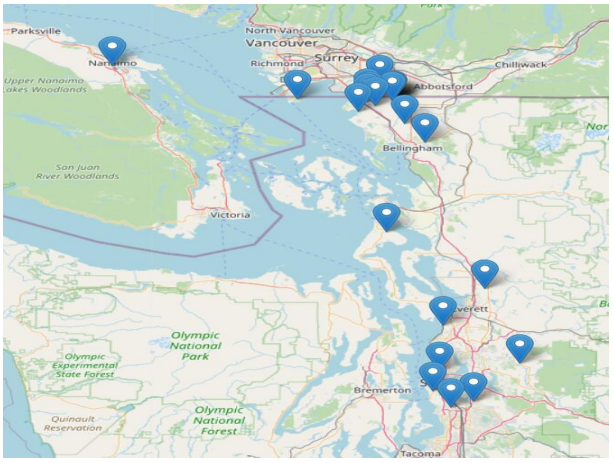


Figure 10. Visualization of 25 reports with the highest score on a geographical map.

negative is defined as a report with high confidence; the report marked as unverified or unprocessed is defined as a report with low confidence.

5.1 Data of High Confidence

Since the data of the positive category is very scarce, we must make full use of it after we get a new report of positive, such as its images, location, and detection date, which will definitely help to track the trail of Vespa mandarinia. At the same time, negative reports can help CNN architecture better identify pictures of insects and indirectly strengthen the ability to find error reports. In addition, the densely appearing bee colonies in negative reports often represent the absence of Vespa mandarinia, because Vespa mandarinia will feed on other bee colonies. All in all, when the report is calibrated by researchers as positive or negative, we must retrain our network with the new data set. The specific method is as follows:

- Temporal spatial features: Rerun the clustering algorithm and calculate the ST_{score} of each point as discussed in Section 3.1.2.
- Textual features: Simply update the term frequency in CF, DT and MP corpus with the report notes, especially important for positive reports,
- Recognition: Divide the data set again using the new label of each image. The image attached to the negative report can be obtained in lab Comments column. Then load the existing model weight to continue training and set a small learning rate(*e.g.* 5e-5) until the model converge again on the new data set.

5.2 Data of Low Confidence

Unverified and unprocessed data is very difficult to use for most models. An important reason is that it does not have a clear label, so we cannot predict its impact on the effect of the model. But START model has the ability of clustering points, so it can naturally identify information points and noise points. Based on this idea, we introduce the concept of model entropy to represent the stability of the model.

Entropy is often interpreted as the degree of disorder or randomness in the system. The term is used in diverse fields, from classical thermodynamics to the principles of information theory. We define the formula for model entropy as follows:

$$Model\ Entropy \quad = \quad \frac{-\sum_{i=0}^{N-1} log(Score_{ST_i})}{N} \tag{13}$$

where N is the total number of points.

Then we manually make up some fake data in positive ID. An example is shown in the Table 4 below. The detection date is 1 day after the true report and the geographical distance is $0.01°$ in both longitude and latitude, which ensures that the new report can be divided into the cluster.

Status	Detection date	Lab status	Longitude	Latitude
True	2020-10-1	positive ID	-122.582465	48.983375
Fake	2020-10-2	positive ID	-122.592465	48.973375

Table 4. Manually make up fake data to check the stability of START model.

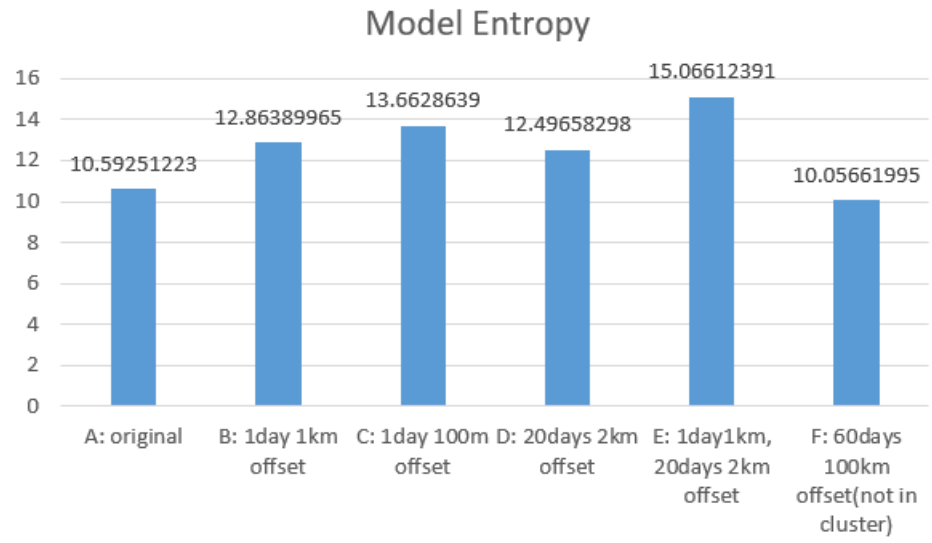


Figure 11. Model entropy with different fake data.

We conduct five sets of controlled experiments, and the result is shown in Fig. 11. We find that the closer the newly inserted data is to the original data, the greater the entropy of the model (BCD); if two points are added to the cluster (E) at the same time, the model entropy will increase significantly. And if the newly added data is not related to the original data (F), then the model entropy will not increase, or even slightly decrease due to the increase of the number of points.

So, we safely conclude that once new reports come in, we can calculate the model entropy to measure the stability and accuracy of the START model. If the entropy increases sharply, we will regard the newly inserted point as positive data. Otherwise, it is a negative point. Based on

empirical experience, when the new entropy is equal to 1.5 times the original entropy, we update the original model, and the updated plan refers to Section 5.1.

6 Problem 5: Active Threshold

In this section, we focus on the end of pest spread. We will discuss this issue in conjunction with the scoring mechanism of the START model and the life habits of the Vespa mandarinia.

According to the discussion of the third question, we will draw a line at the position of the score of the last positive report, and use this line as the active threshold of our system. Any report below this baseline score will be treated as negative. If a positive report appears in the updated data and the score is lower than the threshold, then we will re-update it to the lowest scoreline of the positive report scores. If there is no data that exceeds it for a long period of time, *i.d.* two months, we believe that Vespa mandarinia has been completely extinct in Washington State. Two months is inaccurate, because we also need to consider Vespa mandarinia hibernation.

According to the report [9], when Winter arrives, the current seasons' nests die out and only overwintering queens survive. Overwintering queens emerge in the spring and they find safe areas to begin building a nest. By August the hive is fully mature, with up to 100 worker bees. The queens produce drones in September. Drones and queens leave the nest to mate in October and early November.

So if the positive reports do not appear for two months in Winter, we cannot conclude that the pest has been eradicated in Washington State. We just skip the winter report. Starting from the next spring (roughly in March), if there is no report that exceeds the threshold of the system within two months, we can announce that the Vespa mandarinia is extinct. The pseudo-code is as follows.

Algorithm 1 Framework of constituting evidence that Vespa mandarinia is eradicated.

Input: The new reports including detection date, sighting location, notes, lab comments, lab status and images.

Output: The conclusion whether Vespa mandarinia is eradicated or not.

```
initial Counter=0, bottomline = +∞
if Counter == 0 in two month then
    return eradicated
end if
if report submits in Winter then
    continue
else
    if report.value < bottomline and report is not positive then
        continue
    else
        Counter += 1
        if report.value < bottomline then bottomline = report.value
        end if
    end if
end if
```

7 Strengths and Weaknesses

7.1 Strengths

- **Strength 1**
  We use multi-modal technology to process data in different formats, including text, images, and geographic information.
  - We use STDBScan to cluster temporal spatial data to find the degree of correlation between reports.
  - We use lab comments to annotate image data categories and use LDA and TFIDF models to analyze the word frequency of notes to help us understand why people mistake other insects for Vespa mandarinia.
  - We use CNN to train the annotated image data with 10 categories and achieve a good result on the validation set. The accurate recognition rate of Vespa mandarinia reaches 80%.
- **Strength 2**
  We use a scoring mechanism to integrate the above three scores, and the report with the highest score is considered the most likely positive sighting.
- **Strength 3**
  We introduce the concept of model entropy and regard the impact of new data on model entropy as an important factor in whether the model is worth updating.

7.2 Weaknesses

- **weakness 1**
  Except for the first question, our model does not focus on biological characteristics in most cases, but only focuses on data features, such as temporal and spatial distribution, text description, and image appearance.
- **weakness 2**
  The problem of data imbalance limits the accuracy of our model.
- **weakness 3**
  Due to time constraints, in the text processing dictionary, our description dictionary of the Vespa mandarinia is not very complete.

8 Conclusions

From December 2019, the corpse of a Vespa mandarinia made people alert. With the arrival of a large number of reports carrying text, pictures, and geographic information, a new multi-modal solution needs to be proposed to find the most informative data from the complex data.

We use temporal spatial data clustering to track the trajectory of the Vespa mandarinia, and use the lab comments in the report to label the image data with category information. Then, we analyze

the word frequencies of report notes to find out the mistaken classified reason. Further, a CNN network is used to process the annotated pictures. We fuse temporal spatial data, text feature, and visual feature scores to get the final report score. Based on this, we find the most valuable reports and submit our memorandum to WSDA.

9 Memorandum

Date: February 9, 2021
To: Administrator, Washington State Department of Agriculture
From: Team #2112243

Chief Administrator,

From December 2019, the discovery of Vespa mandarinia has made us all very alert and worried because it will harm local beneficial insects. Fortunately, the Washington State government announced a series of response measures in time, including allowing the public to submit reports of seeing Vespa mandarinia. But another problem has arisen. A large number of complex reports have stretched all WSDA staff, so we propose a model that can help you deal with multi-modal report data. Let me introduce the results of our research.

- **Consider sending more faculty to $48°59'3''$N, $57°25'30''$W.**
  We visualize the positive report result on a geographical map and find that the distances in space are less than $\pm 0.5°$ in longitude(around 31 miles) and $\pm 0.2°$ in latitude(around 12 miles). Then we use the STDBScan model to cluster the existing positive reports and get the most relevant reports in the temporal spatial data are concentrated in the $48°59'3''N$ $57°25'30''W$ area. Therefore, we recommend that WSDA send more personnel to the surrounding area to check if there are Vespa mandarinia gathering.

- **Score the report notes with our textual model.**
  Based on the analysis of report notes, we find out the words of strong correlation with high confidence, detailed description and mistaken possibility, and thus able to score the report notes according to their confidence and helpfulness. Our model then may serve as the filter for selecting the reports with high probability before processing by lab researchers and WSDA staffs, prioritizing the most informative and helpful ones.

- **Run our CNN model to identify Vespa mandarinia in thousands of images.**
  We mark the image data based on the existing text data and send the well-annotated data to the convolutional neural network, also named CNN. The number of this data set has reached about three thousand. CNN can achieve 63% accuracy on the validation data set, and can accurately identify 80% of Vespa mandarinia. With this good helper, I think the staff of WSDA will no longer have to worry about a large amount of image data.

- **Apply our aggregation score to get the most informative report quantitatively.**
  The multi-modal model we proposed can find the most informative report among all reports by scoring. Based on the scores of the above three aspects, we add them and then sort them,

and we can easily get the most likely positive reports. With the guidance of this data, you can dispatch personnel to the most critical areas with the highest scores.

- **Use model entropy to process the unverified or unprocessed report.**
  Due to time limitations, the WSDA staff cannot handle all reports in time. Therefore, we propose the concept of model entropy to speed up the identification of unprocessed or unverified reports. We only need to observe whether a new report has a great impact on the model entropy. If it does, then it means that the report has great reference value. Otherwise, it means that the report has a low probability of being informative.

We are really honored to provide you with a model for processing multi-modal data and finding the most informative reports. Hope our suggestions may help.

Sincerely yours,

Team #2112243

References

[1] 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, oct 1973.

[2] Sifeng Liu and Yi Lin. *Introduction to Grey Systems Theory*, pages 1–18. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[3] Deng Julong. Introduction to grey system theory. *The Journal of grey system*, 1(1):1–24, 1989.

[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] A. Peterson and Jorge Soberón. Species distribution modeling and ecological niche modeling: Getting the concepts right. *Natureza & Conservacao*, 10:102–107, 2012.

[6] Jane Elith, Michael Kearney, and Steven Phillips. The art of modelling range-shifting species. *Methods in Ecology and Evolution*, 1(4):330–342, 2010.

[7] Nuñez-Penichet C, Osorio-Olvera L, Gonzalez VH, Cobos ME, Jiménez L, DeRaad DA, Alkishe A, Contreras-Díaz RG, Nava-Bolaños A, Utsumi K, Ashraf U, Adeboje A, Peterson AT, and Soberon J. Geographic potential of the worlds largest hornet, vespa mandarinia smith (hymenoptera: Vespidae), worldwide and particularly in north america. *PeerJ*, 9:e10690, 2021.

[8] Alp Kut Derya Birant. St-dbscan: An algorithm for clustering spatialtemporal data. *Data & Knowledge Engineering*, 60(1):208–221, 2007.

[9] Michael J. Skvarla. Asian giant hornets. `https://extension.psu.edu/asian-giant-hornets`, accessed on February 8, 2021.

[10] Claude Sammut and Geoffrey I. Webb, editors. *TF–IDF*, pages 986–987. Springer US, Boston, MA, 2010.

[11] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):9931022, March 2003.

[12] Geoffrey I. Webb. *Naïve Bayes*, pages 713–714. Springer US, Boston, MA, 2010.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 10971105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[16] Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

Appendices

Appendix A Selected Source Code

st_dbscan.py

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from st_dbscan import ST_DBSCAN

df = pd.read_csv("2021MCMProblemC_DataSet.csv")
df['time'] = (pd.to_datetime(df['time'])-pd.to_datetime(df['time'].min())).map
    (lambda x:x.days)
data = df.loc[df['label']=="Positive ID", ['time','x','y']].values
st_dbscan = ST_DBSCAN(eps1 = 0.3, eps2 = 30, min_samples = 5)
st_dbscan.fit(data)
def plot(data, labels):
    colors = ['#a6cee3','#1f78b4','#b2df8a','#33a02c','#fb9a99','#e31a1c','#
        fdbf6f','#ff7f00','#cab2d6','#6a3d9a']
    for i in range(-1, len(set(labels))):
        if i == -1:
            col = [0, 0, 0, 1]
        else:
            col = colors[i % len(colors)]
        clust = data[np.where(labels==i)]
        plt.scatter(clust[:,0], clust[:,1], c=[col], s=5)
    plt.xlabel('longitude')
    plt.ylabel('latitude')
    plt.show()
    return None
plot(data[:,1:], st_dbscan.labels)

score = []
for i in range(len(st_dbscan.labels)):
    score.append(0)
    for component in st_dbscan.components:
        score[i]+=component[i]
    score[i] /= len(st_dbscan.labels)
    # score[i] = -score[i]
    score[i] = 1/score[i]
print(score)
```

GM(1,1) grey differential model predicting program

```
import numpy as np
import math

class GM:
    def __init__(self,X:np.ndarray):
        self.X = X
        assert self._assert(self.X),""
        self.a = None
        self.b = None
    def _grade_ratio(self,X: np.ndarray):
        lambda_k = np.array([X[k - 1] / X[k] for k in range(1, len(X))])
        return lambda_k
    def _assert(self, X: np.ndarray):
        lambda_k = self._grade_ratio(X)
        lower = math.exp(-2 / (len(X) + 1))
        upper = math.exp(2 / (len(X) + 1))
        lambda_k = (lambda_k < upper) * (lambda_k > -lower)
        if np.sum(lambda_k) == len(lambda_k):
            return True
        else:
            return False
    def _increasing(self,X):
        aggregate_result = []
        sum = 0
        for k in range(len(X)):
            sum += X[k]
            aggregate_result.append(sum)
        return np.array(aggregate_result)
    def _decreasing(self,X:np.ndarray):
        X_tilde = [X[0]]
        for k in range(1,len(X)):
            X_tilde.append(X[k]-X[k-1])
        return np.array(X_tilde)
    def _mean_generating_series(self,X:np.ndarray,alpha = 0.5):
        return -np.array([alpha*X[k]+(1-alpha)*X[k+1] for k in range(len(X)-1)
            ])
    def _solve_grayeq(self,X:np.ndarray,Z:np.ndarray):
        B = np.vstack((Z,[1]*len(Z)))
        Y = X[1:]
        return np.linalg.inv(B.dot(B.T)).dot(B).dot(Y.reshape((len(Y),1)))
    def _solve_whiteeq(self, X: np.ndarray, theta:np.ndarray):
        a,b = theta
        X_tilde = []
        for k in range(1,len(X)):
            X_tilde.append(float((X[0]-b/a)*math.exp(-a*k)+b/a))
        X_tilde.insert(0,X[0])
        return np.array(X_tilde)
    def _final_check(self,X_tilde):
        print("\noriginals:",self.X)
        print("model values:",X_tilde)
        print("residual:",self.X-X_tilde)
        print("relative error:",np.abs(self.X-X_tilde)/self.X*100)
```

```
        print("ratio error:",self._decreasing(self._grade_ratio(X_tilde)))
    def train(self):
        X = self._increasing(self.X)
        Z = self._mean_generating_series(X)
        theta = self._solve_grayeq(self.X,Z)
        X_tilde = self._solve_whiteeq(self.X,theta)
        X_tilde = self._decreasing(X_tilde)
        print("prediction result:",X_tilde)
        self._final_check(X_tilde)
        self.a = theta[0]
        self.b = theta[1]

X = np.array([52,51,51,51,50,50,50,50,52,51,50,51,54,51])
gm = GM(X)
gm.train()
```