Problem Chosen
C

2021
MCM/ICM
Summary Sheet

Team Control Number
2124786.

Confirming the Buzz about Hornets: Spread Distribution Model and Sightings Classifier

Summary

The Asian Giant Hornet(AGH), which is viewed as a kind of pests since it is the voracious predator of many insects, appears in Washington State in recent years. To protect the local honeybee populations from the nest being destroyed by the Asian Giant Hornet, the State of Washington has created a website where the public can report sightings of this hornet. However, the sightings are not always about the Asian Giant Hornet. It has a lot of look-alike species, such as European Hornets. This leads to many classification mistakes existing in the public report data, which results in the waste of resources. A way to predict the correctness of the report will benefit the efficiency of the resource allocation.

To help predict the status, the AGH Distribution Model is built to simulate the pest's spread. The Model utilizes the Species Distribution Model concept that numerical magnitudes indicate the likelihood of a queen establishing a new nest at a certain location. At the same time, the Sigmoid Curve and Multivariate Normal Distribution variant is the Model's keys to imitate the general nesting behavior. And the three influential factors, the environmental factor, the human factor, and the natural enemies factor, improve the Model accuracy and precision.

Based on the AGH Distribution Model, the Sighting Classification System is created for Washington's government to screen reports most worthy of further investigation. A trained Random Forest Model fitted with up-sampling data is set as the inner algorithm for the System. After processing the input data into a certain form with its original features and a feature generated by the AGH Distribution Model result, the System will output the report's likelihood to be positive sighting.

The two models are applied to the unprocessed reports in the data set to check their accuracy. They give similar results, which demonstrate the correctness of the two models.

To find the potential positive cases, the Sighting Classification System firstly processes the new reports. A prioritizing investigation list is gained by sorting the probability of the reports to be positive returned by the System. The spread range given by the AGH Distribution Model provides a reference to adjust the list. The government can decide the allocation of resources according to the final priority list. Both the Model and the System are easy to use with high accuracy and low update requirements, which solve the government's problems.

Keywords: Asian Giant Hornet; Species Distribution Model; Random Forest

Contents

1 Introduction

1.1 Background

Asian giant hornet (Vespa mandarinia) is the world's largest species of hornets. (Simplify, we will call it "AGH" in the following paper.) In December 2019, the Washington State Department of Agriculture(WSDA) received and verified two reports of Asian Giant Hornet near Blaine. These are the first-ever sighting in the United States.

AGH will attacks and destroys honey bee hives. A few AGHs can destroy a hive in a matter of hours. The AGHs enter a "slaughter phase" where they kill bees by decapitating them. They then defend the hive as their own, taking the brood to feed their own young. They also attack other insects but are not known to destroy entire groups of those insects. They will attack people or pets when they are threatened. Their stinger is longer than that of a honey bee, and their venom is more toxic than other bees. They can also sting repeatedly.[3]

Due to if it becomes established, this hornet will have negative impacts on the environment, economy, and public health of Washington State. The State of Washington has created helplines and a website for people to report sightings of these hornet.

1.2 Problem Restatement

Build a model to advise the government on a strategy: How can we respond to public reports given the current resources of government agencies. According to the reporting data, we found that most of them are unverified. So, the problem firstly asking for an analysis of the spreading trend of the Asian Giant Hornets. We should then build another model to analyze the probability of classification mistakes in the reporting data, which also helps us lock the Asian Giant Hornet's nests.

Based on the two models we built, we can then present how we can prioritize the government resources to eradicate this species in the area and how we can know whether certain nests be eradicated or not.

1.3 Our Work

The original topic asks us to solve a Resource maximization problem. We got public suspecting AGH report data from the government, and we need to build our models to simulate the spreading of AGH. Then, to predict the mistaking report rate to help the government manage their investigation. We build two models to completer the task. The first one is the AGH Distribution Model; the second one is the Sightings Classification Models. In the AGH distribution model, we predict AGH's possible spreading based on the spreading of their nests. In this model, after assigning valid values to environmental factors, we can give each position in the map an approximate new nesting probability. In the Sightings Classification Model, we build it based on the data. We extract several features: detection date, location, imported images, note, and range. We will explain them one by one in the Sighting Classification Model section. Using this model, we can predict whether one report data is positive, negative, or unverified. We used this model to make predictions on the "unprocessed" data in the data file, and the result perfectly matches our spreading model, which is the AGHs Distribution Model. The Figure 1 below is the overview on what we did for solving the problem.

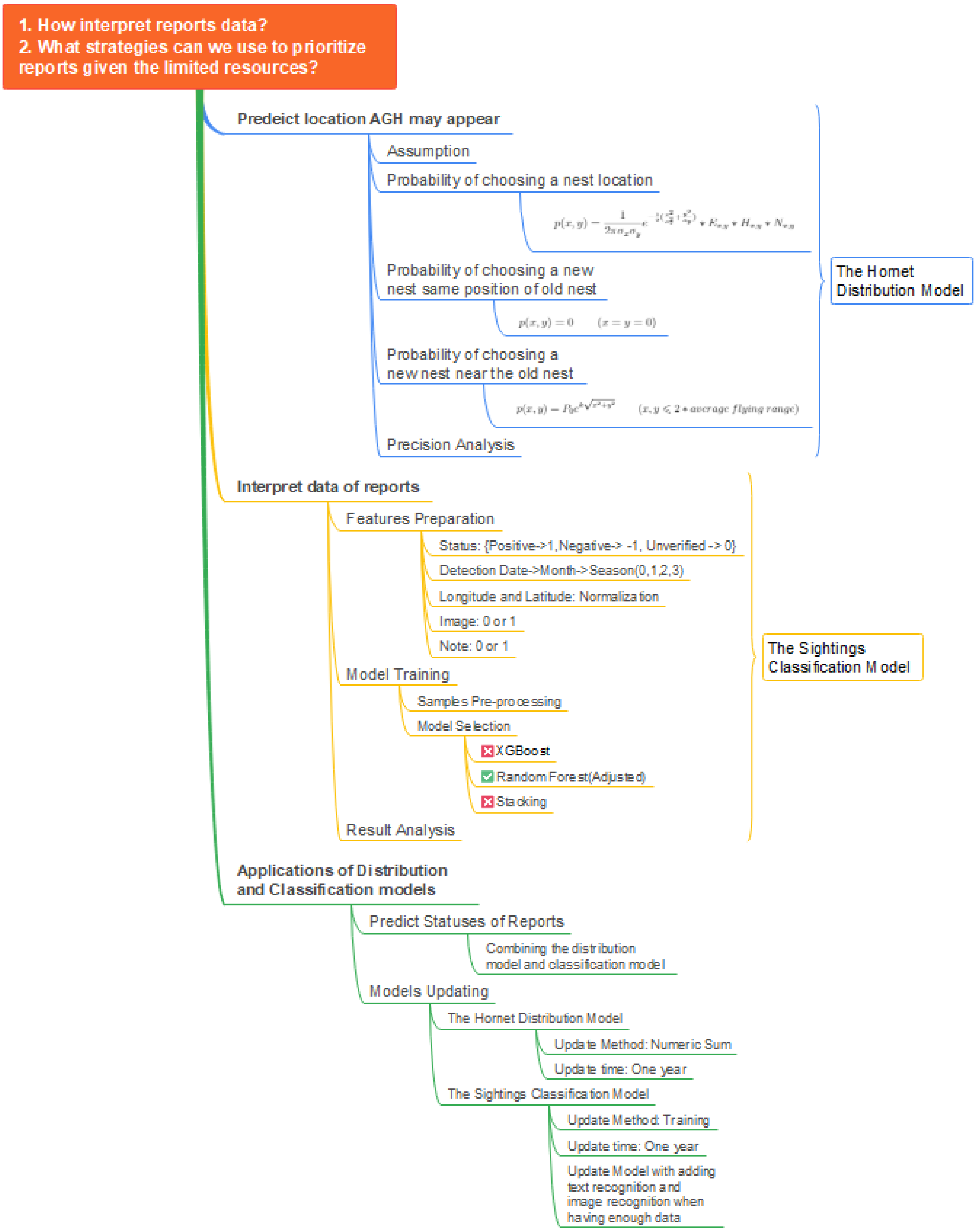

1. How interpret reports data?
2. What strategies can we use to prioritize reports given the limited resources?
Predeict location AGH may appear
Assumption
Probability of choosing a nest location
$p(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2})}*E_{x,y}*H_{x,y}*N_{x,y}$
Probability of choosing a new nest same position of old nest
$p(x,y)=0 \quad (x=y=0)$
Probability of choosing a new nest near the old nest
$p(x,y)=P_0e^{-k\sqrt{x^2+y^2}} \quad (x,y\leqslant 2*average\ flying\ range)$
Precision Analysis
The Hornet Distribution Model
Interpret data of reports
Features Preparation
Status: {Positive->1,Negative-> -1, Unverified -> 0}
Detection Date->Month-> Season(0,1,2,3)
Longitude and Latitude: Normalization
Image: 0 or 1
Note: 0 or 1
Model Training
Samples Pre-processing
Model Selection
☒ XGBoost
☑ Random Forest(Adjusted)
☒ Stacking
Result Analysis
The Sightings Classification Model
Applications of Distribution and Classification models
Predict Statuses of Reports
Combining the distribution model and classification model
Models Updating
The Hornet Distribution Model
Update Method: Numeric Sum
Update time: One year
The Sightings Classification Model
Update Method: Training
Update time: One year
Update Model with adding text recognition and image recognition when having enough data

Figure 1: Framework of Modeling

For the question of how to prioritize the investigation, we suggest that firstly apply the Sighting Classification Model to have a list of predicted positive cases and some unverified cases but with some probability to be positive. Then we apply our AGH Distribution Model to evaluate further and modify this priority list. We use our first model to test the priority of elements in the list. If the case is in the range of high possible new nesting position, then its priority goes up; otherwise, its priority decreases.

We come up with several important updating points to each model in view of how we should update our models. For the AGH Distribution Model, we should consider several nests' effects on one possible new nesting position as the positive data number increases. In this report, our first model predicts the spreading only based on the old nest's position since now the spreading extent is limited. We explained in later sections what it is like when several old nests make contributions together. For the Sighting Classification Model, we can update it by keeping collecting report data, especially positive cases. If there are more valid image information or text information, we can process them and increase our second model's accuracy.

For the question of what constitute the evidence that pest has been eradicated, We propose that we can use our model one to predict the probability of a new nest near the old nest's place. If the range of the most considerable likelihood new nest position does not have any positive confirmed report in several years, then it is a high possibility that that nest has already been successfully eradicated. Then, we will use our model2 to double-check the existence of the nest. Applying our second model, we can know the predicted result of every reported case. We will check the report cases near the old nest and get their probability to be a positive case. If the probability is high, then it is more likely we do not kill the nest completely. If the probability is low and all cases near the old nests were tested as negative reports, we more likely eradicated it.

2 Assumptions

The following assumptions were made to simplify the environment's complexity and eliminate unpredictable stochastic effects. All prominent assumptions and their rationality will be restated when they are utilized later in modeling sections.

- The spread of the pest can be interpreted by the dispersal of their nests.
- The time unit of the spread of the pest is one year, from spring to next spring.
- We regard two positive in-sights as two nests when their distance is beyond 2km.
- The probability of new nests that place outside of 30km range or at the same place with its old nest is 0
- There are approximate 70 fertilized queens produced in one old nest. Part of them will survive from winter and emerge in spring to built nest.
- Regarding massive loss of honey bees as indicator of the existence of AGH

3 List of Notation

Table 1: The List of Notation

Symbol	Meaning
x	New nest's location x-coordinate $-$ Old nest's location x-coordinate
y	New nest's location y-coordinate $-$ Old nest's location y-coordinate
$P(x, y)$	The probability of possible new nests locate at <x, y> from old nests
σ_X	Standard deviation of X
σ_Y	Standard deviation of Y

k	Model parameter that can be calculated case by case
P_0	The probability of possible new nest when x and y approach to 0
M	The peak probability of possible new nest
$E_{x,y}$	The environmental factor at distance <x, y> from old nests
$H_{x,y}$	The humanistic factor at distance <x, y> from old nests
$N_{x,y}$	The suppressing enemy factor at distance <x, y> from old nests

4 The AGH Distribution Model

4.1 Model Preparation

AGH's habits and characteristics are the direct influence factor of the spread of the pest. By summarizing and analyzing some vital habits, assumptions are made and shown in the following table (Table 2):

4.2 Modeling

Modeling Overview

This model gives a probability function on how AGHs spread across regions. The model imitates probability distribution that a location with a higher value represents the higher probability of being chosen as the nesting site. Taking the old nest as the origin of coordinates, we will calculate the probabilities of locations being chosen by a queen, produced in the old nest, to build her new nest. According to one AGH's vital habit that AGH queen can only find its new nest in thirty kilometers from its old nest, and data that shows most new nests are near the old one, we come up with the idea that the probability of a new nest is decreasing when their distance from an old nest to new nest is getting larger and larger. The probability approaches 0 when the distance is beyond thirty kilometers. Meanwhile, we also notice that the queen never place its new nest in the same place as its old one. Thus, when the distance approaches 0, the probability there approaches 0 as well.

Based on the above observations, we build a AGH Distribution model that follows the normal distribution basing on that AGH queens place their new nest randomly, but with the rules I mentioned above. However, we see some varies from normal distribution according to AGH's habit. First of all, the probability that the new nest is on the old nest is 0. Secondly, AGH queens prefer smaller distances. Thirdly, the density of the human population might have harmful effects on AGH's growth. Last but not least, AGH prefers specific environments. For example, according to (Penn extension), AGH queen tends to build a nest in the forest or rural regions. They cannot build the nest in water(river, lake, and sea).

According to the above observations, we build our model in three steps. We will discuss them in order in the following sections. We build a three-dimensional coordinate to express our model. The origin of our model curve is the AGH's old nest's position. The x and y denote the new nest's distance in-unit kilometers from the old nest in latitude direction and longitude direction. The z coordinate denotes the probability value of the new nest's possible position. Below are our three steps for building this model:

Table 2: AGH's habits and model assumptions

Habits description	Associated assumption	Addition note
Except for some queens, the life cycle of a AGH is starting in the spring and ends in the winter[1]	The time unit of our model is one year	The time unit has not been used in the AGH distribution model. The reason is given later in detail.
AGH can only be seen outside of the nest when they are hibernating or in the spring before workers have emerged.	The spread of the pest can be illustrated by the dispersal of the nest.	This assumption is the basis of our AGH distribution model
Only fly 1-2km on average and never more than 8km from the nest in search of food[5]	We regard two positive insights as two nests when their distance is beyond 2km	This assumption was used in the AGH distribution model we analyze data.
The queen can only find its new nest no more than 30km from its old die out nest, and the queen move its nest every year	The probability of new nests that place outside of 30km range or at the same place with its old nest is 0	This assumption was used in the AGH distribution model for depicting the distribution curve[8]
One AGH's nest had the potential to spawn almost 200 queens[4], where results in a large percentage (up to 65%) not being fertilized. Furthermore, one new nest will be built by one queen.	There are approximately 70 fertilized queens produced in one old nest. Part of them may survive from winter and emerge in spring to build a nest.	The truth gives a potential number of nests in the second generation inherited from one old nest.
Do worst damage to honey bee colonies that are less than 1km from the AGHs' nest	Regarding massive loss of honey bees as an indicator of the existence of AGN	This allows us to include data with ID. 13B67BCB-AFCE-4100-AD2B-76EF178BA228 as positive insight

Modeling Step One

We assume that without any impedes and consideration of the environmental concerns, the probability of choosing a nest location will be as the bivariate normal distribution that the nesting probability decreases with increasing distance. That forms our basic function:

$$p(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}e^{-\frac{1}{2(1-\rho^2)}[\frac{(x-\mu_x)^2}{\sigma_x^2}-2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}+\frac{(y-\mu_y)^2}{\sigma_y}]} \tag{1}$$

Since the setting of x and y are simulated to the location, the correlation ρ between X and Y should be 0. Furthermore, the old nest is the origin, which says $\mu_x,\ \mu_y$ also should be 0 in the function. The variances are depended on the movement range of the species. For the AGHs, since the queen can only find its new nest no more than 30km from its old nest die out, the variances should let the probability for places 30km far away from the old nest close to 0. Therefore, we could simplify our function into the new form:

$$p(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y})} \tag{2}$$

And its simulation is shown in the follow picture (Figure 2).

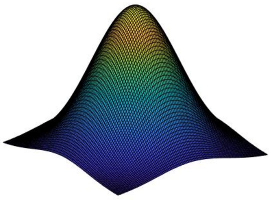


Figure 2: The simulation of the basic AGH Distribution Model

Later, by considering the environmental effects, the influences caused by human activities, and natural enemies, which are independent of the queen's random flight trajectory, the environmental factor $E_{x,y}$, humanistic factor $H_{x,y}$, and $N_{x,y}$ are adding into the basic function as multiplicative factors. The decision to apply multiplication rather than addition is made because of the product theorem. Here, the three factors $E_{x,y}$, $H_{x,y}$, and $N_{x,y}$ should be the evaluations of how do the surroundings affects the probability of the queen to nest given by the real situation of the nearby. For each x,y, $E_{x,y}$, $H_{x,y}$, and $N_{x,y}$ are independent and identical. All three factors should be numbers around 1 based on the value of the basic function. Because after multiplying the factors, the return value should not be bigger than 1.(Even though it does not really matter, controlling it not to be bigger than 1 helps people understand.) Thus, we gain our final function of step one:

$$p(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y})}*E_{x,y}*H_{x,y}*N_{x,y} \tag{3}$$

Table 3 are some possible value for the factors in the AGHs case for reference concluded in the given information. More precise value selection can be based on detailed topography, climate, population, and animal preferences.

Table 3: Possible values for the factors for the AGHs case

Examples for the three factors		
Factor	Location's situation	Suggesting value
$E_{x,y}$	A lake where AGHs won't nest	0
$E_{x,y}$	Forest	1.2
$E_{x,y}$	Snow mountain	0.5
$H_{x,y}$	A central business district (CBD)	0.7
$H_{x,y}$	Suburban district	1.1
$A_{x,y}$	A place near the Japanese honey bee's nest	0.6

Modeling Step Two

According to the information that the new nest will never be placed in the same place as the old nest, we derive that when the distance <x, y> from the old nest is 0, its correlating probability is 0.

$$p(x,y) = 0 \qquad (x = y = 0) \tag{4}$$

Modeling Step Three

We assume that the probability of possible new nests will increase under a certain distance when the distance increases. Based on one AGHs' habit, they never build new nests on their old ones to avoid resource disadvantages. Also, AGH's average flying range is 1-2km from their nests, and we assume that two average flying range from the old nest is the peak probability position of the new possible nest. The inference comes from the assumption that AGHs prefer that the new nest's position is closer to the old one but not too close to get enough food. So, step two is to find a model for distance from <0,0> to <2*average flying range, 2*average flying range>. In this range, the probability increases when distance increases. That forms us a second basic function:

$$p(x,y) = P_0 e^{k\sqrt{x^2+y^2}} \qquad (x, y \leqslant 2 * average\ flying\ range) \tag{5}$$

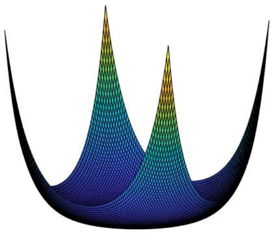


Figure 3: The simulation of the AGH Distribution Model's step three

We mirror the natural growth process to simulate our probability function model. When the distance is one average flying range, the probability increases the fastest. The range is when the distance is right beyond the AGH's average flying range from the old nest. Then it keeps increasing to approach the peak probability. We mirror the natural growth process in the case that P_0 is approaching 0, which correlates to the case in step 2, that the probability at the origin approaches 0.

We are not including the environmental factor($E_{x,y}$), the humanistic factor($H_{x,y}$), and the potential enemy factor($N_{x,y}$) here in this case since the distance range of this case is very short, it is only two average flying ranges from their old nest. Thus, the new position's external factors are very similar to the situation of their old nest. This means those external factors are optimal and tend not to affect the probability much.

In general, the following function is our complete model function:

$$p(x,y) = \begin{cases} 0 & x = y = 0 \\ p_0 e^{k\sqrt{x^2+y^2}} & 0 < x, y < \theta \\ \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2})} * E_{x,y} * H_{x,y} * N_{x,y} & x, y \geq \theta \end{cases} \tag{6}$$

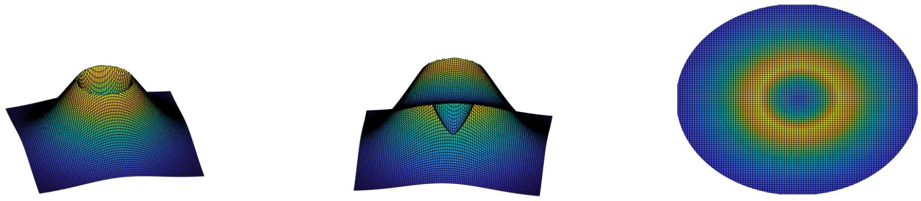


Figure 4: The three views(front, down, and top sides) of the simulation of the AGH Distribution Model where the three factors are all ones

4.3 Precision Analysis

There are three unknown factors in our model, which indicates the real-world situation. $E_{x,y}$ represents environmental suitability, $H_{x,y}$ represents whether and what effects human population factor might put on AGHs' nesting law, and $N_{x,y}$ denotes the effects by AGH's potential enemies. These are factors that might strongly affect the final probability in a real-world situation. These factors are important but mostly contribute to special situations. For example, $E_{x,y}$ is larger than when the environment powerfully meets AGHs' preference.

Because of the lack of information on AGH's habits, we will firstly ignore these three factors and test our precision level based on the function without them. These factors never hurt the precision if they were assigned to the right values. If we can get decent precision based on our general function without considering this three-factor, we can absolutely get more accurate precision when considering all factors.

We apply our model to the real data. According to our model, the peak probability holds when distances along x and y coordinates are two times AGH's average flying range, which matches the data. The figure below is the overall data distribution from 2019 to 2020. The one on the left denotes 2019 positive cases, and the one on the right denotes

2020 positive cases. We calculate the distances between the 2019's three positive cases, and they are all less than one average flying range, so we assume that these three cases are from the same nest. We calculate the distance between different positive points in 2020, and we compare those distances to AGH's average flying range to circle the approximate nest's position. As the figure shows below, all possible nests' locations are quite near the nest's position in 2019. We calculate it, and it is around two times AGH's average flying range from the old nest, which is the <x,y> location for the peak probability. Thus, our model explains 2020 data well.

When using the real-world model, the three factors we suggest affect the final probability much. They affect our precision a lot. Also, there might be many other real-world factors that affect the probability, and this needs more researches on AGH's habit.

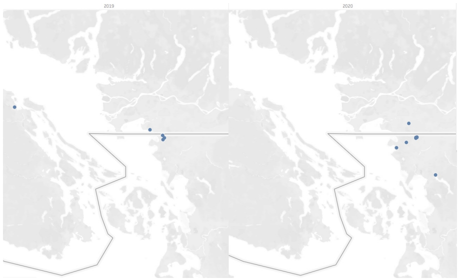


Figure 5: The positive cases in 2019 and 2020

5 The Sightings Classification Model

5.1 Model Overview

Learning from the previous data, the Sightings Classification Model built an inner algorithm to predict how the Lab will label a report based on the report's detection date, location, imported images, and note. Given a report, our Sightings Classification Model can output the probability of the report's three status, Positive ID', 'Negative ID,' and 'Unverified.'

5.2 Features Preparation

Every report contains 4 information: *Detection Date*: when the situation was found, *Longitude* and *Latitude*: where the situation locates, *Image*: how the situation looks like, and *Notes*: addition notes about the situation. In addition to the *Detection Date* and the *Longitude* and *Latitude* that each sample contains and is uniformly formatted, *Image* and *Notes* are more personal information with uncertainty. About the class, the lab labels each sample to four statuses: *Positive ID*, *Unverified*, *Negative ID*, and *Unprocessed*, which have 14, 2342, 2069, and 15 samples respectively (Figure 6). Of which, *Unprocessed* belongs to the invalid category. We drop all *Unprocessed* samples when we train our model. And only *Positive ID*, *Unverified*, and *Negative ID* are the target categories.

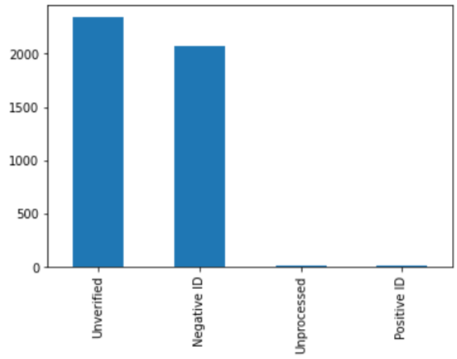


Figure 6: The distribution of *Lab Status*

The following table (Table 4) is the list of the features and the label we used in modeling and their descriptions. Detailed explanations of how we gain the features will be presented after the table.

Table 4: The Features and the Label used in our model

Features/Label	Type	Description
Season	Categorical	We classify the samples by their detection date based on the AGHs' life cycle. (0: March, April, May, June; 1: July, August; 2: September, October; 3: November, December, January, February)
Normalized Latitude	Quantitative	Since the change among the latitudes is inconspicuous compared to other features, we normalized it.
Normalized Longitude	Quantitative	The same as the longitude
Image	Boolean	Whether or not image(s) is(are) provided
Note	Boolean	Whether or not note is provided.
Range	Boolean	Whether or not the report is located in the high probability range of discovering the AGHs
Label	Categorical	1: Positive ID; 0: Unverified; -1: Negative ID

Image

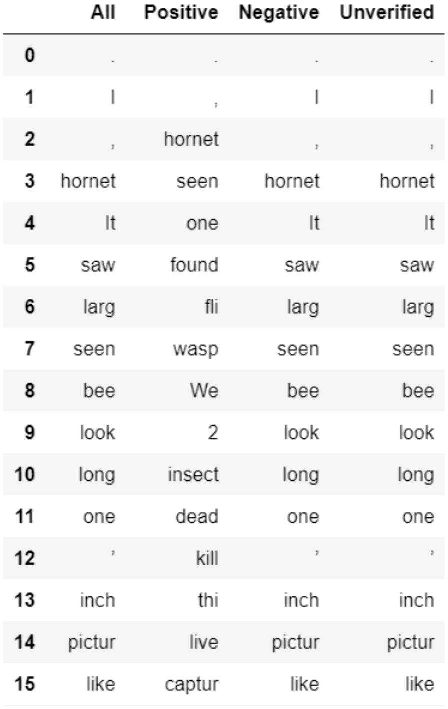
After browsing all the images, we found that most of them are of low quality resulting in the inability to obtain valid information by the computer reading the image content. Therefore, we will not consider the effect of image content on classification. Also, we found that there seems to be some connection between the availability of images and classification (shown in Table 5). Therefore, we will finish one of the features: whether the reporter provides images or not.

Table 5: The two side table of *Lab Status* and *Image*

	Positive ID	Unverified	Negative ID
Image provided	11	73	2043
No image provided	3	2269	26

Note

We grabbed all the notes information and split each paragraph into individual words. After removing all word variations, we counted the frequency of all words. We found the 50 used words with the highest frequency within each class, respectively(The below figure only shows the top 15 limited to space).



	All	Positive	Negative	Unverified
0	.	.	.	.
1	I	,	I	I
2	,	hornet	,	,
3	hornet	seen	hornet	hornet
4	It	one	It	It
5	saw	found	saw	saw
6	larg	fli	larg	larg
7	seen	wasp	seen	seen
8	bee	We	bee	bee
9	look	2	look	look
10	long	insect	long	long
11	one	dead	one	one
12	’	kill	’	’
13	inch	thi	inch	inch
14	pictur	live	pictur	pictur
15	like	captur	like	like

Figure 7: The top 15 used words within different classes

We found that the words with the highest frequency appearing in different dictionaries were basically the same. This indicates that the reporters' information on remarks basically cannot be used as a criterion for classifying the samples. Therefore, we do not consider further processing of textual content. Instead, we only take if the reporter provides notes as a feature.

Range

Based on our AGH Distribution Model's general result, addresses close to areas where target AGHs have been present will be more likely to show positive results. Therefore, we designed a new feature: whether the sample occurred within 5 km of a location where positive reports had occurred in the past, where 5 km is the maximum radius of AGHs' activity range.

5.3 Model Training

5.3.1 Samples Pre-processing

The past reports gives imbalanced samples that the number of positive cases is far less than the number of negative and unverified cases. The insufficient positive data would affect the learning of positive behavior. To avoid making an invalid classifier, we re-sample our data by up-sampling the positive cases. [2]

5.3.2 Model Selection

We tried multiple algorithms for our training. We chose the XGBoost algorithm at first. The XGBoost algorithm can enhance the predictive model because XGBoost adds a regular term to the cost function to control the model's complexity. The regularization term contains the number of leaf nodes of the tree and the sum of squares of the absolute value of L2 of score output on each leaf node. From the perspective of the bias-variance trade-off, the regularization term reduces the model's variance, which makes the learned model simpler and prevents overfitting. However, the XGBoost depends on the "Image" feature too much.

Then, We tried random forest algorithm. Random forest algorithm is a well-known ensemble learning method that integrates multiple trees through ensemble learning. Its basic unit is a decision tree algorithm. "Random" has two meanings in this algorithm. One is to randomly select the same amount of data from the original training data as the training sample; the other is to randomly select some features from the full features to establish the decision tree. These two kinds of random procedures reduce the correlation between the decision tree and further improve the models' accuracy. Due to the imbalance amount between labels, we adjust the parameter in the random forest algorithm. We use "balanced" mode to modify class weight. The "balanced" mode uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data. Thus, we can reduce the impact of imbalance classification.

At this point, we already have two individual models, and both of them have pretty good performance. Then, we want to combine them using Stacking Ensemble Machine Learning. Stacking is a method that can combine the results of other individual machine learning blocks. The stack uses the predictions of some basic classifiers as the first level (base), and then at the second level uses another model to predict the output of the earlier first-level predictions.

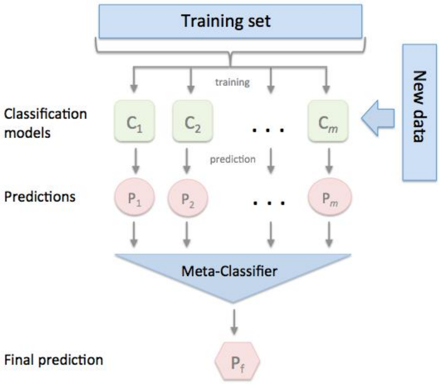


Figure 8: Schematic of a stacking classifier framework.[6]

The performances of adjusted XGBoost algorithm, random forest, and Stacking shows below:

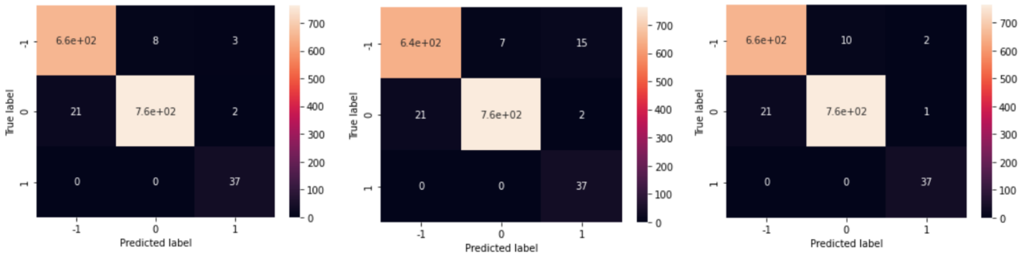


Figure 9: From the left to right are confusion matrix of adjusted random forest, XGBoost algorithm, and Stacking in test set

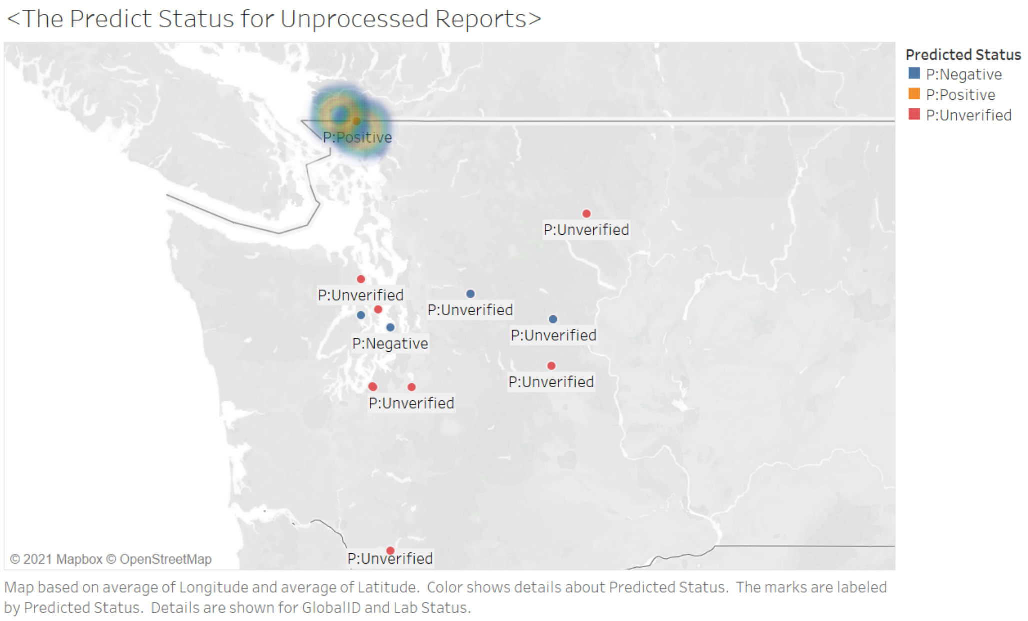
5.4 Result Analysis

We chose adjusted random forest as our final model after comparing those three models' prediction on unprocessed data and the trade-off on weights of features. As shown in Figure 9, it is obvious that all three models have well performances on the test data set. However, they have different weight on features. For example, the "image" feature has up to 90 percent of the total features. However, we are looking for a model that applies as many features as possible instead of only need one feature. The adjusted random forest model is the model we want. Its weight is more balanced on features than other models. Although the adjusted random forest model has a bit of lower accuracy, it also means a lower risk of overfitting.

We used our model to predict the data labeled "unprocessed." Our output includes the probability of each label and the type for each data. We also drew our prediction on the map. The result is shown on Figure 10. The positive point we predicted is near to three positive points. It corresponds to our AGH Distribution Model since there should be a nest according to our assumption, and it should spread from the hive last year. The circle in the image is the area we predicted for possible nests in the AGH Distribution Model.

We used our model to train the data without the "Image" feature. The accuracy was reduced significantly. Associated with the given data, most of the given image data were determined as "negative," and only several ambiguous pictures or videos were determined as "unverified." Most of the data labeled "unverified" do not have matching images. Therefore, it is understandable why the "image" feature is such important in our models. Thus, providing images is essential for Lab to identify AGH.

The next most important feature is location, to be specific, latitude and longitude. This is also explicable by our assumption. Since the AGH could not be far from their nest, we can speculate about the area they might appear to confirm their nests' location. "Season" is the third most important feature since AGH only appears in certain seasons. However, its habits may be similar to other bees, so that this feature is relatively small in proportion. For example, most bees and wasps hibernate during the winter, then the amount of sightings reports will decrease in winter. Therefore, "season" is not a powerful feature to distinguish AGH and other bees. "Text" is the least important feature. It may be caused by our method when processing this feature. While preparing data, we only focus on whether the reporter provides "Note," but most reports include "note," and its


<The Predict Status for Unprocessed Reports>

Map based on average of Longitude and average of Latitude. Color shows details about Predicted Status. The marks are labeled by Predicted Status. Details are shown for GlobalID and Lab Status.

Figure 10: The Prediction Status for Unprocessed Data: the circles in the image is the area we predicted for possible nests in the AGH Distribution Model

distribution is similar to total data, which means it is hard to classify data using this feature.

Overall our model did a good job and provided an accurate result of predicting. The model has a higher weight on the features "Image" and location that applies more features than other models. In the test data set, our model answers with great accuracy. When we utilized "unprocessed" data, we also get answers corresponding to our model made before.

The weakness of our model is the lack of data. Although we applied several techniques to features in data preparation, it cannot make up we only have 14 labeled positive data in the size of 4440 datasets. The model depends on the "Image" feature too much that is also caused by the unbalanced data. The devastating unbalanced data likely make our model overfitting so that the model still needs to train with more data in the future to improve the data performance of prediction when inputting new data.

6 Applications of Distribution and Classification models

6.1 Predict Statuses of Reports

There are two steps for government agencies to find the prioritizing cases and decide whether they should keep follow-up with the additional investigation: applying the Sighting Classification Model on the reports' data to predict the probability of the cases' status and then using the AGH Distribution Model to verify the results given by the first step.

Applying the Sighting Classification Model: Following the feature preparation steps, the data will be changed into a form that the model can identify. And then, the deformed data can be processed with the trained Sighting Classification Model, which is provided in our paper, and the model will generate the probability that each sample will be judged as positive. By sorting the cases with their probabilities of being positive sightings, the priority list to investigate is formed. Additionally, the next section provides information for updating the model.

Verifying the probability by using the AGH Distribution Model: By checking if the cases with high probabilities to be positive are located in the high possible region given by the AGH Distribution Model, the priority list's correctness will be improved. Once a case in the list is not located in the region, its priority can be reduced.

6.2 Models Updating

The AGH Distribution model gives a probability density function on how AGHs spread across regions. We can predict the location of nests of AGHs using given additional data; then, we can apply our distribution model to calculate the probability of how AGHs spread across regions over time. Our model only focuses on the probability of the region near one nest, but as time pass, AGHs may spreads, so we will use the numeric sum to compute the probability density given a location if it is within the range of multiple nests. The model's update time should be one year since AGHs only build a new nest in spring every year. So AGHs only spread once over one year.

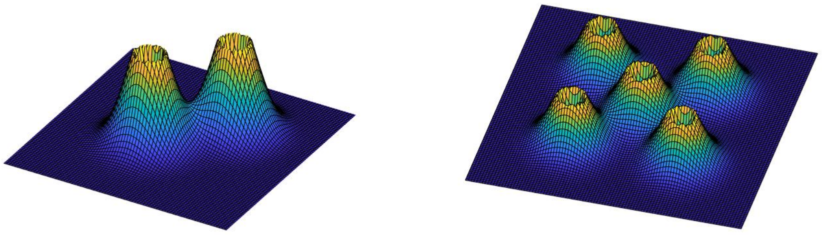


Figure 11: Figure of Interference by Multiple Functions

Our classification models can be updated by training with given additional new reports over time. The model can improve by applying image data and text data in "note," but it is required we have enough data labeled positive. As the AGHs spread and we have more data, we can use neural networks to process image data to achieve image recognition and reduce lab workload. Meantime, we can adopt text recognition techniques to explore whether the reporter's note affects classification. The classification model's update time also should be one year since the location feature is essential on our model, and AGH's active range also update once a year. AGHs' actions are around their nest so that their active range is certain when they have built their new nest. However, we should update our model to implement text recognition and image recognition when collecting enough data.

6.3 Evidence for eradication

In the last application of our models, we try to answer this question: what would constitute evidence that the pest has been eradicated in Washington State?

Here pest was eradicated means the government department adopted some methods to destroy the nests and kill the spreading. However, we need ample evidence to show that our destroy methods are valid. Both our AGH Distribution model and Sightings

Classification model give some evidence.

According to our AGH distribution model, when the distance range is 0 to 2*average flying range, the new nest probability increases with distance increases; while the distance range is 2*average flying range to 30, the probability decreases with distance increases. Thus, we can use old nests' position to predict the position of the new nests. When we try testing whether a certain nest being eradicated or not, we can use our first model to derive some position ranges that new nests most likely positioned. If none of them has any new nests located and not any positive cases being reported around those positions in the following years, we can say that the nest is already be eradicated.

Our Sightings Classification Model can give a prediction on the result of every report. In other words, it can tell us whether certain reports most likely to be positive cases, negative cases, or unverified cases. Applying this model to this question, we can suspect that if the predicted reports around the old nest are all negative cases or unverified cases, then it is highly possible that the old nest has already been eradicated.

7 Strengths and Weaknesses

7.1 Strengths

- The AGHs Distribution Model predicts new nests not only based on the randomness of AGH's nesting but also on their most apparent habits. It accurately predicts 2020's data based on 2019's data when we are not taking full considerations special situation factors such as environmental and humanistic factors. After assigning accurate value to these factors, our precision can even go higher.
- Although we have not found much information on the potential enemy of AGH, we make a factor ($N_{x,y}$) to represent the potential effects from the potential enemy for higher fault-tolerance.
- The Sighting Classification Model utilizes the AGHs Distribution Model's outcome to create a new feature.

7.2 Weaknesses and Extensions

- May lack of considerations on other potential external factors that affects AGH's law of nesting.
- Lack of real data to confirm our assumption that AGH's law of nesting follow normal distribution in general.
- Lack of considering how much every special situation factors($E_{x,y}$, $H_{x,y}$, $N_{x,y}$) and how different they affect the probability.
- Lack of data when training our Sightings Classification Model, especially positive report data.
- Data quality can be better. In our study, three labels ("Positive", "negative" and "unverified")'s amounts are hugely different.

Optimization method: Anytime we find a nest, do some detailed research on what and how exactly the environmental factors affect AGH's nesting law. Then we can modify and specify our external factors($E_{x,y}$, $H_{x,y}$, $N_{x,y}$) to better solve real-world situation. Also, we should keep collecting the data and require every report to include images that are as clear as possible to get more valid information.

8 Conclusion

To help the government rationalize the allocation of limited resources to deal with the AGH problem, we built two sets of models to predict and analyze whether the public's AGHs reported positive cases. Our models well simulate the spread of wasps as well as accurately predict the statuses of reports.

1. The spread of the pest over time can be predicted by the AGH Distribution Model as it simulates how the AGH queens seek place to nest.

2. With information provided by the public, the Sighting Classification Model can predict the chance of a case to be positive sighting which illustrates the likelihood of a mistaken classification.

3. Utilizing the Sighting Classification Model to find the prioritizing investigation list and verifying it with the AGH Distribution Model can help government agencies allocate their resources.

4. Since the time unit of the AGHs to nest is one year, the AGH Distribution Model and the processing of feature *Range* in the Sighting Classification Model are suggested to be updated once a year while the Sighting Classification Model can be retrained every time when a new case has been labeled.

5. Once the range given by the AGH Distribution Model has no positive case found and the Sighting Classification Model 's prediction show no case would be positive, we can assume the pest has been eradicated in Washington State.

Memorandum

To: the Washington State Department of Agriculture
From: MCM Team 2124786
Date: February 8th, 2021
Subject: Suggestions on how to interpret public report data on Asian Giant Hornets to kill the spreading

I am writing to give some suggestions on interpreting public report data on the existence of the Asian Giant Hornets. In addition, based on the interpretation, how to prioritize limited resources to do further investigations and kill the spreading. We build two models to give some insights into the problem. The first one simulates and predicts the possible spreading of the Asian Giant Hornets, while the second one predicts certain report is more likely to be the positive case, negative case or unverified case. Both of them provide many insights on how to help with killing the spread of the Asian Giant Hornets.

Our first model is called the AGH Distribution Model, which predicts the new nesting probability on every position in the map. Based on this, we can simulate the possible spreading track of this Asian Giant Hornets. We test our model one to the real data, and they perfectly match.

Our second model is called the Sightings Classification System, which runs on every report data and can result on their most likely labels. In other words, we will know the certain data most likely is labeled positive case, negative case, or unverified case.

Applying these two models, we come up with a process that helps listing nests' approximate positions from high priority to low priority. The following is the complete procedures:

1. After getting the raw report data from the public, we first apply it with our second model. We can get a list with predicted positive cases on the top, then unverified cases, then negative cases.
2. We apply our first model on the list. Firstly we pay attention to the predicted positive cases. If the case is in a high nesting probability position, its priority increases; otherwise, its priority drops.
3. We then go looking at unverified and negative cases and check if they locate at some places that the hornets prefer. For example, the forest or somewhere has pines. If they locate at these places, we will apply our first model to these cases to see how their priority change.
4. After we were finishing the priority list, we get the cases that we are most interested in. We can then go through all these cases to try finding the nests. If we have extra resources, we can get through some unverified cases, too, since the hornets' law of nesting varies a lot by many conditions, and our model can only give a probability.

We also propose a guide on updating our model to interpret better the data comes in the future. Here is the guide for both of the models:

- We can consider more than one old nests' position's effects on possible new nests when we get more lab positive data for our AGH Distribution Model. In other words, we can extend our AGH Distribution Model to AGH Joint Distribution Model for better prediction. After more data is coming, we can get a better insight into how external environment factors work. This helps our model works better in real-world situations.
- For our Sightings Classification System, we can update our model when new data comes in, especially lab positive cases. Also, suppose data comes with valid images and notes. In that case, we can do further processing on these two features for better prediction.

We also have some advice on how to make data more efficient. Here they are:

- When we were building our second model, we noticed that image is the essential factor for defining the data's possible status. Thus, if most data come with a clear image, our model can be more efficient.
- When we trained our model on the current data, we noticed that we hardly can get any useful information from the notes. Thus, this might be the part where data can improve. If data comes with more detailed information, our model can improve as well.

That's all we have for solving the resources-optimization problem. We really appreciate your time in reading this. Please feel free to contact us if there is any question. Hope we can have a chance to cooperate in the future.

References

[1] Exotic pests. (n.d.). Retrieved February 08, 2021, from https://beeaware.org.au/archive-pest/asian-AGH/ad-image-0

[2] How to handle Imbalanced classes in machine learning. (2020, May 23). Retrieved February 08, 2021, from https://elitedatascience.com/imbalanced-classes

[3] Invasive AGHs. (n.d.). Retrieved February 07, 2021, from https://agr.wa.gov/AGHs

[4] Madani, D. (2020, November 10). 'Murder AGH' Nest had potential to spawn almost 200 queens in Washington state. Retrieved February 07, 2021, from https://www.nbcnews.com/news/animal-news/murder-AGH-nest-had-potential-spawn-almost-200-queens-washington-n1247302

[5] Michael J. Skvarla Assistant Research Professor of Arthropod Identification Expertise Arthropod identification Arthropod survey. (2021, ry 17). Asian giant Hornets. Retrieved February 08, 2021, from https://extension.psu.edu/asian-giant-hornets: :text=However%2C%20ASIAN%20giant%20hornets%20only,nests%20further%20away%20may%20be

[6] Raschka, S. (n.d.). StackingClassifier. Retrieved February 07, 2021, from http://rasbt.github.io/mlxtend/user_guide/classifier/StackingClassifier/

[7] Sofaer, H., Jarnevich, C., Pearse, I., Smyth, R., Auer, S., Cook, G., . . . Hamilton, H. (2019, June 05). Development and delivery of species distribution models to INFORM DECISION-MAKING. Retrieved February 08, 2021, from https://academic.oup.com/bioscience/article/69/7/544/5505326

[8] Zach Barth; Thomas Kearns; Elizabeth Wason. (n.d.). Animal diversity web. Retrieved February 08, 2021, from https://animaldiversity.org/accounts/Vespa_mandarinia/36D30A73-3A4F-11E2-ABAB-002500F14F28

[9] Zhu, G., Illan, J., Looney, C., amp; Crowder, D. (2020, October 06). Assessing the ecological niche and invasion potential of the Asian Giant Hornet. Retrieved February 08, 2021, from https://doi.org/10.1073/pnas.2011441117

Appendices

The following are the python code we used for the Sighting Classification Model:

```
#!/usr/bin/env python
# coding: utf-8

import numpy as np
import pandas as pd
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

pip install openpyxl

data = pd.read_excel("/kaggle
                     input(/mcm2021c/2021MCMProblemC_DataSet.xlsx"))
data_image = pd.read_excel("/kaggle/input
                           mcm2021c(/2021MCM_ProblemC_, Images_by_GlobalID.xlsx"))

# ## Feature Exploration & Preparation

# Image
a = list(data_image["GlobalID"].unique())
data["Image"] = 0
data.loc[data["GlobalID"].isin(a),['Image']] = 1

# Detection Date -> Season
data['Date'] = pd.to_datetime(data['Detection Date'], errors='coerce')
data['month'] = data['Date'].dt.month
data['Date_4'] = 0
data.loc[data['month'].isin([11,12,1,2]),['Date_4']] = 3
data.loc[data['month'].isin([7,8]),['Date_4']] = 1
data.loc[data['month'].isin([9,10]),['Date_4']] = 2

# Notes
data['Note_tf'] = 1
data.loc[data['Notes']==' ',['Note_tf']] =0

# Label
data['Label'] = 0
data.loc[data['Lab Status']=='Positive ID',['Label']]=1
data.loc[data['Lab Status']=='Negative ID',['Label']]=-1

# Location
from sklearn import preprocessing
data['Latitude_n'] = preprocessing.scale(data['Latitude'])
data['Longitude_n'] = preprocessing.scale(data['Longitude'])

# Range
lat = data.loc[data['Lab Status']=='Positive ID',['Latitude']]
long = data.loc[data['Lab Status']=='Positive ID',['Longitude']]
lat['min']=lat['Latitude']-0.05
lat['max']=lat['Latitude']+0.05
long['min']=long['Longitude']-0.05
long['max']=long['Longitude']+0.05
```

```
def cl_lat_long(lat, long, p_lat, p_long):
    for i in p_lat.index:
        if ((p_lat.loc[i,’min’]<lat)
            and (lat<p_lat.loc[i,’max’])
            and (p_long.loc[i,’min’]<long)
            and (long<p_long.loc[i,’max’])):
            return 1
    return 0

data[’Range’] = 0

for i in data.index:
    data.loc[i,’Range’] = cl_lat_long(data.loc[i,’Latitude’],
                                      data.loc[i,’Longitude’],lat,long)

# drop all Unprocessed
new_data = data.loc[data[’Lab Status’]!=’Unprocessed’]

#upsampling
from sklearn.utils import resample
df_majority = new_data[new_data.Label!=1]
df_minority = new_data[new_data.Label==1]
df_minority_upsampled = resample(df_minority,
                                 replace=True,
                                 n_samples=100,
                                 random_state=42)
df_upsampled = pd.concat([df_majority, df_minority_upsampled])

# ## upsampling_Random Forest

from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics
import collections
from sklearn.metrics import plot_confusion_matrix

uX_train, uX_test, uy_train, uy_test = train_test_split(
    df_upsampled[[’Latitude_n’,’Longitude_n’,
                  ’Image’,’Note_tf’,’Date_4’,’Range’]],
    df_upsampled[’Label’], test_size=0.33, random_state=42)

uclf = RandomForestClassifier(n_estimators=50,
                              max_depth=5,
                              class_weight=’balanced’,
                              random_state=42)
uclf.fit(uX_train, uy_train)

# ## Predict unprocessed samples

unpro = data.loc[data[’Lab Status’]==’Unprocessed’]
La_unpro_proba = uclf.predict_proba(
    unpro[[’Latitude_n’,’Longitude_n’,
           ’Image’,’Note_tf’,’Date_4’,’Range’]])
La_unpro_proba
```