Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2126594

Predict the spread and buzz about Asian Giant Hornets in Washington based on VAR and DCNN

The impact of biological invasions is always a global concern. In 2019, four Asian giant hornet workers were found in Washington and it's not clear whether they are established in North America and how widespread they are. The Washington government hopes to better predict the spread of the hornet and carry out investigations under limited resources.

To help the Washington government predict the spread trend, we build models to map the spread path and calculate the total number of the hornet respectively. As to the spread path, we first calculate the distance between sightings by using their coordinate information. We then define two high-frequency areas and re-divide the sample accordingly. Next, we build VaR models to calculate the spread direction in two areas. Last, we map the spread path and analyze its trend. As to the number prediction, we build Fourier (number of terms=3) model to estimate the number of reports in each month of a year. Since the total number of reports is highly correlated to the number of positive reports, we can get a view of how many potential positive reports there will be in future months by this model.

According to the problem 2, create the classification model of the reported sighting and analyze the likelihood of mistaken classification. To preprocess the multimodal data, we convert all images (including frames extracted in the video) into fixed-size image. To create the classification model, we apply the deep convolutional neural network to better extract image as well as video features whose frames are conducted by mean pooling. We use transfer learning to solve the problem of relatively small amount of data for computer vision, and an over-sampling method SMOTE is applied to overcome the difficulty of class imbalance. We also try a language model BERT with dimension reduction method SNE to prove that the "notes" in the dataset have no significant help for classification.

According to the problem 3, we establish an evaluation model to prioritize the investigation of reports most likely to be positive sightings. Apart from the Image Positive Rate (IPR) derived from the classification model, we also define another index Distance Preference Score (DPS) as a measurement of the report's location' favor in being positive sighting, as another decition indicator. With the help of TOPSIS method, we calculate the Relative Closeness Degree (RCD) of each report, which is the only reference index for prioritizing the investigation.

According to the problem 4, we clarify the model updating and optimization process given new reports by discussing its model updating mechanism, regional applicability, timeliness and seasonality; Determine the updating frequency of previously constructed models respectively based on their properties confronting specific data.

As to the evidence for eradication of this hornet in Washington: When the number of supplementary reports or supplementary positive reports in a certain month is 0, it can be considered that Asian giant hornets have been eradicated.

Key Words: Asian Giant Hornets, VAR, DCNN, imbalanced data, multimodality

Contents

1 Introduction

1.1 Problem Background

The impact of biological invasions on agriculture, biodiversity and human activities is always a global concern. As found during the invasion of Vespa velutina in Europe, some hornets can have serious effects on local honey bees and deliver painful stings with venom. In 2019, a nest of Asian giant hornets was discovered and destroyed in western British Columbia, and later on, four workers were found in Washington. Though at present Asian giant hornets are not known to occur outside of Washington state and Vancouver Island, it's not clear whether they are established in North America and how widespread they are. Under such circumstances, the state government set up helplines and a website for people to report sightings of these wasps. But since Asian giant hornets can be easily confused with other species of wasps (e.g. European hornets and Eastern cicada killers), those sightings are often invalid. The State of Washington hopes to screen out sightings where Asian giant hornets are more likely to appear and prioritize public reports for follow-up investigation. To help the Washington State Department of Agriculture achieve their goals, we need to：

- Discuss the predictability of the spread of Asian giant hornet and the accuracy of the prediction.
- Build a classification model and predict the likelihood of a mistaken classification using the data set file and image files provided
- Based on models above, screen out sightings more likely to be positive for priority investigation.
- Explain how to update the models and how often the models should be updated.
- Based on models above, discuss how to conclude that Asian giant hornet in Washington State has been eradicated.

1.2 Our Work

In order to address the problems above, we will proceed as follows:

- Build Vector Autoregression (VaR) model and Fourier (number of terms=3) model to respectively show the spread direction and total number of Asian giant hornets in a certain month in the future.
- Apply deep convolutional neutral network to assess image features and full connection layer for videos; Build a classification model to show both the classification result and the likelihood of a mistaken classification.
- Use Image Positive Rate (IPR) and Distance Preference Score (DPS) as attributes to establish a TOPSIS model.
- Clarify the model updating and optimization process given new reports by discussing its model updating mechanism, regional applicability, timeliness and seasonality; Determine the updating frequency of previously constructed models respectively based on their properties confronting specific data.
- Combine the results of Fourier (number of terms=3) model for number prediction and the classification model, decide whether the pests have been eradicated.

1.3 Literature Review

As to the spread of hornets, several articles have studied this from different aspects. One studies the mechanism of the spread. To predict the spread of Asian hornet in Great Britain, Matt J. Keeling divides the whole process into two parts, which are generation of new queens and dispersal of these queens across a heterogeneous landscape [1]. Another one probes into factors which may influence the spread pattern of hornets. Christelle Robinet shows how a spread model can be used to explore humans' effect on expansion of the invasive yellow-legged hornet in Europe [2]. Besides, lots of studies have also predicted the impacts of hornet invasion. Alberto J Alaniz analyzes the potential impacts of Asian giant hornets on honey bee colonies and calculates relative economic losses in the honey bee industry and bee-pollinated croplands [3]. Different from studies above, our paper not only discuss the spread mechanism of Asian giant hornets, but also provide a plan for the state government to screen out sightings and prioritize investigations, which can be of more practical use.

2 Assumptions and Justifications

We make some general assumptions and explain their rationales for further study:

- **Assume that the active area and total number of Asian giant hornets will not change a lot within a month.** Changes in the location and number of wasps are results of biological activities, which follow a regular pattern and don't change sharply. This assumption makes sure that we can use monthly data for prediction, investigation and updating information.
- **Assume that the spread of the hornet starts from a center, the current active area is affected by the past.** Hornets don't fly for a long distance each time they change their nests, so the spread would be center on one location.
- **Assume that the number of hornets changes in a certain pattern, and the current number is between the number of the previous month and the next month.** This is reasonable since number of Asian giant hornets changes with season. This assumption makes sure that linear interpolation can be used to supplement the missing month data.
- **Assume that the number of reports is linearly proportional to the number of positives found, and the number of positive reports is proportional to the number of potential wasps.** This assumption simplifies our model for number prediction.

3 Predict the Spread of Asian Giant Hornets

3.1 Determine the Spread Direction

3.1.1 Data Processing

Assuming that the spread of Asian giant hornets starts from a center, we first measure the distance between the 14 positive sightings in the sample. We obtain two high-frequency regions for Asian giant hornets to appear by observing distribution pattern of those 14 positive sightings. The distance calculation formula is as follows:

$$d = 2r*\arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right)+\cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right) \tag{1}$$

r is the radius of the earth, φ_1, φ_2 represent the latitude of two points, and λ_1, λ_2 represent the longitude. If the distance is greater than 30km, the two places would be considered as relatively far away from each other, in which case the spread of hornets is unlikely to happen.

- The earliest positive report happened in Vancouver, September 19, 2019. Its geographic location is more than 30km away from the other 13 positive sightings. Among the 4440 reports, only one positive and one negative sightings are within 30km of this earliest location. Therefore, we can conclude that the hornets found in this location has not spread to the Washington State, so we eliminate this location in our further study.
- The second earliest positive report occurred on September 30, 2019. Among the remaining 12 positive sightings, except for the sighting that occurred on June 12, 2020, other sightings are all within 30km of the second earliest location. Therefore, we can conclude that 30km within this location is a high-frequency area for hornets to appear. Let this area be "Area A". In addition, there are 7 positive sightings within 30km of the location reported in June 6, 2020. Therefore, 30km within this location can be considered as the second high-frequency area, which is defined as "Area B".

We re-divide the samples according to the two high-frequency areas A and B above. In order to form continuous time series data, we convert daily data into monthly data and use linear interpolation to fill in missing values. Taking Area A as an example, we first include all the reports in this area into its database (with samples before 2018.09 been deleted), and then calculate the value of month t (2019.05 is the first month), the formula is:

$$(Latitude_t, Longitude_t) = \frac{\sum_i^{N_t} W_{ti} * (Latitude_{ti}, Longitude_{ti})}{\sum_i^{N_t} W_{ti}} \tag{2}$$

N_t is the number of samples contained in the database in month t. W_{ti} is the weight of sample i in month t, which depends on the "Lab Status" of the sample (see formula 3). $N_P/(N_P+N_N)$ represents the proportion of positive samples in all verified samples in the database of region A, which in this case is 8.70%.

$$W_{ti} = \begin{cases} 1, & \text{"Lab Status"} = \text{"Positive"} \\ 0, & \text{"Lab Status"} = \text{"Negative"} \\ N_P/(N_P+N_N), & \text{"Lab Status"} = \text{"Unverified"} \end{cases} \tag{3}$$

3.1.2 Modeling and Results

We use *Vector Autoregression (VAR)* model to estimate the spread path for the following reasons:

- By testing the stationarity of the time series, we find that both the latitude time series and the longitude time series are non-stationary series. They turn stationary after the first-order difference processing, so the *Autoregression (AR)* model should be adopted.

- By analyzing the correlation of the data, we find that the correlation coefficient of "Latitude" and "Longitude" is -0.77, which shows a strong negative correlation. Therefore, instead of adopting two AR models respectively, a VAR model should be adopted.

We first determine the lag order of the model. Take Area A as an example, according to the "lag order selection criteria", the lag order should be 6 because significant indexes are the most and AIC is the smallest. Meanwhile, in order to avoid excessively high freedom, we remove lag terms with a lag order of 4 or 5. Same for Area B.

The regression results (see Table 1) show that both equations for both models have passed the F test. For Area A, the equation with Longitude as the dependent variable is significant at the 10% level, while the other one is significant at the 5% level, meaning the equation as a whole is meaningful. Both R-square are above 90%, indicating the model fits well. For Area B, both equations pass the F test with significance of 5%.

Table 1: R-square and P Value of Two VaR Models (For Area A and Area B)

	Area A		Area B	
Equation	R-square	P > F	R-square	P > F
Latitude	0.9497	0.0229	0.6863	0.0363
Longitude	0.9130	0.0634	0.7466	0.0164

We map the spread path of Asian giant hornets in 2021 in the two areas based on the VaR model:

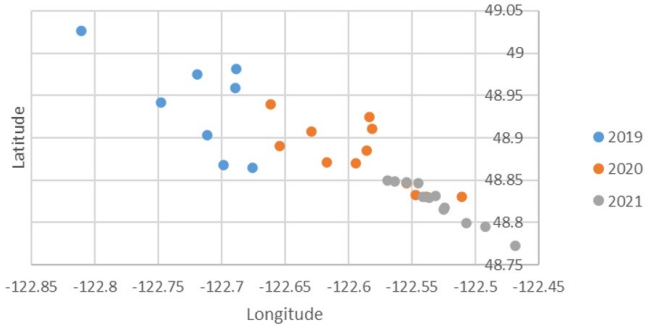

Figure 1: Spread Path of Area A in 2021

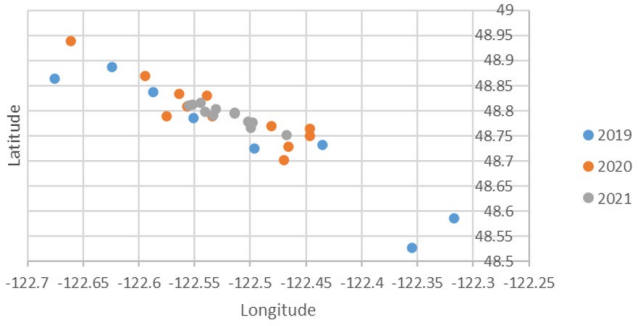

Figure 2: Spread Path of Area B in 2021

In Area A, we find that Asian giant hornets have a tendency to migrate to the southeast in recent years. Although the predicted active locations of the hornet in 2021 are still within the range of Area A (within 30km), most of the predicted locations are close to the boundary. This indicates that in 2021, Area A should no longer be considered as a high-frequency area, and the corresponding center should also move to the southeast. While hornets in Area B will spread more toward the center, so this area should still be considered important for hornet investigation.

3.1.3 Robust Test

First, the unit root tests below show that the VaR models for both Area A and Area B are stable, and the following Granger causality tests are meaningful (See Figure 3)

Then, Granger causality test is performed on the two models respectively. In the model for Area A, the assumption that longitude is Granger reason for latitude is accepted at a significance level of 5%, and the assumption that latitude is Granger reason for longitude is accepted at a significance level of 10%. This indicates that longitude and latitude are mutually cause and effect for each other. The test results for Area B are similar.

Finally, we use impulse response analysis to further test the stability of the model. As shown in Figure 4 in all cases, system disturbance converges to 0 in about one year under the impact of exogenous latitude or longitude changes. This proves that the prediction results of the model are stable.

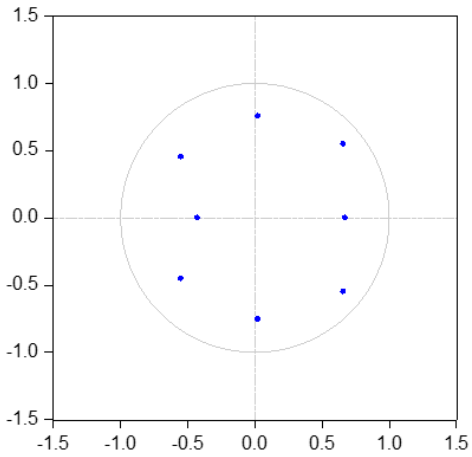

Figure 3: Unit Root Test (Area A)

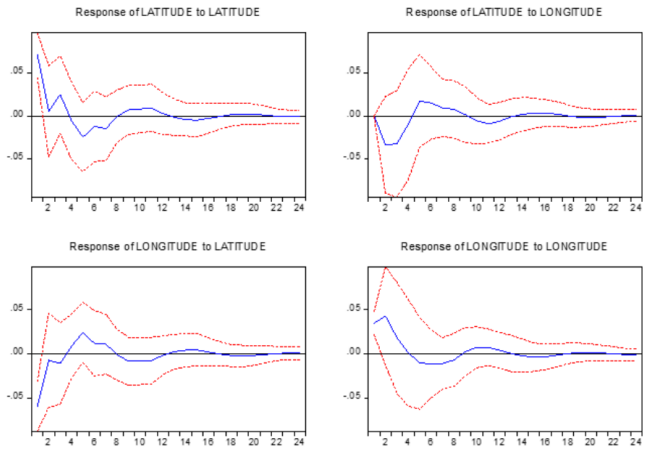

Figure 4: Impulse Response (Area A)

3.2 Calculate the Total Number

3.2.1 Data Processing

In order to predict the total number of hornets in a specific month, we assume that the total number is highly relative to the number of positive reports in that month. We also assume that the number of positive reports is linearly proportional to the total number of reports, so we can estimate the number of positive reports by predicting the latter.

Considering the shortage of reports in 2018 and before, the monthly number of reports in 2018 and before is not proportional to the number of positive ones, so we only select reports in 2019 and 2020 as basis for constructing the prediction model.

Since the time span of the sample is only two years, it is not enough to predict the annual growth trend. But because of the colony cycle of Asian giant hornets, number of reports show a strong seasonal trend (as shown in Figure 11-12). So we can estimate how the number of reports varies from month to month in a year.

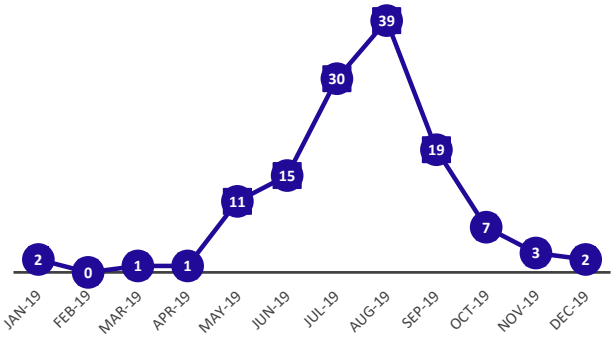

Figure 5: The Number of Reports in 2019

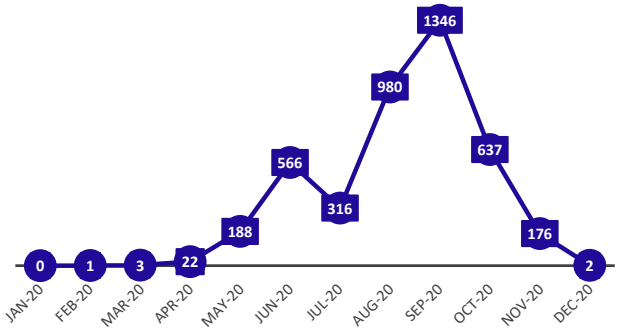

Figure 6: The Number of Reports in 2020

By observing the historical data, we find the number of reports is always close to 0 from November to December and from January to February, while reaches its peak in August. Therefore, we normalize the monthly data of each year (take the value of August as "1"), and define the result as “Relative positive ratio” (RPR_{jt}). It measures the ratio of the number of positive reports in month t to the peak (the number of reports in August) in the year j (based on the assumption that the number of positives is proportional to the number of reports), namely:

$$RPR_{jt} = \frac{\mathrm{Re}portsNum_{jt}}{\mathrm{Re}portsNum_{j8}} \tag{4}$$

It is easy to know that when the RPR_{jt} get larger, the number of potential hornets during the month t is closer to the upper limit of that year.

In addition, because queens spent the winter underground and all workers died in the nest, there are few observations from November to February. So we remove this part from the sample and only consider reports from March to November.

3.2.2 Modeling and Results

Since the sample of 2020 is more sufficient than that of 2019, we choose the former to fit the curve, and use the latter as dataset for testing. After the normalization process above, we use different function models to perform nonlinear regression on the time series data of 2020. Then we obtain 3 possible models under the standard of “R2>0.9”. We also use data of 2019 to further see whether the fitting effect of the curve. The details are as follows:

Table2: Fitting Effects of Three Models (2020)

	SSE	R^2	Adjusted R^2	RMSE
Sum of sine (number of terms=2)	0.08112	0.9122	0.7659	0.1644
Fourier (number of terms=2)	0.05038	0.9455	0.8546	0.1296
Fourier (number of terms=3)	2.262e-06	1	1	0.0015

Table2: Fitting Effects of Three Models (2019)

	SSE	R^2	Adjusted R^2	RMSE

Sum of sine (number of terms=2)	0.09836	0.8838	0.8838	0.1185
Fourier (number of terms=2)	0.06064	0.9284	0.9284	0.09308
Fourier (number of terms=3)	0.05911	0.9302	0.9302	0.0919

From the four evaluation indicators (SSE, R2, Adjusted R2, RMSE) above, *Fourier (number of terms=3)* performs best among three models whether using the original dataset or the exogenous one. So we choose *Fourier (number of terms=3)* as the model for predicting the relative positive rate of a month. Its specific form is:

$$y = 0.4159 + 0.3725*\cos(0.84x) + 0.1079*\cos(2*0.84x) + 0.1834*\sin(2*0.84x) + 0.1036*\cos(3*0.84x) + 0.001812*\sin(3*0.84x) \quad (5)$$

x is the month, and y is the corresponding RPR. All the estimated coefficients above are significant at the 10% level and are applicable to all months after 2019. The more general estimation of RPR is as follows:

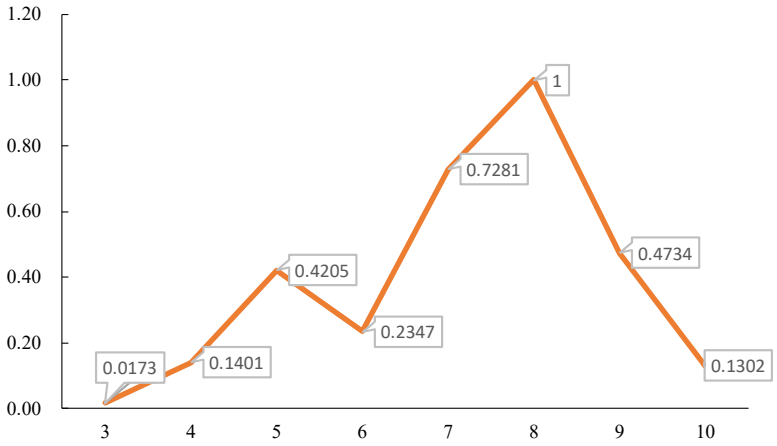

Figure 7: RPR from March to October

RPR can be used to estimate the number of reports in future months. For example, if 50 reports are received in March 2021, it can be estimated in advance that about 2887 (≈50/0.017318) reports will be received in August. Furthermore, the number of Asian giant hornets can also be roughly estimated. For example, if 5 hornets are found in April, it can be estimated that around 15 (≈10*0.4205/0.1401) hornets may be observed in May.

We also notice that, different from our intuition, the number of hornets doesn't follow a logistic expansion pattern. In fact, hornets observed in June are less than those in May. A possible explanation for this is that in late April, while inseminated queens start to search for nesting sites, the uninseminated queens do not search for nests since their ovaries never fully develop. They continue to feed, but then disappear in early June. The sharp rise is July is because hornet workers do not begin to work outside of the hive until July[4].

4 Build a Classification Model

4.1 Data Processing

- Purpose: Convert an image of any size to a fixed size.
- Method: Use bilinear interpolation to convert the size of input image to 256x256, and then crop a region with the size of 224x224 from the image center. Finally, normalize the pixel values. For videos, every frame of the video is extracted, and then the same preprocessing is done with the images.

4.2 Classification Model

In order to better extract image features, Resnet50 [5] is used as the image classification model. First, image features are extracted through deep convolutional neural network, and then the probabilities that the image is a positive example and a negative example are predicted through the full connection layer. For video classification, each frame of the video is input into the model separately, and mean pooling is conducted for all frames extracted from the model as the feature representation of the video, then the probabilities are predicted through the full connection layer.

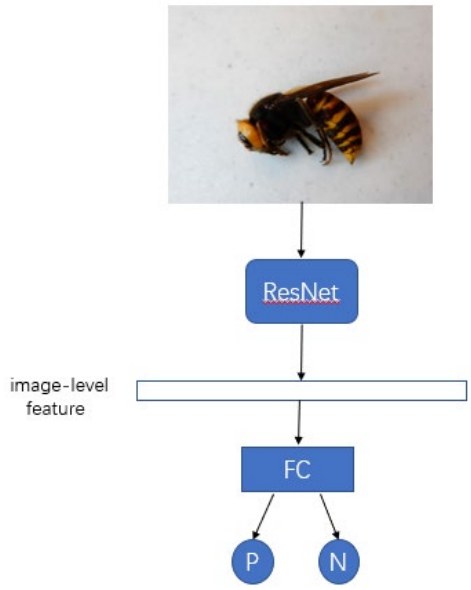

Figure 8: Image Classification Model

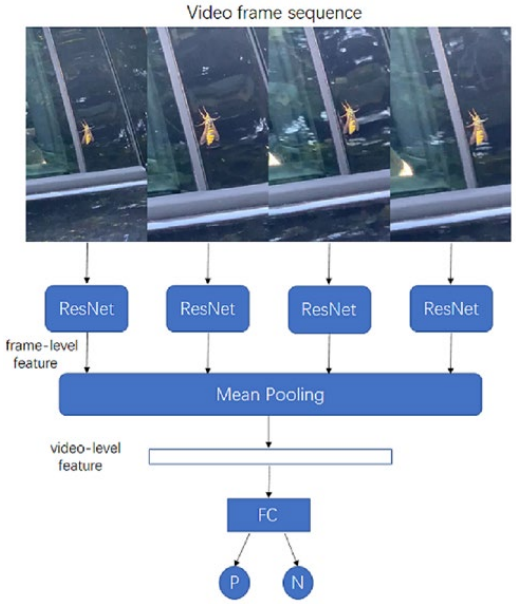

Figure 9. Video Classification Model

4.3 Difficulties and Solutions of the Problem

4.3.1 Small Amount of Data

The dataset contains about 4000 images and 100 videos, which is far less than ImageNet[6] and Microsoft Coco [7] , which contains millions and tens of millions of images and are commonly used datasets for computer vision. A small amount of data is easy to lead to overfitting, which reduces the prediction performance in real scenarios.

In order to solve this problem, we refer to the idea of transfer learning to train our own model on the basis of the pre-trained model using the provided data sets. Specifically, we try two schemes. In the first one, Resnet50 [5] pre-trained on ImageNet [6] is used as the feature extractor for images and videos. All parameters of the feature extractor were unchanged, and only the parameters of the full connection layer are adjusted through gradient descent. The second method uses the parameters of ResNet50[5] pre-trained on ImageNet [5] as the initialization of the model parameters, and adjusts all parameters of the model through finetuning.

4.3.2 Class Imbalance

After analyzing the data set, it is found that the ratio of positive and negative examples is about 1:150, which is extremely unbalanced and lead the classification model to prefer predicting the negative examples.

In order to solve this problem, a over-sampling method, SMOTE[8] is adopted to construct the positive example by linear interpolation of the k-nearest neighbors of the positive sample, so that the ratio of positive and negative examples is 1:1, thus alleviating the problem of unbalanced sample category.

4.4 Experiment

4.4.1 Dataset Partition

There are 6194 examples in total, and the training set and test set are divided according to the ratio of 7:3. The number of examples in the training set is 4335, and the number of examples in the test set is 1859.

4.4.2 Experimental Results on the Test Set

Table 3: Accuracy, Precision and Recall on the Test Set

Model	Accuracy	Precision	Recall
Feature Extract	0.9911	1.0000	0.9821
Finetune	0.9989	1.0000	0.9978

It is worth noting that we reject to use the language descriptions of discoverers provided in the dataset in our model, because these language descriptions are not clearly distinguishable and do not strongly correlate with the true category of the examples. We also test it.

First, we extract the vector representations of positive and negative language descriptions using BERT [9] (a pre-trained language model), and reduce the vector dimensions to using T-SNE [10] (a random neighborhood embedding method based on T distribution), then visualize the vectors. The visualization results are as follows:

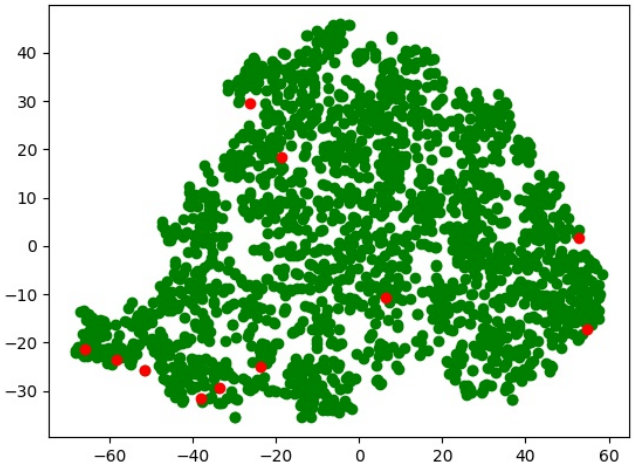

Figure 10: Visualization of Sentence Features

As we can see from the image, there is no obvious distinction between positive and negative examples in the same representation space, and even many positive and negative language descriptions are very close.

The reason for the imprecise language description is that the classification problem is fine-grained, that is, positive and negative classes belong to the same big category and have many same or similar features, which need to be distinguished through more detailed information. Unlike experts, the public don't know well about Vespa Mandarinia. It is difficult to grasp the key points for the public. Most of these descriptions are the characteristics common to Vespa Mandarinia and other wasps or bees. Therefore, the use of such textual data is likely to be unhelpful or even counterproductive to model performance.

5 Determine the Investigation Priority

5.1 Distance Preference Score (DPS)

In order to build an evaluation model, besides the Image Positive Rate (IPR) obtained from the classification model, we also need an index to measure the preference of positive examples at the time of report.

For every report, we calculate the average distance from its location to the center of every high-frequency area. Then this value is mapped to [0,1] as follows (see equation 6), so we get the Distance Preference Score (DPS) of report i in year t:

$$DPS_{ti} = 1 - \frac{AverageDistance_{ti}}{\max(AverageDistance_{ti})} \tag{6}$$

DPS measures the proximity of a reported sighting on a specific detection date to the average of the center of the high frequency area. We assume that a positive report is more likely to occur if the sighting is closer to the center of a high-frequency region. Therefore, we can conclude that the greater DPS is, the more likely it is to predict a report as a positive one.

5.2 TOPSIS Method

Then we use Image Positive Rate (IPR) and Distance Preference Score (DPS) as attributes to establish a *TOPSIS (The Technique for Order of Preference by Similarity to Ideal Solution)* model. The specific steps include:

- First, make scale transformations of IPR and DPS, and denoting them as a_{i1} and a_{i2}

$$a_{i1} = \frac{IPR_i}{\sqrt{\sum_i IPR_i^2}} \tag{7}$$

$$a_{i2} = \frac{DPS_i}{\sqrt{\sum_i DPS_i^2}} \tag{8}$$

- Among these limited plans, the optimal plan is $A^+ = (a_{i1}^+, a_{i2}^+)$ and the worst plan is $A^- = (a_{i1}^-, a_{i2}^-)$, where a_{ij}^+ and a_{ij}^- are the maximum and minimum values of the index j.
- Next, we calculate the distance between each evaluation plan and the best plan D_i^+, and also the worst plan D_i^-, where:

$$D_i^+ = \sqrt{w_1 (a_{i1} - a_{i1}^+)^2 + w_2 (a_{i2} - a_{i2}^+)^2} \tag{9}$$

$$D_i^- = \sqrt{w_1 (a_{i1} - a_{i1}^-)^2 + w_2 (a_{i2} - a_{i2}^-)^2} \tag{10}$$

- w_j is the normalized discrimination F_j of attribute j. And $F_j = 1 - E_j = 1 + \frac{\sum_i^m r_{ij} \ln r_{ij}}{\ln m}$, where m is the number of alternatives and $r_{ij} = \frac{a_{ij}}{\sum_i^m a_{ij}}$, then E_j is the information entropy of attribute j. when the values of attribute j of each report are the same, $r_{ij} = \frac{1}{m}$ does not make any difference. When the property value R_ The greater the ij difference, the greater the difference among r_{ij} is, the smaller E_j is, the more important the attribute j is to distinguish the good from the bad_ J , and the greater discrimination F_j and weight w_j is.
- Finally, the Relative Closeness Degree(RCD) of each evaluation object was calculated, i.e. $RCD_i = \frac{D_i^-}{D_i^- + D_i^+}$. The calculation formula shows that $RCD_i \in [0,1]$. The larger RCD is, the better the evaluation result is. Therefore, we calculate RCD of all unverified or unprocessed reports, and always give priority to the report with the highest RCD.

5.3 Real Case Analysis

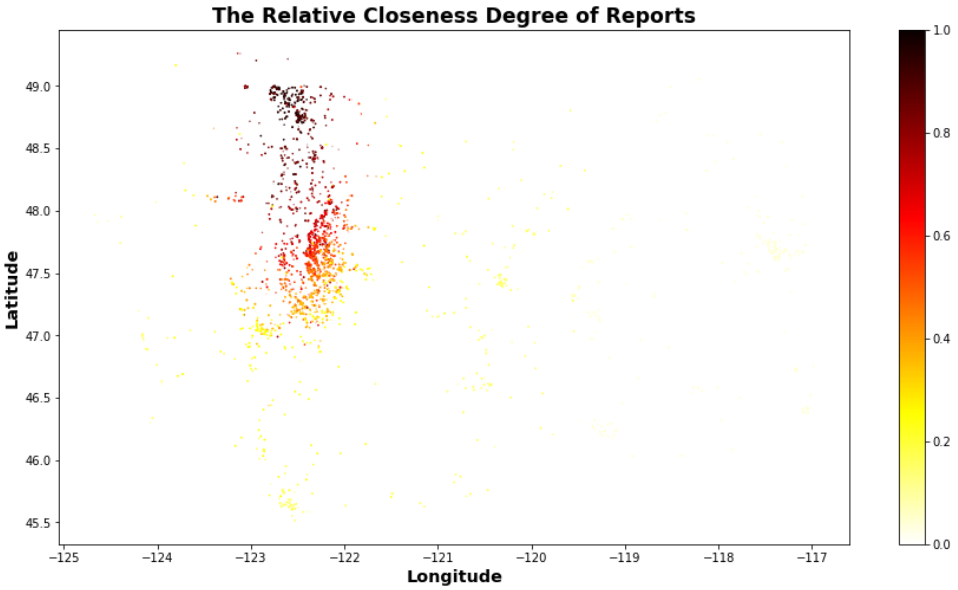

Figure 11: The Relative Closeness Degree of Reports

C_i of the original data distributes unevenly in [0,1]. In order to highlight the significant differences between levels and the geographical distribution characteristics of different levels while maintaining enough levels, we map Ci of the original data to the percentile in its array one by one. We then map the data between (10*k)-th percentile and (10*k+10)-th percentile to the value of (10k+5)/100 and use this value to make the figure above, so that we can Maximize the performance of the color bar.

From the figure above, we can see that through combination of IRF and DPS, the possibility of report being positive still shows obvious rationality. The percentage of positive cases reported in the two discovered activity areas is also significantly higher. For reports outside the active area, since there were no abnormalities in the southeast, we can temporarily infer that the Asian giant hornet has not spread in southeastern Washington.

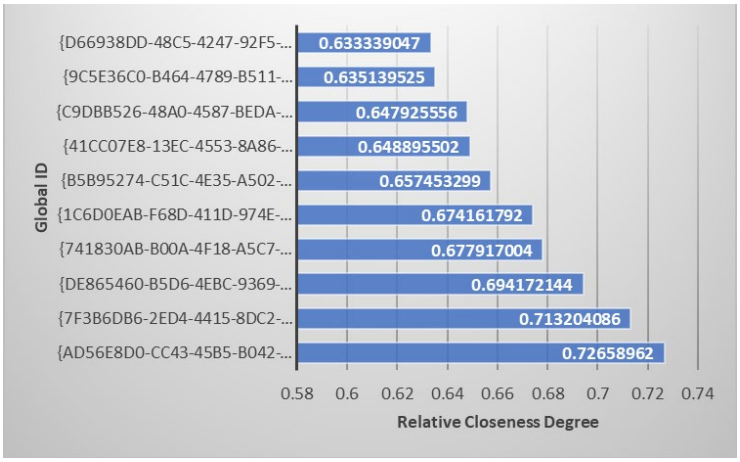

Figure 12: Relative Closeness Degree of Top 10 Reports

We have counted the ten reports with the highest Relative Closeness Degree with December 2020 as the observation point. According to the principle of prioritizing the largest report of Ci, the next report to prioritize has a global ID which is "AD56E8D0-CC43-45B5-B042-94D1712322B9".

6 Update of the Models

This part contains two subproblems, to address the approach to update our model given new reports over time and the frequency of updates.

Since the superior calculation efficiency of our model makes sure that it can handle with the spur of new reports under the current circumstances, there's no need to worry about the computation cost of the model which usually ends as a huge problem. Hence, all we need to care about the precision of our model's prediction. In short, given the new report we should take the updating mechanism of each model, timeliness and applicability in different regions into consideration.

6.1 Model Updating Mechanism

As to the first subproblem, the approach to update our model should be strictly based on the property of our model correspondingly.

For the VAR model used to predict the spread path and the Fourier series curve (number of terms=3) for estimating the number of the Asian Giant Hornets over time within a year, since they both require monthly data as a new sample, the updating frequency of these models should also go with the month or more. That is to say, by the end of this month, calculate this month's number of reports and pre-defined hornets' active location in each high-frequency area, and update the VAR model. And by the end of the year, after collecting all the monthly data in this year, update the Fourier 3 model.

For the classification model, the situation should be divided into two steps.

- First, when there comes a new report, the classification model will be just used to predict its possibility of being a positive sighting, and store the result into the sample database. By helping the report get the Relative Closeness Degree, the result will only have very little possibility to affect the priority of the next investigation.
- Secondly, when the report is marked "negative" or "positive", then the training set or test set of the classification model gets expanded, which will lead to the update of the classification model.

Hence, the frequency of the update of the classification model goes with the frequency of the emerging cases that a report about Asian Giant Hornet sighting is officially verified, which is equivalent of the lab status getting marked "Positive" or "Negative".

6.2 Timeliness

Usually the value of the data has strong positive correlation with its freshness, except those obvious belonging to the negative group by the result of classification model, every new report should be given more significance than the previous counterpart.

By affecting the pre-defined the average longitude and latitude of hornets' active location's with the lab status marked "unverified", the new report comes as a necessary shock to the estimated activity track

of the Asian Giant Hornet. Nevertheless, most of the shock will last a very short time since most of them will be marked "negative" soon and so the estimated location will come back to its original location.

Not every new report is overwhelmingly vital. The timeliness also accompanies with the seasonality of the report. Since most of the queens are hibernating under the ground, the sighting occurs in the winter will be less likely to be positive than other seasons.

6.3 Regional Applicability

Every time there comes the new report, we just need to add it to sample database. Only by the response from the lab, are we able to change the lab status of this report, and thus update the data in the sample database.

However, the discovery of an "outlier" positive sighting would change the whole model radically, which without doubt makes all the reports occurring around its location relatively more convincing then before. Since the construction of our model is heavily based on the assumption of the high-frequency area from the location of the verified positive sightings, our model tends to give much more attention to the place in the high-frequency area. Thus, when it comes to be a positive sighting that is out of the 30km reach to any already-defined high-frequency area, we should immediately take a look back at all the reports' data previously used, construct the brand-new high-frequency area around this "outlier", and rebuild all of our models except the Fourier curve.

7 Evidence for Eradication

7.1 Based on Previous Prediction

When the number of supplementary reports or supplementary positive reports in a certain month is 0, it can be considered that Asian giant hornets have been eradicated. The reasons for this are from two aspects:

- First, in the third part we predict the number of Asian giant hornets by building a *Fourier (number of terms=3)* model, which represents the relative relationship between the number of hornets in different months of a year. If the number of new reports in a certain month is reduced to 0, according to the model prediction, the number of new reports afterwards will also come close to 0. Then we can roughly conclude that the number of positive reports in the following months is close to 0, and the number of newly discovered hornets is also 0 as a result.
- Second, in the fourth part, we built a classification model to identify positive reports. If there is no positive report screened out by the model, we can conclude that the number of hornets is actually equal to 0.
- The results of these two models confirm each other and strengthen the conclusion that Asian giant hornets are eradicated in Washington.

Considering that observations from the public may be biased, and the received reports may not necessarily represent the actual situation, we can lengthen the observation time. For example, if there is no new report or no new positive report for three consecutive months, we can conclude that Asian giant hornets have been eliminated. Meanwhile, there may also be a time lag between the report and the destruction of hornets. Although it seems that there are no new positive reports in that month, there may

be historical nests or hornets that have not been destroyed. So in order to build a more effective feedback mechanism, the government should update the investigation information in time and strengthen the investigation of suspicious locations. In this way, the results predicted by the models would be closer to reality.

8 Conclusions

In this paper, we first discover two high-frequency area for Asian giant hornets to appear and then use *VaR* models to predict the future spread of the hornets based on these two areas. We find that in Area A, Asian giant hornets have a strong tendency to migrate to the southeast, while hornets in Area B will spread more toward the center. This suggests different regulation strategy for these two areas. Then we apply *Fourier (number of terms=3)* to calculate the number of Asian giant hornets in each month, and we discover that number of hornets will reach its peak in August, while decline to almost zero from November to February. Next, we discuss how to build a classification model and find that compared to language descriptions, images have more distinctions and are more useful for classification. Then, we use Image Positive Rate (IPR) and Distance Preference Score (DPS) as attributes to establish a *TOPSIS* model, this model helps us to understand how to prioritize investigation. Last, we discuss how to update the models above and how to define that Asian giant hornets have been eradicated.

MEMORANDUM

To: Washington State Department of Agriculture

From: Team #2126594

Date: Feb 9th, 2021

Subject: Predict the spread and buzz about Asian Giant Hornets in Washington

Dear governors of Washington State Department of Agriculture, we are honored to present our study and recommendations to you.

The impact of biological invasions on agriculture, biodiversity and human activities is always a global concern. Recently, Asian giant hornets have been found active in North America. In 2019, four hornet workers were discovered and destroyed in Washington, which brings a lot of concerns.

As we know, some hornets can have serious effects on local honey bees. Asian giant hornets switch from other prey sources to honey bees. Unlike Japanese honey bees, which have coevolved with Asian giant hornets, western/European honey bees are much more likely to be killed by the hornet. What's worse, when three or more hornets from the same nest attack the same honey bee hive, a "slaughter phase" may occur, which may pose threat to agriculture, biodiversity and honey bee industry. Besides, Asian giant hornets can also put harm on human's well-being when they are defending their nests or food source. They may deliver painful stings with venom, which may cause skin necrosis and hemorrhaging. Under such circumstances, it's critical to investigate into this problem and get control of the situation.

We've done comprehensive study on Asian giant hornets and our findings in detail is as below:

Prediction of the spread direction: Based on historical reports, we find that the earliest positive report happened in Vancouver, September 19, 2019 and its geographic location is more than 30km away from the other 13 positive sightings. This indicates the hornets found in this location have not spread to the Washington State, so we can put less focus on this area in further investigation. Then we discover two high-frequency areas (Area A and Area B) for Asian giant hornets to appear. The center of Area A's coordinates is (48.99°N, 122.70°W), which is reported on September 30, 2019. And the center of Area B's coordinates is (48.78°N,122.42°W), which is reported on June 12, 2020. Through VaR model, we estimate the spread path in Area A and B. We find that in Area A, Asian giant hornets have a strong tendency to migrate to the southeast in recent years. This indicates that in 2021, Area A should no longer be considered as a high-frequency area, and the corresponding center should also move to the southeast. While in Area B, hornets will spread more toward the center, so this area should still be considered important for investigation and control.

Prediction of the total number of Asian giant hornets: As to our calculation, the total number of reports is highly correlated to the number of positive reports, so by estimating the number of reports, we can get a view of how many potential positive reports there will be. Just like the number of Asian giant hornets which change in a seasonal pattern, we find that the total number of received reports also follows a seasonal change. We build Fourier model to estimate this change and this model can be used to roughly predict the number of Asian giant hornets. We also find that the number of reports is always close to 0 from November to December and from January to February, while reaches its peak in August. However, different from our intuition, the number of hornets doesn't follow a logistic expansion pattern. In fact,

hornets observed in June are less than those in May. As a result, government should adopt different strategies at different times and allocate resources flexibly.

Classification Model: After analyzing the dataset, at first we find that the ratio of positive and negative examples is about 1:150, which is extremely unbalanced and lead the classification model to prefer predicting the negative examples. Then we adopt SMOTE [8] to solve the problem of unbalanced sample category and find that the ratio of positive and negative examples is 1:1. We also discover that there is no obvious distinction between positive and negative examples in the same representation space. The reason for this is that the classification problem is fine-grained, that is, positive and negative classes belong to the same big category and have many same or similar features, which need to be distinguished through more detailed information. Therefore, the use of such textual data is likely to be unhelpful or even counterproductive to model performance. So we suggest that when building a classification model, more image resources should be provided in order to raise accuracy.

Ways to update models: In order to optimize our model to avoid its precision decreasing greatly with time, we must make great advantage of new reports to help us update the sample database as well as heighten the precision of our model. There are three aspects associated with the data from reports that need to be carefully paid attention to: model updating mechanism, timeliness(including seasonality) and regional applicability. Not only should we respect the properties of each model and customize their updating mechanism to specific data, we should also care about the fact that the reliability of the data used to update should follow the periodicity of the Asian Giant Hornets apart from its freshness. Moreover, any "outlier" positive sighting is a serious warning that the current model need great change immediately.

Evidence for eradication of this hornet in Washington: When the number of supplementary reports or supplementary positive reports in a certain month is 0, it can be considered that Asian giant hornets have been eradicated. Both our Fourier model and classification model can support for this. However, considering that observations from the public may be biased, and the received reports may not necessarily represent the actual situation. Meanwhile, there may also be a time lag between the report and the destruction of hornets. So in order to build a more effective feedback mechanism, the government should update the investigation information in time and strengthen the investigation of suspicious locations.

References

[1] Keeling, Matt J., et al. "Predicting the spread of the Asian hornet (Vespa velutina) following its incursion into Great Britain." Scientific reports 7.1 (2017): 1-7.

[2] Robinet, Christelle, Eric Darrouzet, and Christelle Suppo. "Spread modelling: a suitable tool to explore the role of human-mediated dispersal in the range expansion of the yellow-legged hornet in Europe." International Journal of Pest Management 65.3 (2019): 258-267.

[3] Alaniz, Alberto J., 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 (2021): 104-112.

[4] Wikipedia contributors. "Asian giant hornet." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 7 Feb. 2021. Web. 8 Feb. 2021.

[5] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[6] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009.

[7] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

[8] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

[9] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[10] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).