Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2126887

In this report, our main goals are: 1) prioritize the follow-up investigation of new reported sightings of Asian giant hornets and 2) predict the location of their nests in the near future.

To prioritize the follow-up investigations, we propose a majority-vote scheme that predicts whether a reported sighting is likely to be verifiable and positive by analyzing previous labelled data based on the following three dimensions:

1. **Text** We performed the logistic regression on texts associated with each reported sighting and predict whether a report is verifiable or unverifiable. Specifically, the verifiable samples consist of both positive AND negative labels, while the unverifiable ones consist of unverified labels. This classifier gives us a testing accuracy of 70.6%.
2. **Image** The images are used to differentiate between verifiable and unverifiable submitted reports. Information based on the images of each report consists of two sub classifiers:
   (a) **Image Availability** The availability of Images corresponds to a higher likelihood of the report being verifiable.
   (b) **Image Classification** For the reports with at least 1 images attached, we applied the InceptionV3 model to predict whether an image is more likely to represent a verifiable reported sighting or not. This classifier gives us a testing accuracy of 68.2% .
3. **Location** We analyze the location of previous positive reports to estimate a probability density function of where nests will occur on a given year in the future by using the Monte Carlo method.

While the method of combining the information sources above can be flexible, we propose a **majority-vote scheme** with total 5 votes to combine these classifiers as follows:

- The text classifier has 1 vote. If the text classifier predicts a new report to be verifiable based on its attached note, it votes *Yes*.
- The image classifier has 2 votes, one for Image Availability and one for Image Classification.
- The location analysis has 2 votes and the number of *Yes* votes depends on the likelihood that the classifier outputs.

In sum, a new reported sighting will get examined by lab if it received at least 3 votes. The probability of misclassifying a verifiable sample as unverifiable is between 0.03% and 48.3% depending on whether or not they test positive in the location model.

Our model provides some additional insights. For example, the government might want to increase the probability of a verified report among all reported sightings by asking mandatory images or simply adding more prompts when the public tried to upload a text-only-report.

Contents

1 Introduction

In this report, we consider the problem of predicting Asian giant hornet nests given the data of labelled public report. More specifically, our work can be seen as the combination of the two following tasks:

- A majority-vote based model which prioritized the government's limited resources when a new public report is uploaded;
- A prediction model which indicates the likelihood of Asian giant hornet appearing in a certain region near previous positive spot.

Generally speaking, we analyzed the location, texts, and images based on previous labelled public reports to build one classifier for each. The detailed explanation for these classifiers can be found later.

1.1 Background

Asian giant hornet, originated in East Asia, is the largest hornet species in the world. Starting from 2018, however, they have been frequently reported in United States (Washington) and Canada (Vancouver island) as invasive species. Asian giant hornets are harmful to the local environment because they feed on European honey bees and therefore causing huge loss in honey production [1][2][6]. In addition, their stings are also toxic and have been reported to injure or even kill people [3][4][5].

To settle the Asian giant hornet issue, many past studies have devoted to better analyzing their effects on the ecosystem and predicting their expansion in North America and Among them, Alaniz , Carvajal, and Vergara[1] analyzed the National Agriculture Statistics data and found that the Asian giant hornet colonies could potentially spread to the east of Washington and Oregon state and threaten around 95216 honey bee colonies leading to an economic loss of $114 million. Another study [8] modeled the niche of Asian giant hornets as regions with relatively warm temperature and high precipitation. By doing dispersal simulation taking environmental factors into account, they concluded that the Asian giant hornets could prevail along the coast line of British Columbia, Washington, and Oregon, if not controlled properly [8].

1.2 Problem Statement

The key motivation lies in the contradiction between the limited government resources and the huge amount of mistaken public reports. Given the previous labelled public reports, we want to build a model which helps prioritize investigation of the reports. To be specific, the model should be able to predicts the likelihood of a mistaken classification. Furthermore, since the Asian giant hornets might propagate each year and move around, we want to predict the spread of the pest over time with high precision. This requires taking the geographical constraint into consideration. For example, the Asian giant hornets are not able to live on the ocean, and this fact should be considered.

Our model provides logically coherent and statistically reasonable solution to all the above problems. Along the way, we will discuss how our model can be updated given new reports over time. We shall also discuss what might constitute evidence that the Asian giant hornet has been eradicated in Washington state.

1.3 Assumptions

We make the following assumptions in our models:

1. Wind does not affect the motion of Asian giant hornets;
2. Asian giant hornets can equally likely inhabit anywhere on the earth if no initial distribution is given;
3. There is no aided dispersion of Asian giant hornets;
4. distribution;
5. The spread of Asian giant hornet nests is discretized into period of a year;
6. The reported sightings in the dataset are assumed to be in the same year period;
7. The prediction of nest locations of Asian giant hornets are only based on the reported sightings of Asian giant hornets.

1.4 Our Model

Our model consists of three classifiers built from the analysis of location, text and images:

1. **Location Classifier**
   We analyzed the location of the confirmed positive sites of Asian giant hornets to predict the location of the hornet nest based on the probability density function for the location of a nest in the following year. This model aims to create a probability density of where nests will appear in future years. It is also used as an index of measurement to order the testing under limited resources. In addition, this model is also used to predict the spread of Asian giant hornet nest in the future.
2. **Text Analysis**
   Text analysis relies on the frequencies of words in the Note column of the dataset of reported sightings to predict whether there is enough information for the lab to make a classification on the sample. This analysis is used as an index of measurement to order testing priority.
3. **Image Analysis**
   Image availability model predicts the likelihood for a sighting to get successfully classified based on its number of images or videos submitted. This method is used as an index of measurement to order the priority of testing. In addition, if the image is available, we perform image classification trained by the existing images, with two labels: Unverified and Verified. This aims to predict the likelihood that the image represents a verifiable setting. This method is used to order the testing.

In order to prioritize the testing of some samples due to the limited number of resources, we propose two different ordering methods:

1. **Majority Voting**
   In the majority voting scheme, we take a majority vote on location model, text analysis, image availability analysis and image classification. Details of voting scheme will be discussed in section 3.1 below.

In both of the prioritization schemes, we can use location model only to predict the likely hood of a sighting being actually positive.

2 Models

2.1 Location Model

The location of a reported sight is a key feature for both evaluating whether a reported sighting is likely to be the positive sight of Asian giant hornets, and prediction of their future expansion. For this model we must assume

- Asian giant hornets will not go further than 8km away from its nest[7]
- Every year the new queens make a new nest within 30km from the last nest[7]
- The positive reports represent a positive spotting of Asian giant hornet when it is out of its nest.

2.1.1 Spotting a Asian giant hornet

It is very rare for the nests of Asian giant hornets to be above ground and therefore very hard to spot.[7] This means the vast majority of reports of Asian giant hornets are going to be when the hornets are out of their nest and looking for food. From this, we know that a positive reporting of a Asian giant hornet means there is a nest within 8km, which is the maximum distance the hornets will fly away from their nest when in search of food.[7] Without precise information on how far from the nest the a Asian giant hornet will be when in search of food, we will assume the distance a hornet is away from its nest follows a uniform distribution with a range of 0km to 8km around the location of the report. This results in a uniform distribution in a circle with its center at the location of the report and a radius of 8 km. The distribution has the following probability density function(PDF):

$$g(d) = 1/64\pi \tag{1}$$

2.1.2 Where will a nest be next year

We know that the Asian giant hornet is an annual species and when winter arrives the nests will die and the only ones to survive is the queens. The queens will hibernate in the ground until spring. After hibernating the queens will then go out and establish a new nest within 30km of their previous nest.[7] Without precise information on how far the new nest will be made in relation to the original nest, we will assume this distance follows a uniform distribution with a range of 0km to 30km around the location of the original nest. This results in a uniform distribution that is in the shape of a circle with a center at the location of the original nest and a radius of 30km.

The PDF of the nest location in the following year is shown by:

$$h(d) = \frac{1}{900\pi} \tag{2}$$

where d is the distance of the new nest from the original nest .

2.1.3 Yearly Timestep

Since we assume a Asian giant hornet is spotted when they are out looking for food and all hornets except for the queen die in the winter, a spotting will never happen while any of the queens are out establishing a new nest. This means to predict the location of where a nest will be next year, we can use a timestep(yearly) that occurs in the spring while the queens establish new nests. We call the position

of a reporting as p_1 and use the PDF $g(d)$ to predict the position of the current nest, which we call p_2. Given the position of p_2, the PDF $h(d)$ is used the predict the position of next years nest, which we call p_3. Since we don't know the position of p_2, but rather the PDF of where p_2 will be, we combine the PDF's of $g(d)$ and $h(d)$ which is called $f(d)$.

2.1.4 Calculating $f(d)$

Since the location of p_3 is dependent upon the location of p_2 and both PDF's, $g(d)$ and $h(d)$, are uniform distributions, the PDF of f(d) can be calculated with the area of overlap of the circles around p_1 and p_2.

$$A(d) = \begin{cases} (r^2\pi), & 0 \le d \le R - r \\ \left(r^2 \cos^{-1}\left(\frac{d^2 + r^2 - R^2}{2rd}\right) + R^2 \cos^{-1}\left(\frac{d^2 - r^2 + R^2}{2Rd}\right)\right) & \\ \quad -\left(\frac{1}{2}\sqrt{(-d + r + R)\,(d + r - R)\,(d - r + R)\,(d + r + R)}\right), & R - r \le d \le r + R \\ 0, & r + R \le d \end{cases} \tag{3}$$

where,

- $A(d)$ is the area of overlap of two circles
- d is the distance between the centers of the two circles
- r is earths radius of the smaller circle
- R is earths radius of the Larger circle

For $f(d)$ to be a PDF, we multiply $A(d)$ by a constant C.

$$f(d) = \begin{cases} C(r^2\pi), & 0 \le d \le R - r \\ C\left(r^2 \cos^{-1}\left(\frac{d^2 + r^2 - R^2}{2rd}\right) + R^2 \cos^{-1}\left(\frac{d^2 - r^2 + R^2}{2Rd}\right)\right) & \\ \quad -C\left(\frac{1}{2}\sqrt{(-d + r + R)\,(d + r - R)\,(d - r + R)\,(d + r + R)}\right), & R - r \le d \le r + R \\ 0, & r + R \le d \end{cases} \tag{4}$$

where,

- $r = 8$
- $R = 30$
- $C = 1/\int_0^{r+R} A(d)\,dd$

2.1.5 Monte Carlo method

We use the Monte Carlo method to estimate the PDF of where new nests occur in the future. To do this we will take the points p_1 and then plot a large amount of points p_3 based upon $f(d)$ and use the distribution of the points p_3 to simulate the PDF for where new nests will occur in the following year. As long as the number of points of p_3 is large, the density of p_3 will be a close estimate for the PDF of where new nests will occur in the following year. This same method can be used to create a PDF

to represent where new nests will occur in two years. To do this the points p_3, which are calculated from p_1 are then treated as new values for the points p_1 and the same process of creating points p_3 is repeated to calculate the PDF for the second year. Similarly, the PDF for where nests will occur in any year in the future can be simulated.

2.1.6 Calculating the position of new nests

To calculate the position of p_3, the distance in kilometers is calculated according to $f(d)$. A random value from 0 to 360 is given to θ which represents the direction from p_1 to p_3. Since the position of p_1 is given in latitude and longitude, A Mercator Projection is used to map p_1 into a 2d plane with units in kilometers, indicated by Eq.5 and Eq.6 .

$$y = Esin(\frac{2L_a\pi}{360}) \tag{5}$$

$$x = \frac{L_o}{360(2\pi E)} \tag{6}$$

where,

- L_a is latitude
- L_o is longitude
- E is earths radius in kilometers

Now that p_1 has a position of x and y in kilometers, δx and δy must be calculated using d and θ to find the position of p_3, as shown by Eq.7 and Eq.8

$$\delta x = dcos(\theta) \tag{7}$$

$$\delta y = dsin(\theta) \tag{8}$$

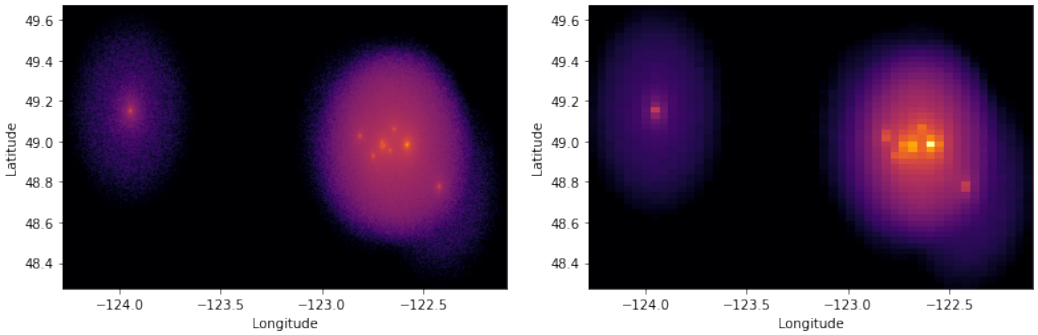
For the purpose of the Monte Carlo method, this process of calculating p_3 for one of the points p_1 is repeated a large number of times. This is then repeated for each p_1. The result is a large quantity of points p_3 which represent the PDF of where next years nest will be. Once all values of p_3 have been calculated, either the values of p_3 can be used to estimate the following years PDF or the values can be converted back into latitude and longitude by the inverse of the Mercator Projection.

2.1.7 Results

The result of the model shows a PDF of where nests will occur for a given year. The resulting PDF is continuous but can be turned into a discrete PDF by integration. This is useful since giving a specified area of land can be useful for the purpose of investigation.

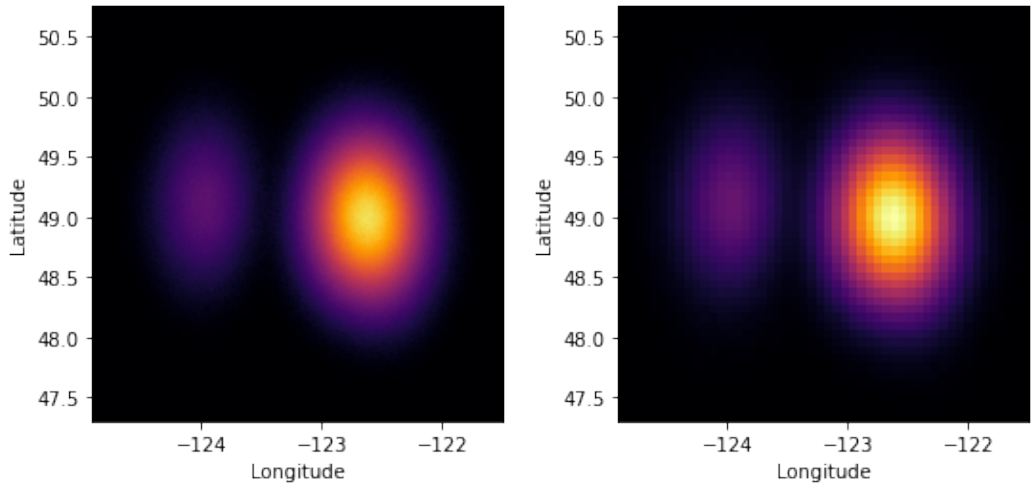
Fig.1(a) show the estimated PDF of where nests will be in 1 year as a continuous PDF and Fig.1(b) shows the discrete PDF which is sectioned into squares of width .1 longitude and height of .1 latitude. Similarly, Fig.1(c) and Fig.1(d) show the continous and discrete PDF of where nests will be in 4 years.

From the nature of this model, as the number of years increases in the future that the model estimates for, the models accuracy decreases. One downfall of this model is it does not take into consideration the geography. For example there is many points that are put into the water even though a nest can't be established in the water. As new positive sightings of the Asian giant hornet are reported, those data points can easily be added as new points for p_1 before the simulation begins.



(a) Spread of Asian Giant Hornets after 1 year, Continous

(b) Spread of Asian Giant Hornets after 1 year, Discrete



(c) Spread of Asian Giant Hornets after 4 years, Continuous

(d) Spread of Asian Giant Hornets after 4 years, Discrete

Figure 1: spread plots

Report Status	**Verifiable**	**Unverifiable**
Verifiable	331	100
Unverifiable	155	299

Table 1: Confusion Matrix of Text Analysis Testing

2.2 Text Analysis

Text analysis is important because the comments by reporters may be useful to determine whether the reported sighting is likely verifiable.

2.2.1 Principle

When performing text analysis, we analyse the words from "Notes" in each of the reports, and use a logistic regression model to perform training. In all of the reports, there are in total 9692 different words. After excluding all the reports that are in the lab status of "Unprocessed", we group the reports of both positive ID and negative ID as verifiable, as the rest as unverifiable. Then, we partition 20% of the data into testing set and the rest into training set. In total, we have 885 training data and 3540 testing data.

First, we encode the training set. In our training set, there are in total 9663 case insensitive different words in total. This means that each input note is encoded into a vector of dimension of 9663, where the magnitude at each dimension corresponds to the frequency of the word. Here, we ignore all spaces and delimiters during encoding, and the abbreviations are counted as one word. For example, "I'm" is a word. If we encounter some notes that are empty, they are represented by a zero vector in dimension 9663.

Then, we encode the testing set into vectors in dimension 9663 in similar fashion. However, if we encounter a word that did never appear in the training set, we ignore it.

Finally, we apply the logistic regression model from sklearn module to fit the training data.

2.2.2 Result

Table 1 shows the testing results of text analysis, where the rows represent the real category of the sighting, and the column represents the prediction with the logistic regression model. We find that the total testing accuracy is 70.55%.

2.3 Image Analysis

Image analysis is important because it provides information on whether an Asian giant hornet sighting is verifiable. Therefore, it is one of the models in the majority vote scheme to prioritize the verification of some possible sightings of Asian giant hornets.

Category/ Availability of Images	Available	Unavailable
Verifiable	2054	29
Unverifiable	68	2274

Table 2: Image Availability

2.3.1 Principle

When performing image analysis on reported sightings of Asian giant hornets, we group both positive IDs and negative IDs into verifiable dataset. The image analysis process is subdivided into two parts:

1. Image availability Analysis
   Based on the Asian giant hornets sighting dataset, we find that the sightings have different number of images or videos available. Here, we examine the number of images or videos available of each reported sighting, and explore its connection with the likely hood of a successful verification.
2. Image Classification
   Among the reported sightings with images available, we perform image classification to examine whether an image more likely corresponds to a verifiable sighting or unverifiable sighting.
   Through image classification, we first resize all the image data into squares of (512,512). In total, there are 3079 verified images and 73 unverified images, where 20% of the images are assigned into testing set and the rest into training set. Due to the much lower availability of unverifiable dataset with images available, we augment the number of unverifiable dataset in training by repeating each of the images for 40 times. Then, we apply InceptionV3 model with softmax activation function to train the images.

2.3.2 Result

Here we discuss the results of image analysis on the dataset:

1. Image Availability Analysis
   While some of the reported sightings contain more than one images, we categorize category each reported sighting as either with available images or without available images. Table 2 hows the distribution of image availability of both verifiable and unverifiable images.

   Based on Table 2, if we assume the lab processes data in the same fashion, which means a same availability distribution in the future, we have relationships as shown in Eq.9 and Eq.10

   $$P(\text{Verifiable} \mid \text{Image Available}) = 0.967 > P(\text{Unverifiable} \mid \text{Image Available}) = 0.032 \quad (9)$$

   $$P(\text{Unverifiable} \mid \text{Image Unavailable}) = 0.987 > P(\text{Verifiable} \mid \text{Image Unavailable}) = 0.013 \quad (10)$$

   Therefore, we can conclude that if a reported Asian giant hornet sighting includes an image, it is much more likely that the sighting is verifiable. However, if a reported Asian giant hornet sighting does not have available images, it is much more likely that the sighting is unverifiable.
2. Image Classification
   With InceptionV3 model of image classification, the Confusion matrix of testing after 20 epochs of training is shown in Table 3

True Label/ Tested Label	Verifiable	Unverifiable
Verifiable	423	197
Unverifiable	5	10

Table 3: Confusion Matrix of Image Classification Testing

Based on Table 3, we find that the testing accuracy of image classification is 0.682, which is mediocre. However, the precision of the testing is 0.988, which makes it confident to conclude that if an image is tested as verifiable, then it is highly likely it is verifiable. However, if it is tested to be unverifiable, there is not much information to say about the true verifiability, partly due to the small testing unverifiable sample size.

3 Prediction

3.1 Majority-Vote

3.1.1 Description

In our study, a majority-vote is applied to prioritize whether a reported sighting needs to be tested soon. Here, we assign the text classifier 1 vote, image classifier 2 votes, and location analysis 2 votes.

If the text classifier predicts a new report to be verifiable, then it votes *Yes*. If the reported sighting has at least one available image, then it votes for one *Yes*. If it has an available image, and the image classifier predicts it to be verifiable, and additional *yes* is voted. For the location analysis, the number of *Yes* votes is calculated by the following rule: The new report receives 0, 1, or 2 votes, if the predicted likelihood of being a positive sighting is < 50%, between 50% and 75%, or > 75% respectively.

In sum, a new reported sighting will get examined by lab if it received at least 3 votes.

3.1.2 Accuracy

Here, we calculate the classification accuracy of majority voting. In the majority voting scheme, the most important consideration is to lower the probability of categorizing a sample as unverfiable but actually its is verifiable. While there is not much information about the probability of a new report being a positive sighting without testing with the location model, we calculate the range of probability of misclassifying a verifiable sighting into an unverifiable sighting (i.e. false negative rate).

Whether a sample is going to be tested not only depends on its verifiability but also its likelyhood of being a positive sighting. Eq.11, Eq.12 and Eq.13

$$P(\text{Unexamined} \mid \text{Likelyhood of sighting being positive} > 75\% \wedge \text{verifiable}) = 0.3\% \quad (11)$$

$$P(\text{Unexamined} \mid 50\% \leq \text{Likelyhood of sighting being positive} \leq 75\% \wedge \text{verifiable}) = 8.6\% \quad (12)$$

$$P(\text{Unexamined} \mid \text{Likelyhood of sighting being positive} < 50\% \wedge \text{verifiable}) = 48.3\% \quad (13)$$

While we do not have the information on the total proportion of testing samples producing positive result, since the likelyhood of sighting being positive would be one of the cases above, the average false

negative rate would be between 0.3% and 48.3%. Also, when the likelyhood of the reported sighting being positive is over 75% based on the location model, almost all such verifiable sights will be tested by combining image and text analysis, as shown by eq.11, because one of them showing veriable would make the testing proceed on this sample. On the other hand, if the location model shows that the sighting is more likely negative, less than half of them are expected to be tested, shown by eq.13

3.2 Future Spread Prediction

We base on the location model on future spread prediction. Results are discussed in section 2.1 above. Based on Fig. 1, we can see that in four years, the Asian giant hornets would eventually be spreading and prevalent on the border between US and Canada in the west coast if not treated properly. However, states east of washington and south Oregon would not be affected quickly.

4 Discussion

We summarized the flow of our model in this section. Suppose a user upload a report sighting, our model will first extract the three pieces of information, i.e. location, text and image, and put each into corresponding classifier. The model adopts the majority-vote scheme. It has 5 votes in total distributed into the three classifiers.

The detailed voting scheme is as the following. The text classifier has 1 vote. If the text classifier predicts a new report's to be verifiable based on its attached note, it votes *Yes*. The image classifier has 2 votes. If a new report does not have an image available, then the sighting receives 0 vote, and otherwise 1 vote for its Image Availability. If there is at least 1 image and the prediction of the Image Classification is verifiable, it receives an additional 1 vote. The location analysis has 2 votes and the number of *Yes* votes is calculated by the following rule: The new report receives 0, 1, or 2 votes, if the predicted likelihood of being a positive sighting is < 50%, between 50% and 75%, or > 75% respectively.

By performing this majority-vote scheme, all reported sightings will be ranked according to how many *Yes* notes they get (with lots of ties, of course, for there might be more than 1 reported sightings get, say, 4 *Yes*). We believe our model spares no effort to pick out those reported sightings that are most likely to be verifiable and thus help prioritize the limited government resources.

Our model spares no effort to save the unnecessary time for examining reported sightings that are likely to be unverifiable and thus help prioritize the limited resources. At the mean time, our model adopts a high level of flexibility as the majority-vote scheme can be easily adapted into a weighted sum form. Furthermore, given additional labelled reported sightings, the three classifiers can be easily retrained so that it captures the new situation.

References

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

[2] Jacqueline R Beggs, Eckehard G Brockerhoff, Juan C Corley, Marc Kenis, Maité Masciocchi, Franck Muller, Quentin Rome, and Claire Villemant. Ecological effects and management of invasive alien vespidae. *BioControl*, 56(4):505–526, 2011.

[3] Ya-nan Cheng, Ping Wen, Shi-hao Dong, Ken Tan, and James C Nieh. Poison and alarm: the asian hornet vespa velutina uses sting venom volatiles as an alarm pheromone. *Journal of experimental biology*, 220(4):645–651, 2017.

[4] KG Juliana Jeyanthi. Fatality by multiple asian giant hornet stings: two case reports from south india. *Journal of South Journal of South India Medicolegal India Medicolegal Association Association*, 2(2):68–71, 2010.

[5] Zheng Liu, Xiang-Dong Li, Bo-Hui Guo, Yi Li, Ming Zhao, Hai-Yan Shen, Ying Zhai, Xue-Li Wang, and Tao Liu. Acute interstitial nephritis, toxic hepatitis and toxic myocarditis following multiple asian giant hornet stings in shaanxi province, china. *Environmental health and preventive medicine*, 21(4):231–236, 2016.

[6] Beverly McClenaghan, Marcel Schlaf, Megan Geddes, Joshua Mazza, Grace Pitman, Kaileigh McCallum, Samuel Rawluk, Karen Hand, and Gard W Otis. Behavioral responses of honey bees, apis cerana and apis mellifera, to vespa mandarinia marking and alarm pheromones. *Journal of Apicultural Research*, 58(1):141–148, 2019.

[7] Michael J. Skvarla. Asian giant hornets. *PennState Extension*, 2020.

[8] Gengping Zhu, Javier Gutierrez Illan, Chris Looney, and David W Crowder. Assessing the ecological niche and invasion potential of the asian giant hornet. *Proceedings of the National Academy of Sciences*, 117(40):24646–24648, 2020.

Memorandum

To whom it may concern,

We are team 2126887 of The Mathematical Contest in Modeling 2021. We are writing to propose our solution to predict the Asian giant hornets spreads in the Washington state. We hope the model will help the government prioritize investigation of the hundreds, if not thousands, of public reports. We notice that many reports turn out to be "unverified", which means lab has investigated the report but no conclusion could be made. We believe one bright spot of the our solution lies in that it helps identify whether a report is likely to be classifiable or not. In other words, our model spares no effort to save unnecessary investigation into the public reports.

Our model adopts the majority-vote scheme which consists of 3 classifiers built on images, text and location analysis. For the image classifier, we performed training on labelled images using the InceptionV3 model. We also discover the correlation between the number of images uploaded and that particular report's verifiability. This observation is also included into our majority-vote scheme. For the text classifier, we performed the logistic regression model on the text that the public typed along with their report. Through the key-word analysis, we see that words such as "trapped" implies a higher probability of verifiability, while words such as "woods" implies a lower one. For the location classifier, we made several justifiable assumptions on the Asian giant hornets' behavior. Given a new reported sighting, we calculate the probability of it being positive by calculating the distance between the new reported sighting and previous positive labels.

The models we applied in this study has yielded relatively high accuracy. For the text analysis, we get a total accuracy of 70.5%. For Image classification, we get a total accuracy of 68.2%. If we combine these classifiers in the majority scheme we have proposed in the paper to prioritize some of the testing, the likelihood to misclassify a verifiable sighting report as unverifiable given that it is predicted positive in location model ranges from 0.32% to 8.67%. Therefore, we can confidently say that our majority voting scheme would not likely miss the opportunity to examine a positive sight.

As more positive sightings are updated into the dataset, our model can easily be updated into the current setting just by retraining the text, images, and locations of the dataset, which could be accomplished relatively fast. In the future, if the likelihood of an areas containing at least one Asian giant hornet nest falls below 5%, then we can confidently say that the Asian giant hornets are most likely eradicated, since the likelihood of the area having at least 1 Asian giant hornet nest is more than 95%, which is highly likely.

Our model is not perfect, as we have not investigated into the habits of Asian giant hornets. For example, they mostly likely do not inhabit in the ocean, on the plateau, or in the mountain [8]. To solve this issue, the geographical landscape could act as a constraint. In a refined model, we can assign the probability of an Asian giant hornet nest moving into the ocean as zero, and into a high mountain as a lower probability. Also, the effects of the wind-aided or human-aided dispersion are not clearly examined. To solve this issue, we need to do more analysis on the effects on wind or carriers on Asian giant hornets. Other deficiencies of our study include a low sample size of Asian giant hornets. If more positive sightings of Asian giant hornets are found in the future, we could potentially perform

machine learning directly on classifying whether a reported sight is positive or negative, which would also save the government a lot of time.

In addition to refining our model by considering the niche and detailed dispersion patterns of Asian giant hornets, we can also put the classifiers together in a different way. For example, instead of majority-voting as proposed in the study, we could take the weighted average the likelihood of the testing result of each classifier, and then rank them from the highest to lowest to order the testing timeline.

Another method that could make the testing efficient while also expansive is to encourage the reporters to include more high resolution images and more detailed text description. Based on the previous data, we find that the reported sightings that contain at least one image is almost 76 times more likely to be a verifiable report sighting than those with no image available. A way to facilitate more effective reporting would be to provide some monetary award to those reporters who provide positive sightings of Asian giant hornets.

As students, we are truly worried about the ecological health of America, and the interest of people who are affected by the Asian giant hornet crisis. This report proposed a flexible and efficient model for reporting sightings and prediction on the future of the Asian giant hornets. Lastly, we would like to thank you time reading our report and your kind consideration

Sincerely,
Team 2126887
The Mathematical Contest in Modeling 2021