Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2125661

TO BEE OR NOT TO BEE

In this paper, we attempt to increase the efficiency of Asian Giant Hornet report classification and answer questions posed by the Washington State officials in their effort to combat the spread of invasive species Asian Giant Hornet (AGH). In order to do so, we first create a classification analysis that provides suggestions for researchers on whether or not reports are correct sightings of AGH.

First, we analyse and extract information from the given data set by separating negative reports to mis-identified species reports in order to add variability in the trained model.

Then, we create species-matching indexes for each report based on the image and report note information. By creating a library of reference images, we perform SIFT matching for each report image and obtain an image-matching index. Similarly for report notes, we create a library of high value words for each species using the tf-idf algorithm, matching each report note we obtain a note-matching index.

Combining image-matching index and note-matching index, we are able to find a species-matching index through conjugate gradient optimization. We then create a growth-observation index based on classic animal expansion models that take into consideration repulsion factors. Since the spread of AGH in Washington State is influenced significantly by human activities, our attempt focuses on analysing observations of AGH rather than its growth.

Combining the species-matching index and growth-observation index, we create a function that models the relationship between positive (species-match to AGH, growth-observation index) and negative (species-match to other species) factors. Using SANN optimization, we are able to finalise a classification index for each report that measures their likelihood of being a negative sighting.

We use the classification index to determine two key threshold indexes to help identify most likely positive sightings. Observing simulations growth-observation, we give numerically informed reviews on the current and future states of AGH and present them in our memorandum. We find that currently we are unable to claim AGH eradication in Washington State, if no unexpected situations occur, we expect to be able to claim eradication within a year.

Throughout the process, we record and analyse each assumption and calculation in order to achieve maximum preciseness in our final model.

Keywords: SIFT; tf-idf; Expansion Model; Multivariable Optimization

1 Introduction **3**
1.1 Our work 3

2 Model construction **3**
2.1 Species matching 3
2.1.1 Image matching index 4
2.1.1.1Assumptions 4
2.1.1.2 Scale-Invariant feature transform 5
2.1.2 Note matching index 7
2.1.2.1 Term Frequency and Inverse Document Frequency 8
2.1.3 Species identification results 10
2.1.3.1 Result analysis 11
2.2 Growth estimation 11
2.2.1 Assumptions 12
2.2.2 Simulation Process 12
2.2.3 Distance Factor 13
2.2.4 Repulsion factor 13
2.2.5 Area smoothing/averaging 14
2.2.6 Results 14
2.3 Classification analysis 15
2.3.1 Assumptions 15
2.3.2 Model 15
2.3.3 Training system 16
2.3.4 Results 17
2.3.5 Threshold computation 17

3 Predictive analysis **18**
3.1 General trend 18
3.1.1 Eradication 18
3.2 Updating model 18
3.3 Strengths 20
3.4 Weaknesses 20

4 Conclusion **21**
4.1 Future improvements 21

5 References **22**

Memorandum **23**

1 Introduction

Like most invasive species, Asian Giant Hornets (AGH) pose great threats to the local ecosystem it is introduced to. Specifically, by being an aggressive species that prey on honeybees and other hornets vital to a healthy ecosystem, AGH rises above many to be one of the most dangerous invasive species if left uninterrupted. In fall of 2019, the two nests of AGH each found in Vancouver Island, British Columbia and Washington State alerted officials to take action in preventing further spread of the pest, residents to report possible sightings of AGH. Amongst thousands of reports filed, only a fraction of them appeared to be correct sightings of AGH, while the majority mistook other species of bees/hornets as AGH. Since professionals can only judge the report manually using limited information, it is both time-consuming and costly to continue such high-intensity search. Yet, as there is no certain proof of AGH's eradication in Washington State, it is dangerous to stop the search. To combat this, we create a system that identifies possible false sightings, thereby highlighting possible correct sightings. We then produce a detailed analysis of the preciseness of our model and its implications in real-world settings.

1.1 Our work

Our work first creates a classification analysis that provides suggestions for researchers on whether or not reports are correct sightings of AGH. Since we have limited positive sightings, it is hard to recognise a false sighting purely based on its relationship with AGH. We therefore consider a probability model where we assess the likelihood of a sighting belonging to another species, therefore not AGH. We also consider the natural behaviour of AGH and its implications on the pest's local spread. Using SANN optimization, we create a function outputting estimation for false reports. Using thresholds computed by simulations of optimized function, we give a final classification index.
We then use the classification model to give evaluations of AGH's future trends, creating a memorandum for the Washington State of Agriculture. Throughout the process, we record and analyse each assumption and calculation in order to achieve maximum preciseness in our final model.

2 Model construction

2.1 Species matching

Amongst negative sightings, we notice that researchers are able to give clear negative indications when they are able to identify the report as another type of animal or hornet species. Therefore, by creating a system that estimates the most likely species of a report, we are able find the probability of a report being false. We do so by examining two aspects of a report: report image and report note. These choices will be justified later in their corresponding sections. Using matching indexes found for image and report note, we estimate the species-matching index.

We find the “correct” species by examining lab comments of negative sightings. Since almost all lab comments give a clear indication of mis-identified species, we are able to compose a list of species “true values” which will later be used to train and test our species-matching index. Here we have a list of most commonly miss-identified species:

1. Golden digger bee 2. Sawfly 3. Neus erocerus 4. Bald faced hornet 5. Cicada killer 6. Yellow jacket	7. Bumble bee 8. Paper wasp 9. Syrphid fly 10. June bug beetle 11. Jerusalem cricket 12. Robber fly

For each species, we create a reference index that is broken into two parts: image match index and notes matching index. We are then able to compare each image to the species reference index to find its most likely species. Below we explain how each index is found, starting with image matching index.

2.1.1 Image matching index

Examining the datafile given, we find that only 1.3% of negative sightings do not have images, while 96.9% of unverified sightings do not have images. It is logical to infer that researchers base their ability to pinpoint a report to be false mostly on images.
Immediately, we face the challenge of non uniform data. Since our images were uploaded by individual reporters taken under varying lighting and backgrounds, two images appearing to be similar may have vastly different quantitative features. Therefore, our method needs to be able to identify the main target desired, bees/hornets/flies, while also being immune to lightscale changes and image variables such as difference in rotation and size.
The Scale Invariant Feature Transform (SIFT) algorithm designed by David Lowe satisfies the requirements for changes in lightscale and other image variables. We further improve upon the results by initialising noise cancellation and subject detection algorithms to report images prior to implementing SIFT in order to minimise environmental influences.

2.1.1.1Assumptions

1. Only species listed are relevant
2. Gaussian smoothing performed does not give aliasing sample, pre-blurred assumed to have at least sigma value of 0.5

2.1.1.2 Scale-Invariant feature transform[1]

The SIFT algorithm detects image features as key points. By comparing key points of two images through the Brute Force algorithm, we are able to find matching points. We set the similarity threshold of 0.75 as maximum euclidean distance between matched points and obtain a final percentage similar result. Using the list of species found through frequency tests, we identify three images (each characterizing a different angle/feature of the species) to create a reference index for each species. We then compare SIFT features of each report image to the reference index of each species and obtain a similarity index. We normalise the sum of similarity index of all species to create an image-matching index.

Species Reference Index SR_i for Species S_i

$$(kp_{ref}, dsp_{ref}) = SIFT(SRI)$$

where SRI is species reference image using SIFT algorithm, kp_{ref} is reference keypoint, and dsp_{ref} is reference descriptor.

For each report r,

$$(kp_r, dsp_r) = SIFT(r_I)$$

where r_I is report image, kp_r is keypoint of report image, and dsp_r is descriptor of report image.

$$M_r = BF(dsp_{ref}, dsp_r)$$

where M_r is matched points using brute force matching

$$I_r = \sum_{i=1}^{s} \frac{M_r}{kp_{ref_i}}$$

where I_r is the image index of report i

We test our results by checking with our “correct” species label extracted earlier from lab comments. Results are shown below.

Figure 1 shows the average matching rate for all training pictures and reference pictures is 0.2828+-0.18. We looked into each reference species' pictures and graphed the comparison across the training pictures groups. We use the true positive groups, the similarity index of the same labeled species and reference species, as an indication for our model accuracy. Though the true positive group is not always the highest similarity index, it usually lies among the top 3 among all the species groups.

[1] Statistics were done using Python 3.9.0 opencv-contrib-python 4.5.1.48 (Open Source Computer Vision Library)

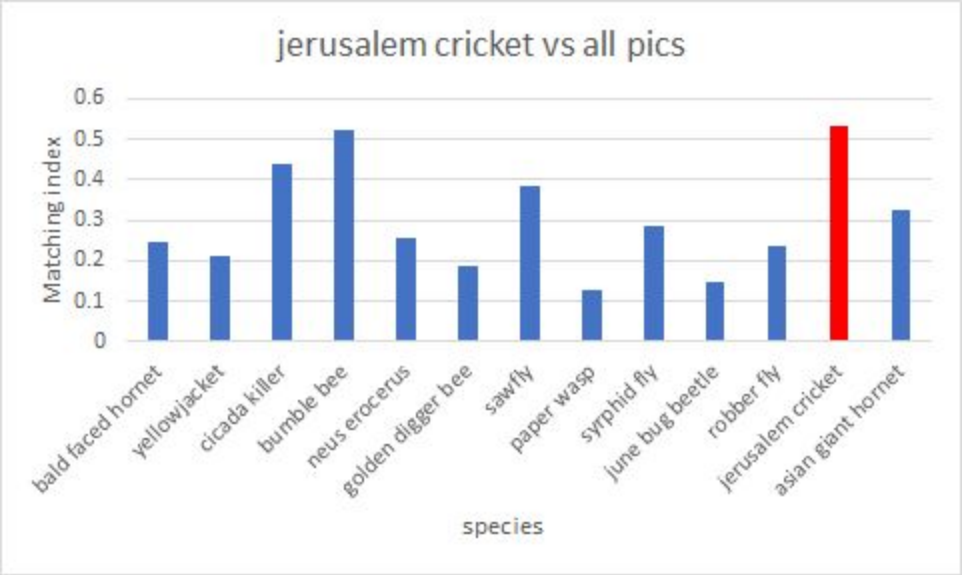

Figure 1: The average similarity index of all possible species on the pictures that have been labeled as "jerusalem cricket".

The red bar is the true positive group in this comparison and we can see that it is the highest among all the species.

In the same way, we fix one of the species groups in our training images to see their performance across all the reference species.

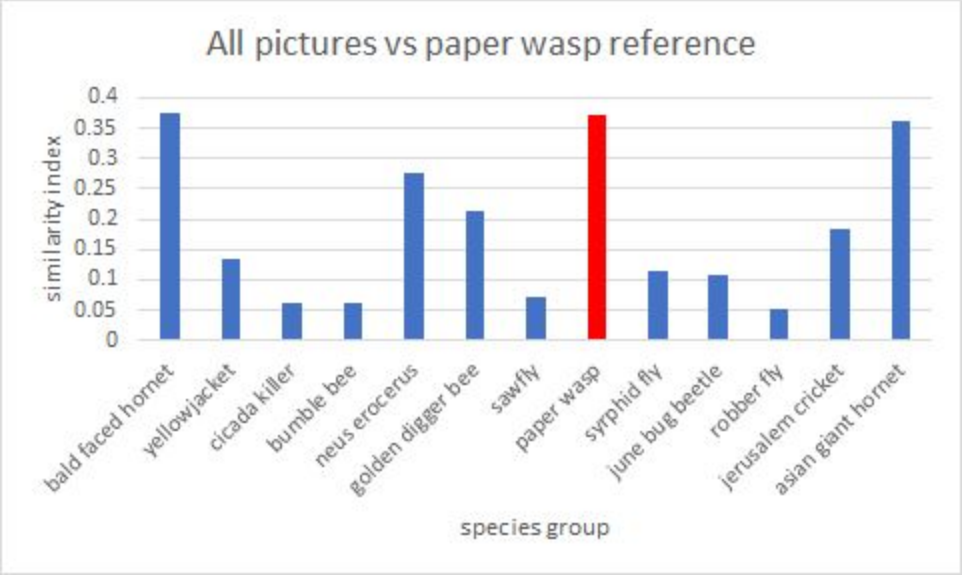

Figure 2. The average similarity index of the yellow jacket reference on all pre-labeled training pictures.

The red bar is the true positive group in this comparison and we could see that it ranks the second among all the species. However, we see an analogous distribution of similarity index on

different groups of species, where the bumble bee and jerusalem cricket is always the top 2 among all groups. So it is logical to infer that the performance is sensitive to the reference images. The reason might be the fluctuation of noise level across each reference image, and angles. And we double checked the reference photos to validate our prediction.

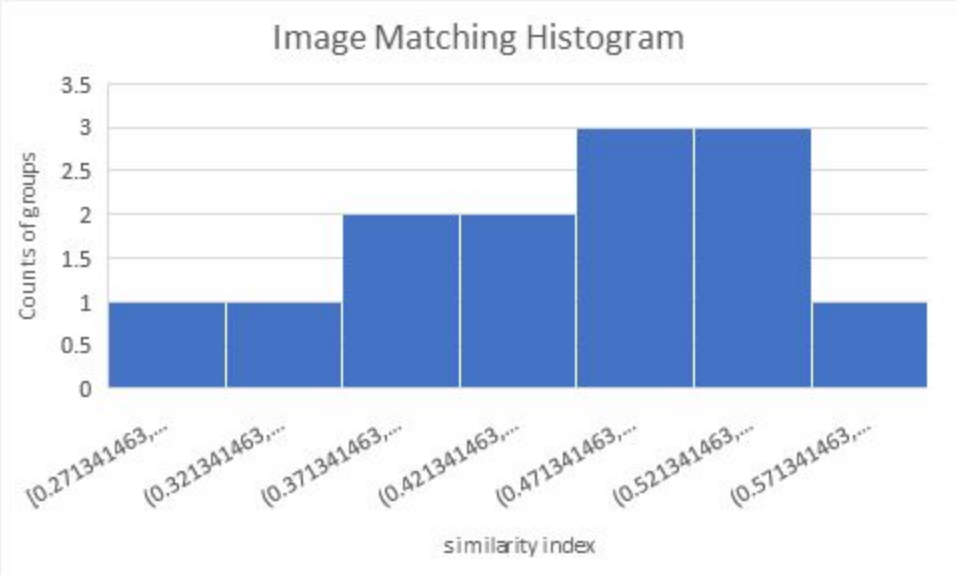

Figure 3: This histogram of the distribution of true positive similarity index for the 13 species' groups.

We can see from the histogram above that the majority true positive index lies in the interval of 0.4-0.5 which is above the 3rd percentile in this dataframe. This is another strong evidence to show that our model can better match the true positive pictures than others.

2.1.2 Note matching index

Examining unverified reports, we see that though images are the primary consideration for a researcher to mark a report as negative sighting, report notes still provide valuable information on a rough guess of the species. This is especially prominent in reports with vague/blurry images accompanied by report notes, research comments often will give a suggestion based on verbal descriptions of the insect. Therefore, though report note descriptions are not the definitive factor in defining the species of a report, it is able to enhance its accuracy, thus justified to be included. We expect the weight of note matching index to be significantly lower in the overall species matching model compared to the image matching index.

Acknowledging that information gathered from large-quantity text relies heavily on word frequencies, we use frequency related features to estimate the reported species. Since report notes are often non-formal and consists of mostly day-to-day language with little professional terms, our method needs to be able to ignore commonly used words such as “a” “I” “the”, at the same time highlight characteristic words that appear more often in reports of a certain species. Therefore, our algorithm should be able to identify:

1. Frequency of each word present across all reports for each species
2. Uniqueness of each high-frequency word compared to words of other species.

2.1.2.1 Term Frequency and Inverse Document Frequency

The Term Frequency and Inverse Document Frequency (tf-idf) method does just that. Calculated by two factors: term frequency and inverse document frequency that offsets each other, the tf-idf score for each word approaches 0 if the word is very common amongst all bodies of texts (high tf score but low idf score), or if the word is uncommon. Applying the algorithm we have[2]:

$$tf(w,r) = log(1 + freq(w,r))$$
$$idf(w,S) = log\frac{N}{count(r \in S : w \in r)}$$
$$tfidf(w,r,S) = tf(w,r).idf(w,S)$$

where w = word, r = reports of a species, S = species

Normalising tf-idf for each species, we obtain an index that is representative of each word's weight in identifying species, and show the final matching index for each report as sum of normalised tf-idf for each unique word.

$$N(r_n \in S) = \sum_{i=1}^{n} w_i * tfidf(w, r_n, S)$$

N= note-matching index,
r_n= report note,
w_i= word i of r_n, $i = 1 \ldots n$ where n=length(r_n)

The average similarity index for all training notes to all species is 0.1+-0.04. The average of similarity index between all notes and Asian Giant Hornet is 0, which is a good indication that our algorithm can successfully predict the false report. We graphed the comparison across the training notes on each species group. Though the true positive group is not always the highest similarity index, it usually lies among the top 4 among all the species groups.

[2] Equation referred extrapolated from (Stephen, 2014)

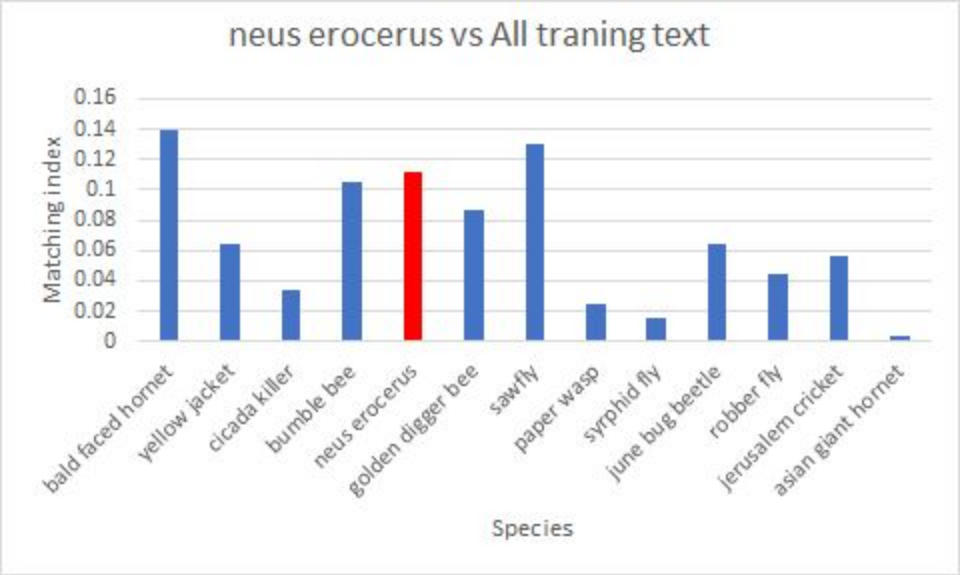

Figure 4: the average similarity index of all possible species on the notes that have been labeled as "neus erocerus".

The red bar is the true positive group in this comparison and we could see that it ranks the third among all the species.

In the same way, we fixed one of the species groups in our training notes to see their performance across all the reference species. However, we see an analogous distribution of similarity index on different groups of species. So the performance is sensitive to the reference images. The reason might be the fluctuation of noise level across each reference image, and angles.

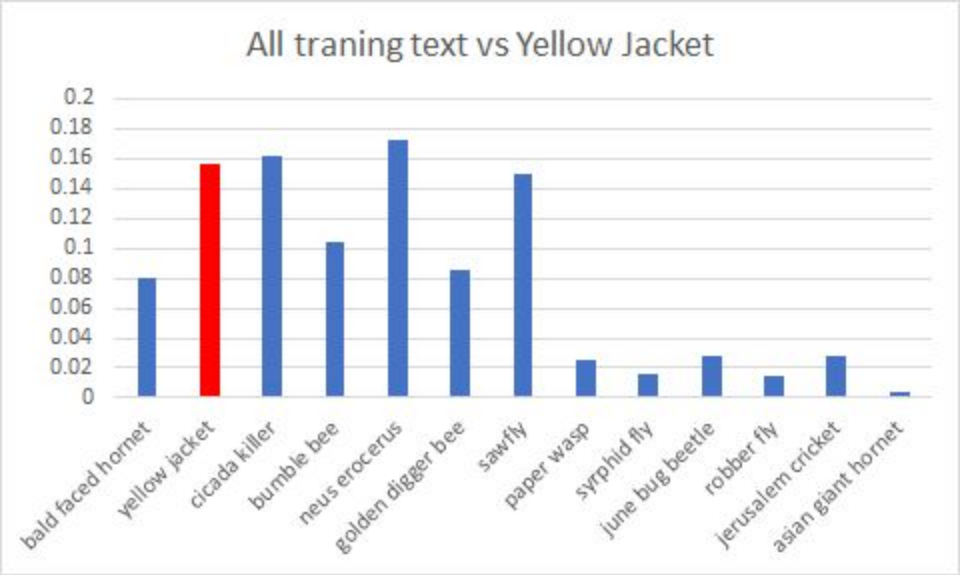

Figure 5: The average similarity index of the "yellow jacket" on all pre-labeled training notes.

The red bar is the true positive group in this comparison and we could see that it ranks the third among all the species.

Unlike the image matching, we did not see any analogous distribution of similarity index on different groups of species. So we are able to conclude that the notes matching is not sensitive to the different species groups.

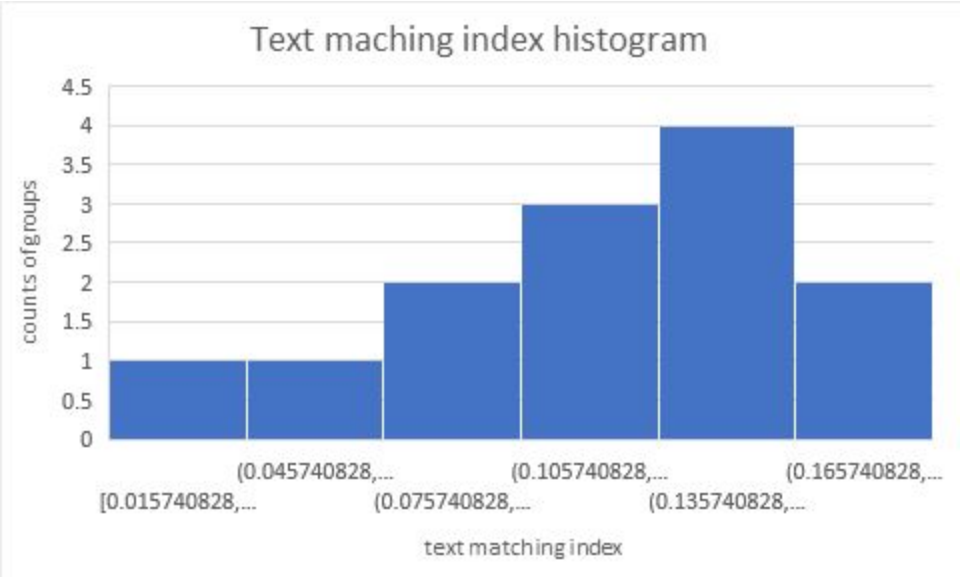

Figure 6: Distribution of true positive similarity index for the 13 species' groups.

This histogram shows that the majority true positive index lies in the interval of 0.07-0.08 which is above the 3rd percentile in this dataframe. This is another strong evidence to show that our model can better match the true positive pictures than others.

2.1.3 Species identification results

With both the image match index and note match index, we can create a model combining the two in forming a species-matching index overall. We design this model by considering the relationship between image index, note index, and identified species. Though the information in images and notes may be connected in reality, where the note may describe the image qualitatively, they are independent mathematically. Therefore, we multiply them each with a weighting factor for the final species-matching index:

$$S = cN^a I^b$$

S = species matching index N = note matching index I = image matching index

Linearizing the function, we get:

$$log(S) = k_1 log(N) + k_2 log(I), \text{ where } k_1, k_2 \text{ are unknown parameters}$$

We find the weight parameters by using a training set to perform a global optimization set to minimise least squared error.

2.1.3.1 Result analysis

The optimal value for the two unknown parameters k_1 and k_2 is 2.546, 0.068 respectively. We compute the accuracy of the species match by testing the model on a test set. Then we analyze the distribution of the overall species-matching index matrix.

The average matching index for all testing data is 0.00356+-0.00172

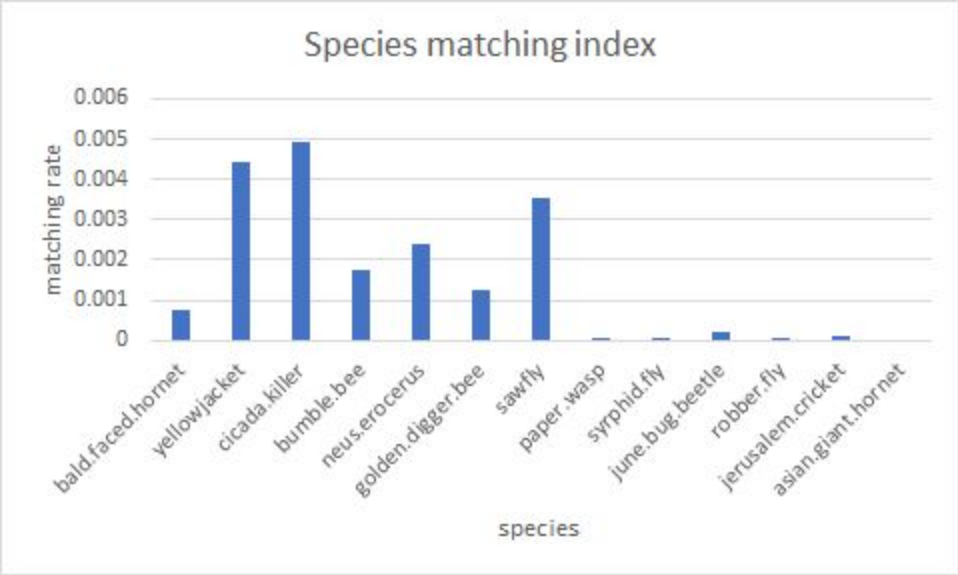

Figure 7: The distribution of the average matching index of each species.

The x-axis is ordered by the number of datapoints of each species, and we could see that there is a significant accuracy drop from the top 7 groups to the bottom 5 groups which is coordinated to the amount of data points drop between the two tiers. The reason behind this drop of accuracy might be the increase of error with less data.

2.2 Growth estimation

To effectively predict the result of dynamic expansion of AGH colonies, we employ a computer simulation program to model the movement of AGHs within two periods of time: 2019.9-2020.4 and 2020.5-2020.10. For each time period, this model yields a matrix of probabilities for detecting AGHs in each small area on the map. By comparing the probabilities and using these data to plot heatmaps, we identify some patterns of the spread of AGHs in terms of geographical distribution and changes over time. The probabilities also contribute to the classification analysis discussed later.

2.2.1 Assumptions

1. Given that a new queen has an estimated range of 30km away from its current nest for establishing a new nest, we assume the distance that an AGH has a maximal moving distance of 30km within a certain time period.
2. Food and other resources around the colony is approximately uniformly distributed so that the new queen will move in a random direction to establish a new nest.
3. Flying requires a considerable amount of energy, so the queen tends to choose a closer position for the new nest among potential positions with similar conditions.
4. The average lifespan of Asian giant hornets is 3 - 5 months.

2.2.2 Simulation Process

For complicated dynamic processes, it is appropriate to use coarse-grained computer simulations for modeling. We differentiate the types of movement according to the moving distance. For frequent short-distance moving (i.e. less than 10km), we regard it as moving within the ambient rectangular space. In comparison, AGHs sometimes but less frequently fly much farther. When AGHs move farther than 20km, we identify such a move as a long-distance move.

Since we need to learn about the spread of AGHs over a large area across Washington State, we should mainly consider long-distance moves and ignore the moves within the ambient space. Therefore, we divide the region that contains public sightings into 50*50 small grids, where the probability of AGH detection is assumed to be the same in every position in the grid given the short-distance moves. Each grid has a length of 0.18 latitudinal degrees (approximately 19.8km) and a width of 0.10 longitudinal degrees (approximately 11km). Then we identify the grids that contain positive sightings of AGHs and assign them with a probability of 1, which indicates that we are certain to claim the existence of AGHs at these grids. For other grids, the probabilities of detecting AGHs are initialized to be "None."

For each moving, we initialize two parameters: moving distance r_0 and autonomous moving direction d_0. Considering Assumption 1, we utilized a normal distribution with $\mu = 25$km and $\sigma = 5/3$ for r_0. By the empirical rule, an AGH has an approximate range of [20,30] for its moving distance. In addition, since food and other resources are uniformly distributed (Assumption 2), AGHs will randomly choose an autonomous moving direction d_0 (a unit-length vector) first, but the probability of making an actual move in direction d_0 is affected by repulsion with other colonies, which will be discussed in detail in the next part.

To observe the changes over time, we divide the dataset of public sightings into two subsets according to the detection date. The two periods are 2019.9-2020.4 and 2020.5-2020.10. We ran 30000 iterations for an 8-month period, which means that AGHs conduct long-distance moving approximately twice a day.

For individuals within a grid, we consider the 8 surrounding grids as the potential destination of a long-distance move since the maximal moving distance is assumed to be 30km. Likewise, for a grid free of AGHs, those among the 8 surrounding grids with a non-zero probability of AGH detection are likely to affect its probability of AGH detection. We compute a weighted sum of the probabilities that the nearby individuals will move into this grid and interpret it as the probability of AGH detection of this grid over the entire duration of 8 months or 6 months. Let S be the set of ambient grids and P_i be the detection probability of a grid. The formula is:

$$P_i = \sum_{j \in S, P_j \neq 0} P_j \; e_d \; e_p$$

where P_j is the detection probability of the surrounding grids that have probably already been visited by AGHs, e_d is the distance factor and e_p is the repulsion factor

2.2.3 Distance Factor

By Assumption 4, we define a distance factor e_d to account for the reluctance of moving farther for concerns about energy consumption. The distance factor is calculated using the following formula:

$$e_d = 1 - \frac{r_0}{r_{max}}$$

where r_0 is the moving distance and r_{max} is the greatest possible moving distance.

2.2.4 Repulsion factor

The expansive and diffusive behaviors of animal colonies are not random but influenced by other surrounding colonies. With the natural tendency to avoid overcrowdedness, individuals in a population will have a repulsive interaction with other individuals when they are too close to each other and may interfere with the necessary space for growth and foraging activities. More specifically, individuals attempt to maintain a minimum distance r_p with their peers. According to researchers in zoology, this avoidance is the highest priority in each individual's rule of motion. Similarly, this pattern holds for the distribution of colonies to avoid overly severe competition and guarantee sufficient ambient space for metabolic activities.

In our simulation model, we define a repulsion zone, Ω_p, which is a circle around an AGH colony with radius r_p. The collective repulsive effect of other colonies within Ω_p determines the desired moving direction for individuals in the colony at the center of Ω_p. Let $c_i(t)$ be the

position of this colony at moment t. The desired direction of travel d_i is calculated according to the following formula (P.Liu 2014):

$$\mathbf{d}_i(t) = -\sum_{j\in\Omega_p, j\neq i} \frac{\mathbf{c}_j(t) - \mathbf{c}_i(t)}{|\mathbf{c}_j(t) - \mathbf{c}_i(t)|}.$$

By our assumption, food as well as other resources almost have no influence for the moving of Asian giant hornets. Therefore, the autonomous direction is randomly hypothesized in simulation. However, in reality moves in different directions are not equally likely to happen because of the repulsion from ambient individuals/colonies. From intuition, the more deviant the autonomous direction is from the desired direction, the less likely an individual AGH is to conduct such a move. The angle between these two directions can be calculated through dot product:

$$\theta = arccos(\frac{d_i \cdot d_0}{\|d_i\|\|d_0\|})$$

As a discount factor between 0 and 1, the repulsion factor e_p is relatively closer to 0 when the deviance, which is measured by θ is large. Thus, e_p is negatively correlated to θ but positively correlated to π -θ . We assume that the corresponding random variable Y=π - θ follows a uniform distribution over [0, π]. Then e_p is defined as:

$$e_p = P(0 < Y < \pi - \theta) = \int_0^{\pi-\theta} \frac{1}{\pi} dy = 1 - \frac{\theta}{\pi}$$

2.2.5 Area smoothing/averaging

To obtain a better result, we average the probabilities for each grid location from 3 independent 30000-iteration simulations.

2.2.6 Results

The computer simulation provides us with a probability matrix whose entries are the likelihood that AGHs spread to the grid region and may potentially be detected later. With this simulation model, we can extract some features of AGH dispersal and roughly predict the spread of AGHs. Using this matrix, we generated a heatmap for Washington state. Obviously, the positions close to the positive reports of AGHs, typically lying in the northwestern corner, have a relatively high detection probability. In addition, from the graph we can identify the pattern that AGHs are more likely to move eastward than southward.

Comparing the two heatmaps for 2019.9-2020.4 and 2020.5-2020.10, we have discovered that area with high detection probability shrunk significantly in the latter period, which probably indicates an effective inhibition of the spread of AGHs, a remarkable progress towards the complete elimination of them in Washington State.

Computer simulation offers us a powerful tool to predict the complicated dynamic processes in real life, but this model definitely has its limitations. For instance, even after 30000 iterations, there are still some grids whose detection probability is "None," which means that no AGHs have moved into these grids. It is difficult for us to determine the detection probability for these grids. There is no reason to simply assign a probability of zero for these regions despite the very low probability of AGH detection. We use the mean of detection probabilities in the surrounding grids for approximation.
Moreover, we have not integrated the factor of altitude and terrain into the model, which is related to the habitat preference of AGHs. According to relevant studies, low mountains are the most favorable habitat, while plains and high-altitude mountains are avoided by AGHs. Since we did not differentiate between different altitudes and terrains, the model fails to match an un-negligible factor in the natural expansion of AGHs. Besides, the spread and distribution of AGHs may also be influenced by the configuration of urban and suburban areas. Large-scale human activities in cities are expected to interfere with the natural growth of AGH, which can motivate them to move to the countryside to stay there consistently. This tendency has not been reflected in the model.
All these limitations will affect the precision of our model to some extent and introduce some deviation.

2.3 Classification analysis

Now that we have both the species-matching index and growth-observation index, we arrive at the final stage of classifying our report.

2.3.1 Assumptions

- All lab identified species are true
- All species not identified are irrelevant
- Growth-observation index variation within a time fame is limited

2.3.2 Model

We can break down our goal into two parts:

1. Probability of report being true of AGH
2. Probability of report being another species

We have information on both factors from previous sections. For estimating true of AGH we have:

- Species match of report to AGH
- Growth-observation index representing likelihood of observation of AGH at report location

For estimating other species, we have:

- Species match of other species

We form the following model:

$$F(r \neq AGH) = k_1 P(r_s = AGH) f(r_g) + k_2 f(P(r_s = S_{1-12}))$$

where F = final index, r = report, r_s = species match, r_g = growth index, S_{1-12} = other species

$$f(r_g) = -\frac{1}{log(P_o)}$$

$$f(P(r_s = S_{1-12})) = max\ P(r_s = S_{1-12})$$

Since our goal is to identify false reports, factors related to true sightings will be negatively correlated to our result, while factors of false sightings will be positively correlated, therefore parameter k_1 should be negative, parameter k_2 should be positive.

Since the growth-observation index is very small for the majority of locations, we perform natural log to reflect the order of magnitudes of these values, which gives us a clearer comparison. In order to maintain its properties of always being positive and its relative size comparison across locations, we perform negative inverse.

For each report image, we take its best matched species other than AGH as an indicator for false report. Since our index is a combination of different manipulated probabilities and is not normalised, it is not a standard probabilistic indication. Since we are manipulating the growth-observation index to be negative inverse, we will encounter negative infinity values for when growth-observation=1, which will be our lower bound. Our upper bound will be when species-match AGH or growth-observation index reaches 0, which stands at k_2 f(P(r_s=S_{1-12})) depending on the trained value.

2.3.3 Training system

Since we are using indexes generated by iterative processes based on coarse data, we only attempt to find a rough global minimum for our function to test the predicted features in order to verify our method. Therefore we use the Simulated Annealing Neural Network method to solve our least squared error optimization problem.

$$min(F(r \neq AGH) - true\ negative)$$

$$\text{true negative}= \begin{cases} 1 & \text{, when true species} \neq AGH \\ 0 & \text{, when true species} = AGH \end{cases}$$

It should be noted that 0 and 1 does not represent probabilities but a numeric binary, which we set for the boolean nature of true negative values.

2.3.4 Results

Running the system on a training set containing 1,000 reports, we get k_1= -19.46, k_2=17.27. We split the 14 positive reviews of AGH into 2 groups, with 7 mixed into the training set and 7 in the testing set. For reports within the same bin of positive identification, growth = 1, P_o = -∞, therefore P=-∞, we will consider these cases as highly dangerous.
Our parameter values correspond to the predicted trend of k_1 being negative and k_2 being positive. Observing their ratios, we can say that factors representing true AGH sighting are slightly more valuable than factors representing negative sightings.

2.3.5 Threshold computation

Based on our testing set, we find the mean index of positive reviews to be -0.18 with an average deviation of 1.32, while the mean index of negative reviews to be 3.43 with an average deviation of 1.25. Based on this, we suggest to mark reports with an overall index greater than 4.68 as false reports, consider reports with overall index between 4.67 and 1.15 as suspicious, and thoroughly investigate reports with negative index lower than 1.14. Doing so will allow researchers to prioritize on reports that are most likely positive sightings, reducing cost.

3 Predictive analysis

3.1 General trend

By comparing results from the growth-observation model of two different time frames, we are able to see that the spread of AGH significantly decreased from fall 2019 to fall 2020.

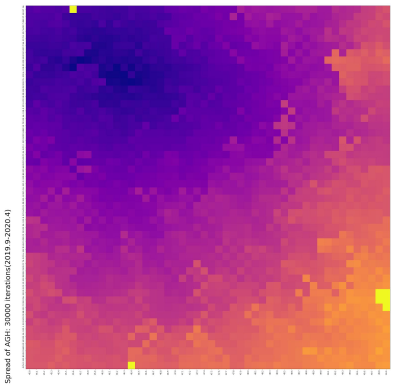

Figure 8: growth-index heatmap for fall 2019

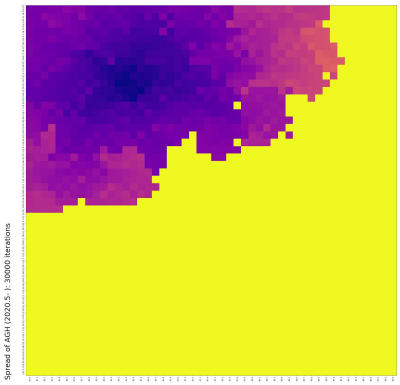

Figure 9: growth-index heatmap for fall 2020

3.1.1 Eradication

Our main point of measure for spread is based on the growth-observation model. By observing the sum of all grids for one time frame, we have a quantified understanding of the sighting likelihood for AGH across the whole of Washington state. Currently, the sum of fall 2019 - spring 2020 is at 6.4, while the sum of spring 2020 - fall 2020 decreased to 3.17. Since 1 is the critical number that represents one colony, we state that if the sum of growth-observation probability is below 1, AGH has been eradicated from Washington State.

3.2 Updating model

New reports bring two forms of value:

1. Based on content
2. Based on quantity

Specifically, reports that bring new content add significantly more value to improving our model than those based on quantity on multiple levels. Since we have a very limited amount of positive AGH sightings, any new reports considered to be a positive sighting will be highly significant. Amongst new positive reports, if the report is located at or around past positive sighting locations (identified via growth-observation model), it provides us with two information:

1. Training information for AGH image matching and note matching
2. Temporal information of the continuation of AGH after summer 2020

The first is important again due to limited information on AGH from current datafiles, retraining AGH matching will allow the model to be more precise. The second is important since it allows us to compute a new growth-observation graph of the next time frame, which allows us to better assess the growth trend of AGH in Washington.
Another possibility is to receive reports located at a location with low growth-observation index. This are two possibilities for this to occur:

1. AGH has spread to the new location without any detection as it spreads
2. A new colony of AGH has established from a new source

Both situations are highly alarming. If this were to occur, we suggest officials immediately take action to track the source of the new reporting to understand its path (eg. by genetic test). If it is from the same colony as ones previously present in Washington State, we suggest to re-examine past reports. Though this is resource intensive, it will only be a short burst of cost compared to the long-term damage if the hive was left unchecked. If the new report belongs to a different colony, we first suggest to run the growth-observation model with the new positive input and identify areas with high chance of positive sightings, then to take action in removing it by searching through those areas. Below we show an growth-observation example if a new positive point were to be added.

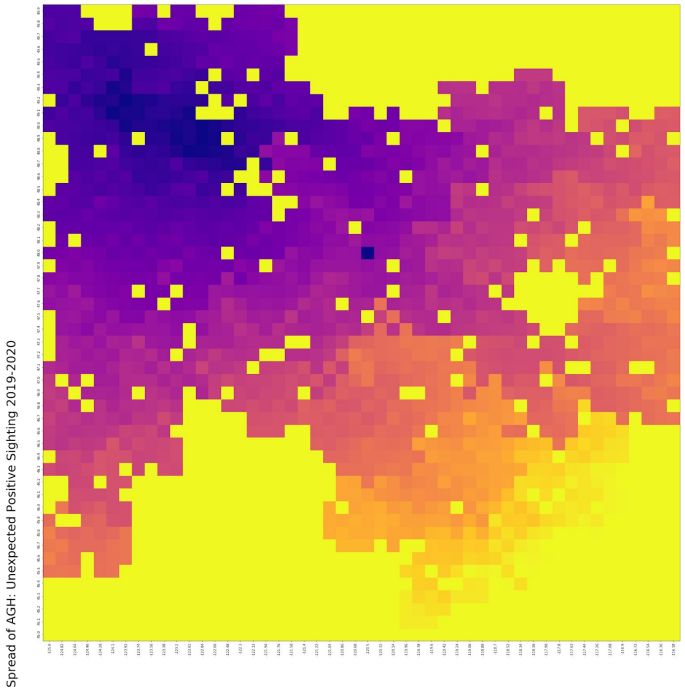

Figure 10: Heatmap of growth-observation if a new positive is added

Non-positive sightings with unique sets of values also provide useful information. We suggest that beyond the basic indexing, to also create a ratio test comparing AGH-match index to growth-observation index. If we have a report with a high AGH-match index but low growth-observation index, we would recommend performing manual check on the report. Doing so may help identify new AGH sightings.

We then consider if no unique sightings were to arise and we only receive negative/unverified results. Since the movements/habits of AGH are highly predictable, namely that they die off over winter and start new colonies in spring, we will suggest officials to re-train our model every year until we confirm eradication. Doing so not only improves the accuracy of future indexing, but also gives officials better understanding of the current state by providing key quantified information that can be plotted and compared.

3.3 Strengths

Our model is strong in that it reduces the impact created by a very small number of initial positive reports. By considering both positive (AGH match and growth-observation) and negative (match with other species) factors, we increase variability in our training model allowing for a more comprehensive result than if we were to only consider reports as negative/positive.
Our overall method is strong in that we are able to give clear answers and provide a systematic routine for researchers to use in their daily assessments. By providing an index and related threshold, we are able to provide a solid basis for further analysis on AGH trends in the future.

3.4 Weaknesses

Though our model does not rely as heavily on variability between positive and negative sightings, we do depend heavily on information obtained from each report to perform accurate species-matching. It is very difficult for our model to give meaningful indexing if a new report only contains images or report notes but not the other, which as we observe do take up a significant number of reports. This however is not necessarily a defining flaw to our model, but rather to the process of identification itself since researchers aren't able to identify purely based on notes either. This limits the application of our model to being a first-round sift that provides recommendation information to researchers rather than a complete classification model that can stand on its own.

Our growth-observation model is very sensitive to initial settings, therefore should be recalculated for every new positive report. This may be computation intensive for researchers to do regularly, however if the spread of AGH follows the analysed trend of decreasing to eradication there will be limited new positive sightings. In this case re-running the model should be acceptable cost wise.

4 Conclusion

Through this paper, we present to you an attempt to automate the report identification process in order to maintain the level of alert needed for invasive species detection while significantly lower its resource cost. By using methods that consider both report-level likelihood as well as large scale distribution, we are able to create a classification system that is applicable and easy to use. Researchers will be able to sift incoming reports through our trained model and have an initial understanding of whether the report is true positive.

Our predictive analysis answers many questions of interest to government officials. First we identify that Washington State cannot claim eradication of AGH yet, and should only do so when the sum of growth-observation index reaches below 1 from the current 3.17. Then we provide an algorithm that classifies reports into negative, suspicious and positive based on thresholds 4.68 and 1.15.

4.1 Future improvements

1. Image-matching
    a. Since our image matching algorithm currently considers the entire image, we find that it is harder to output accurate matches if the image has noisy background or multiple focus objects. A way to combat this is to perform object detection prior to running our image-matching algorithm to identify central areas.
2. Text-matching
    a. Our text-matching training is still at a very rough stage since the number of sightings for each species vary. For some species, like the golden digger bee, that are more commonly mis-identified have a reference-dictionary a lot more well-built than less common species such as robber flies. This difference may have led to differences in match rate. This is why retraining the model when receiving large quantities of new reports will be helpful in the future.
3. Growth-observation
    a. Our simulation model does not take into account the factors of altitude and the potential interference of human activities in urban areas, which brings some deviation from the real situations. We can do further research about the specific locations of low mountain areas and implement a factor to explain the attraction of such places.

5 References

[1] Cloud.tencent.com. 2021. *Classic SIFT Algorithm*. [online] Available at: <https://cloud.tencent.com/developer/article/1081140> [Accessed 9 February 2021].

[2] Lowe, D., 2004. *Distinctive Image Features from Scale-Invariant Keypoints*. [ebook] Vancouver, B.C., Canada: University of British Columbia. Available at: <https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf> [Accessed 9 February 2021].

[3] P. Liu, H. R. Safford, I. D. Couzin, and I. G. Kevrekidis, "Coarse-grained variables for particle-based models: diffusion maps and animal swarming simulations," Computational particle mechanics, vol. 1, no. 4, pp. 425–440, 2014, doi: 10.1007/s40571-014-0030-7.

[4] Robinson, J., 2021. *Term Frequency and Inverse Document Frequency (tf-idf) Using Tidy Data Principles*. [online] Cran.r-project.org. Available at: <https://cran.r-project.org/web/packages/tidytext/vignettes/tf_idf.html> [Accessed 9 February 2021].

[5] Stephen, R., 2004. *Understanding Inverse Document Frequency: On theoretical Arguments for IDF*. [ebook] London, United Kingdom: Journal of Documentation. Available at: <https://www.researchgate.net/publication/238123710_Understanding_Inverse_Document_Frequency_On_Theoretical_Arguments_for_IDF> [Accessed 9 February 2021].

Memorandum

TO: Washington State Department of Agriculture
FROM: Team 2125661
DATE: February 8th 2021
SUBJECT: The State of Asian Giant Hornet in Washington State and Methods of Improving Detection Efficiency

Invasive species pose great threats to their local ecosystem. The invasive species Asian Giant Hornet (AGH) first found in Washington State in fall 2019 is especially harmful since it preys on key local pollinators such as bumble bees that are vital to agriculture. For this reason, the state has been putting a great number of resources into detecting AGH. However, amongst thousands of reported sightings of AGH, only 14 were identified to be positive, while the rest were false reports mistaking other species for AGH. Thus, it is useful for us to 1. identify whether such detection is still necessary and 2. create a less costly system to identify negative reports before manual assessments.

Assessing current state of AGH

After analyzing the observed positive reports, we find that though the spread of AGH has decreased from fall 2019 to fall 2020, it is still not at a point of eradication. Since AGH can cause significant damage quickly if left unchecked, this group suggests for officials to continue detection. Using the overall growth-index obtained from our algorithm, a sum below 1 will suggest that the pest has been eradicated from Washington State. Before reaching threshold level, this group suggests officials to perform yearly rerun of growth-index based on new reports for the year in order to stay informed. Below we show a heatmap of growth-observation index based on longitude-latitude locations across two time frames:

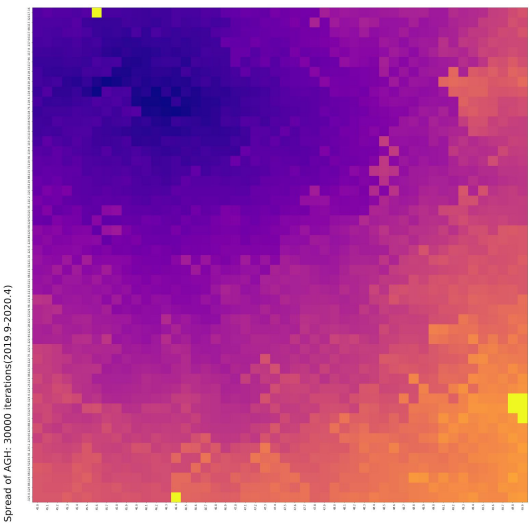

Figure 1: growth-index heatmap for fall 2019

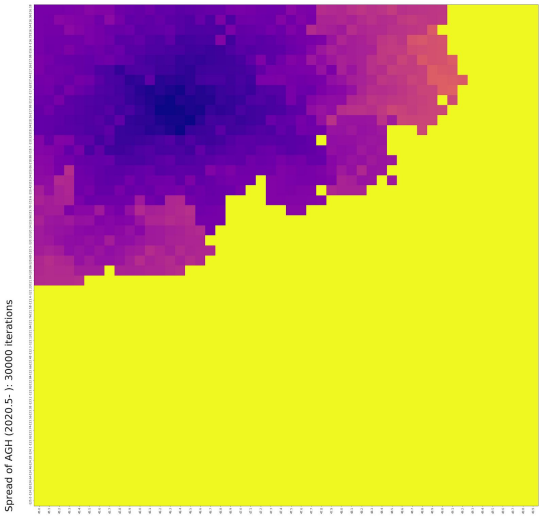

Figure 2: growth-index heatmap for fall 2020

Assessing new reports - data processing

Data provided by public reports have the general trend of being non-professional and highly varying, however it still has many information that can be extracted. We consider the data we can gather from public reports of this problem to be:

1. Time of sighting
2. Location of report
3. Report note
4. Images

We use time and location of report to create the growth-observation model that acts as a biology-informed analysis of the observation likelihood. Using images and report notes, we are able to create references that can be used to compare with new reports.

Assessing new reports - classification index

In order to reduce resources needed to identify positive reports, we create a systematic classification algorithm that provides a classification index. Based on tested simulations, we finalize on two key thresholds to determine the likelihood of a report being true. If the classification index is greater than 4.68, we consider the report to be misidentified; if the index is between 4.67 and 1.15, we consider the report to be suspicious but mostly false; if the index is lower than 1.14, we consider the report to be likely of being positive and should be manually checked. Using this method, we are able to prioritize reports that are more likely to be positive.

Recommendations and warnings

Based on our results, we provide the following recommendations:

1. Continue with the detection of AGH until overall growth index reaches below 1
2. Perform yearly update on grow index to better understand its distribution
3. For future detections, perform classification analysis before manual search
4. Take immediate action to manually analyze reports and perform on-site research if the following occur
    a. Reports are identified with high species-matching index but low growth-observation index
    b. Positive reports are identified at locations with low growth-observation index
    c. Overall growth index increased significantly over a time frame