Problem Chosen	2021	Team Control Number
C	MCM/ICM Summary Sheet	2118089

The Asian giant hornet (AGH) originates from East Asia was first detected in North America in 2019. The potentially invasive species poses a considerable threat to honeybee populations across the United States if allowed to spread. The United States Department of Agriculture (USDA) and the Washington State Department of Agriculture (WSDA) are focused on eradicating the AGH from the Pacific Northwest. Based on citizen reports of AGH sightings, the goal of the WSDA is to determine which reports are most plausible so they may investigate further.

Our model consists of two components that were modeled separately and later combined to form a cohesive model: *Classification* and *Predicting the Spread of AGH.* While analyzing the distribution of our data, we noted the relatively small data set we had to work with and a major class imbalance between user submissions that were positive AGH identifications and negative ones.

Classification: We use both supervised and unsupervised machine learning techniques to identify characteristics as well as common mistakes for AGH sighting submissions that were confirmed to be negative. We use those characteristics to predict the likelihood that a given submission is negative. Because we had a small data set to work with, we opted to use more robust statistical learning techniques rather than deep learning methods. We began by investigating which species the AGH was commonly confused with by comparing lab professional notes with user submitted notes; those species were digger wasps, cicada killers, sawflies, and horntails. From there, we introduced a data pipeline for supervised learning to see if we could gain more insight into what made each of the above categories unique. We utilized NLP tokenization and feature engineering along with both Multinomial Naive-Bayes (MNB) and a linear support vector machine (SVM) and achieved prediction accuracies of around 37% for insect classification and 99% for a negative case classification[1]. Next, to determine if there were defining features of the negative class itself, particularly with regards to the language used in submissions, we isolated the negative results and used a k-means algorithm to cluster our data into groupings with similar features. We were able to supplement these results visually by using both linear and nonlinear dimensionality reduction techniques. The successful qualitative results of k-means suggests that using Gaussian Mixture Models (GMM)— which would produce quantitative results—could offer more useful results.

Predicting the Spread of AGH: Using only the time and locations of positive IDs of AGHs, we develop a model to predict where future positive IDs will be found, thus providing a method for prioritizing reports, *performing better than randomly generating reports.* This spatial model is similar to the reaction-diffusion equation, containing a growth/spreading term, and and a local logistic term. This model takes a discretized version of the locations of the positive IDs and outputs a probability distribution for where we are likely to find positive IDs. The model is physically-derived, and only three parameters are tuned with the existing data. This model can be integrated with a variety of different probability models as inputs, and it can be updated often due to its linear properties. We suggest that this model be used for predicting about 2/3 of the cases that should be prioritized. In addition to the introduction of the above models, we suggest that some prioritized cases are selected through random spatial sampling. Specifically, we suggest that about 1/6 of the prioritized cases are selected within a 20-40 km radius of confirmed or predicted positive IDs, to account for outlier cases of AGHs traveling farther than expected when forming new nests. Another 1/6 of prioritized cases should be selected beyond this 40 km to account for unpredictability.

Ultimately, our model provided valuable insights into the characteristics of user submissions that were negative. We note that the accuracy and robustness of this prediction was limited due to our data set. Based on those prediction probabilities, we were able to predict the locations of future positive IDs to determine a way to prioritize cases.

[1] This result is overall accuracy. We suggest looking at our classification report.

Contents

1 Introduction

The world's largest hornet and an apex predator, the Asian giant hornet (AGH) feeds on medium-large insects like beetles, mantises, and honey bees [1]. AGH queens can grow up to 2 inches in size with a 3 inch long wingspan, while workers can grow up to 1.5 inches [2]. AGHs are social wasps, form their nests underground, and can be identified by their vibrant yellow heads, black thoraxes, and yellow and black or brown striped abdomens [2]. In the United States, they are commonly confused with other hornets, cicada killers, golden digger wasps, and yellowjackets [2].

Since the Asian giant hornet was first detected in North America in 2019, many AGH sightings have been reported in British Columbia, Canada and in the state of Washington and it is unclear how much they have spread. Just handfuls of AGHs can decimate entire colonies of honeybees in mere hours [2]. The presence of AGHs could cause substantial damage to honeybee populations, which add nearly $20 billion of value to U.S. Crop Production [3]. The United States Department of Agriculture (USDA) plans to spend about $1 million on the eradication of the AGH in the Pacific Northwest in the year of 2021 and another $1.3 million on protecting honeybees across the United States [4].

The Washington State Department of Agriculture (WSDA) has created a citizen scientist program, where sightings of suspected AGHs are reported [5]. The vast majority of the processed sightings result in false positives, which are equivalently termed "negative IDs". However, institutional funding is spread thin, posing a need for a more *efficient* method of processing these reports and ultimately predicting AGH locations. An ideal method would prioritize investigating the cases that are most likely to be positive first. To accomplish this, we may use the content of the submitted report, and/or the spatial location of the report.

1.1 Problem Summary

- The widespread presence of AGHs in North America could be devastating for honeybee populations and honeybee farmers. The USDA is prioritizing the detection and eradication of the AGH from Washington state to prevent the potentially invasive species from proliferating.

- Citizens in Washington state are encouraged to report potential AGH sightings to the Washington State Department of Agriculture [5]. Due to limitations in resources, the state must determine which of these reports are most likely to be positive sightings to follow-up with investigation.

- The data extracted from the public submissions can be classified into the following categories: *Images/Videos, Detection Date, Notes, Lab Status, Lab Comments, Submission Date, Latitude, and Longitude.*

- From the given information, the goal is to **prioritizing the investigation of submissions that are most likely to be positive sightings.**

1.2 Our Model

Our model consists of two parts—1. Classification of a report and 2. Predicting the spread of the population— which are then combined to establish a cohesive model.

1. We first classify new reports based on their likelihood of being a negative result by implementing both supervised and unsupervised machine learning techniques on the user reports data. We improve the classification by using techniques for handling small and imbalanced

data sets as found in the literature including considering different representations of the data, as well as implementing both linear and non-linear models [6][7].

2. We then model the spread of the AGH in Washington. By using only the time and locations of positive IDs of AGHs, we develop a model to predict where future positive IDs will be found, thus providing a method for prioritizing reports of AGH sightings.

In this report, we identify common features of negative reports and create a model to predict the spread of AGH populations. We note that the quality of these models is limited by the quantity and class distribution of our data. We propose a framework for how this model should be updated over time as new cases are reported. Lastly, we discuss necessary criteria for the AGH to be considered eradicated.

2 Assumptions

- We assume that submitted images don't affect the classification of the report as determined by our model. The lack of samples would not allow a robust deep learning model. Furthermore, since most of the reports classified as 'unverified' were due to a lack of image submission, we opted to flag any result with no image submission as low priority unless submitted by a verified source, because most of the time it did not contain enough information to make an identification.

- In the classification problem, we give the same weight to each feature. User submitted notes are given the same weight as longitude, latitude, and dates. This is the case by convention for most machine learning algorithms.

- We assume that female bees fly with no intrinsic preference to direction. Male bees do not fly independently. They follow the females within some close distance. According to [2], male AGHs typically only fly a couple of kilometers from their nests.

- The area of interest is geographically homogeneous. That is, we ignore the influence of the pacific ocean, mountainous terrain, and human-made environments.

- The existence of negative IDs within an area does not show the lack of AGH within an area. That is, the negative ID location data should not influence the probability of finding a positive ID there. If we assume that people mistake other insects for the AGH, this may happen regardless of AGHs being in the area.

- The only data we are allowed to use is the given data.

3 Data

3.1 Class Imbalance

Upon inspection of the data available, we first considered the structure and distribution of the data. Our data contains 4440 samples with 8 features. These 8 features are represented in following depiction of our dataset:

Note that we have relatively small sample size in the realm of machine learning. This makes creating a robust model difficult, as there is high bias that will cause overfitting. Because of this limitation, we opted for more statistical methods where there is better generalization. We were

GlobalID	Detection Date	Notes	Lab Status	Lab Comments	Submission Date	Latitude	Longitude
—	12/8/2019	—	Positive ID	—	1/15/2020	48.980994	-122.688503
—	10/30/2019	—	Positive ID	—	1/15/2020	48.971949	-122.700941
—	1/15/2020	—	Unverified	—	1/15/2020	48.939200	-122.661300
—	9/19/2019	—	Positive ID	—	2/4/2020	49.149394	-123.943134
—	8/31/2019	—	Unverified	—	2/14/2020	48.723779	-122.354431
—	10/15/2019	—	Unverified	—	2/27/2020	48.986176	-122.697450
—	2/29/2020	—	Negative ID	—	2/29/2020	48.729596	-122.480035
—	3/1/2020	—	Unverified	—	3/2/2020	48.186024	-122.344680
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

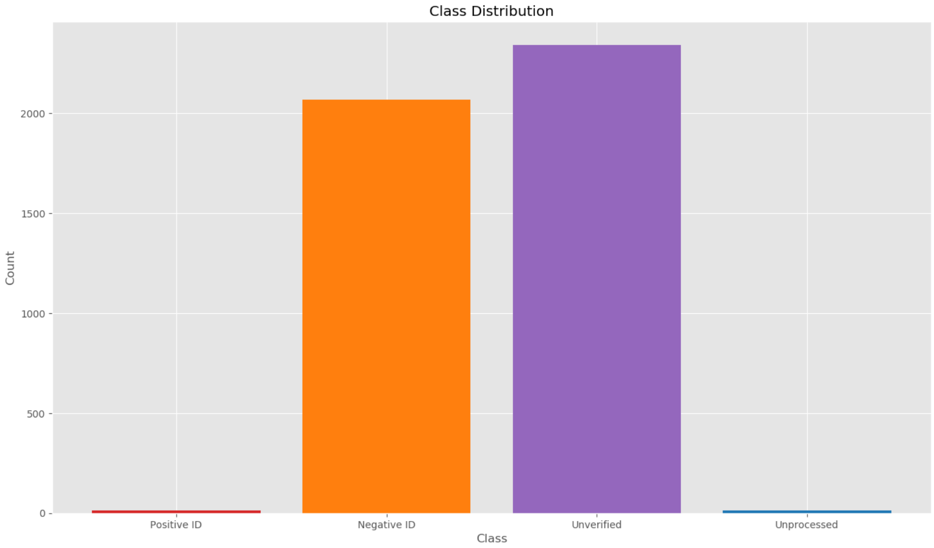

Figure 1: This is a fairly severe class imbalance problem as there is not enough data for meaningful data augmentation and too few data for downsampling.

also given images corresponding to the GlobalID feature; however, we opted not to consider these as many pictures were entirely missing or low quality.

We began by separating our dataset into the four 'Lab Status' classifications of the data: **Positive ID, Negative ID, Unverified, Unprocessed.** 'Positive ID' meant the reported sighting was confirmed as an AGH, 'Negative ID' meant the reported sighting was confirmed as not an AGH, 'Unverified' meant there was not enough information given to make a classification, and 'Unprocessed' meant the submission had not been processed yet. Upon inspection of the structure and distribution of our data, we quickly realized we were presented with a major class imbalance problem. Out of 4440 total samples, we had only 14 positive AGH IDs and 2069 negative IDs. This result can be seen in Fig. 1.

3.2 Lexicon Exploration

Furthermore, we suspected the text from the user submitted notes and lab comments would offer the most insightful information regarding the validity of their finding. In order to explore this, we used NLTK's tokenization package [8]. This allowed us to strip the notes by both the user as well

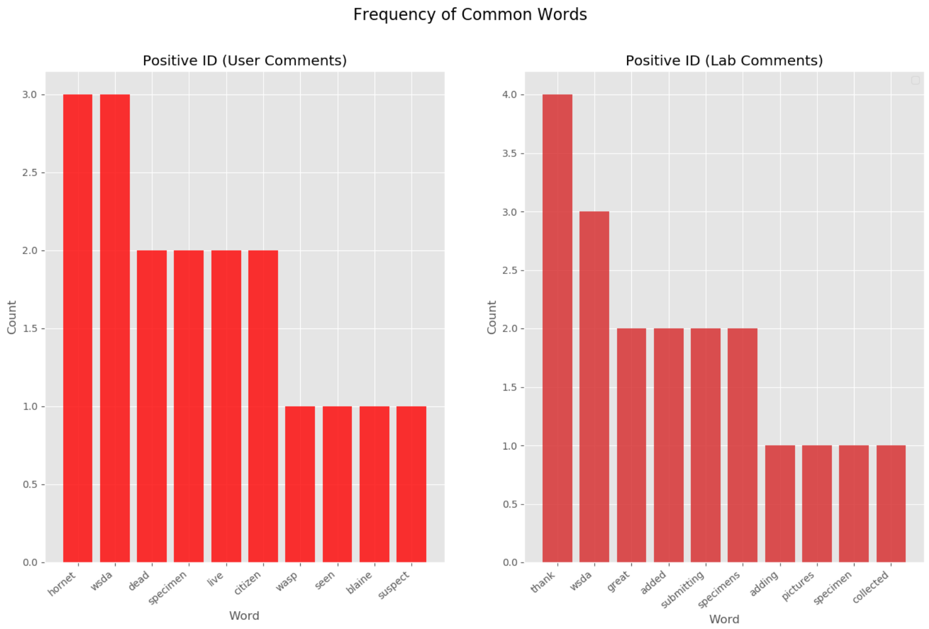

Figure 2: Positive Cases.

as the lab professional to determine if there exist trends among language.

To begin, we can look at the most common words used among the positive cases in Fig. 2.

Here, we note that there are only 14 cases of positive identification and that the most "common" words even have an occurrence of 1. This is, of course, essentially the entire dataset and so should be taken with a grain of salt. Although, we do see words that we would expect such as *wsda*, which stands for Washington State Department of Agriculture.

The following trends were confirmed by inspection:

- There are "citizen scientists" or WSDA staff that have submitted accurate findings

Next, in Fig. 3, we have the negative cases. Already, we know that we have 2069 samples, a dramatic increase from the positive cases. Upon further inspection, we can see that the user comments mainly contained words such as "found", "long", "large", "sure", "like", "bee", and "wasp." From first glance, it appears that maybe the size of the AGH play a role in making people think that it is an AGH. Additionally, words such as "like" or "sure" might signify uncertainty among people who submit their reports. Lastly, "bee" and "wasp" are commonly used suggesting that, again, user might be reporting uncertain evidence of AGH.

Along with the user comments, much can be gleaned from the lab professionals. This is because many professionals might respond in the way that shines light into why the submission was a negative ID. Looking at the common words, "wasp", "sawfly", "digger", "golden", "horntail", "cicada", and "female" appear to stick out. This could represent many of the misconceptions among AGH as users are mistaking them for these other insects or overlooking certain features.

After looking at the dataset, we can confirm the following trends:

- Users commonly mistake *cicada killers*, *golden digger wasps*, *sawflies* and *horntains* as AGH's
- Large insects such as wasps, beetles, and horntails appear to be mistaken just by size
- There are many users unsure of their findings and submitted anyways

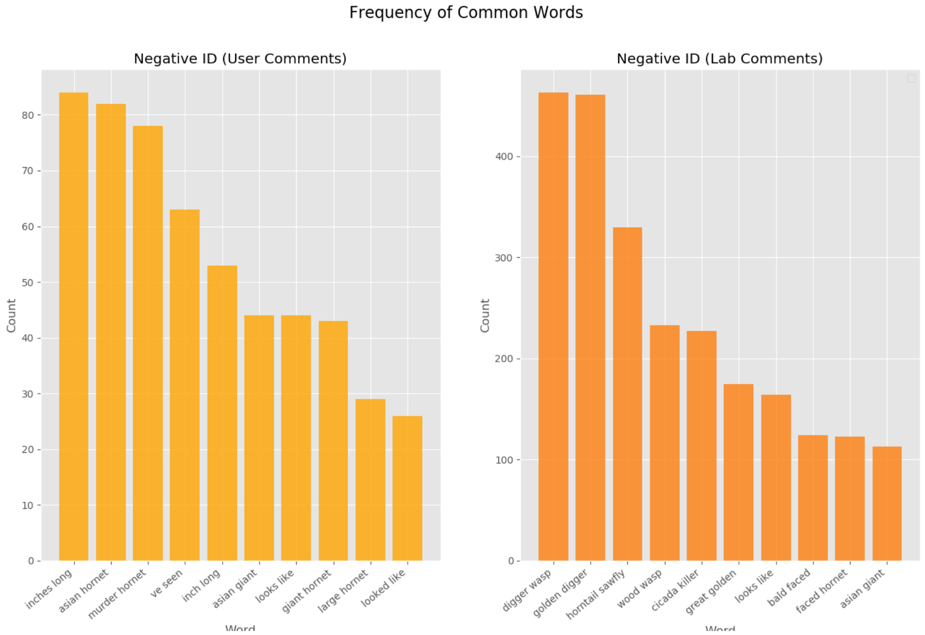

Figure 3: Negative Cases.

Lastly, we will consider the unverified cases. Notice that we are not exploring unprocessed findings there is a small number of samples and less information since lab professionals have not review them. We can view the unverified cases as a bar chart like above to see if there are any correlations within the class. This is depicted in Fig. 4.

On the user side, we see many users citing "large," "long," "orange," "black," "picture," and "flying." Most of these are characteristics of AGH's and warrant further exploration. We also see many lab professionals responding with "photo," "resubmit," "county," "negative," "feel," and "free." Looking at the keywords, it appears that unverified findings are correlated with a lack of a photo, negative counties, and submissions warrant a resubmit.

After looking at the dataset, we can confirm the following trends:

- Users who do not submit a photo can not have their sighting verified
- Counties negative for AGH's make up a large portion of unverified findings.

4 Classification

Due to the class imbalance problem noted in section 3.1, a direct supervised model seemed infeasible. Rather, we opted for finding trends within the negative ID prediction class to shine insight into common errors among users. For this reason, we decided to use *unsupervised learning* as our model basis.

4.1 Data Pipeline

For our data, the following features were Notes, Latitude, and Longitude. The global ID provided no statistical meaning and the lab comments were made from professionals. Since we want to create a model to try to help save resources for the government, we made the assumption that the lab comments would not count towards a feature. We also omitted the pictures as it would require

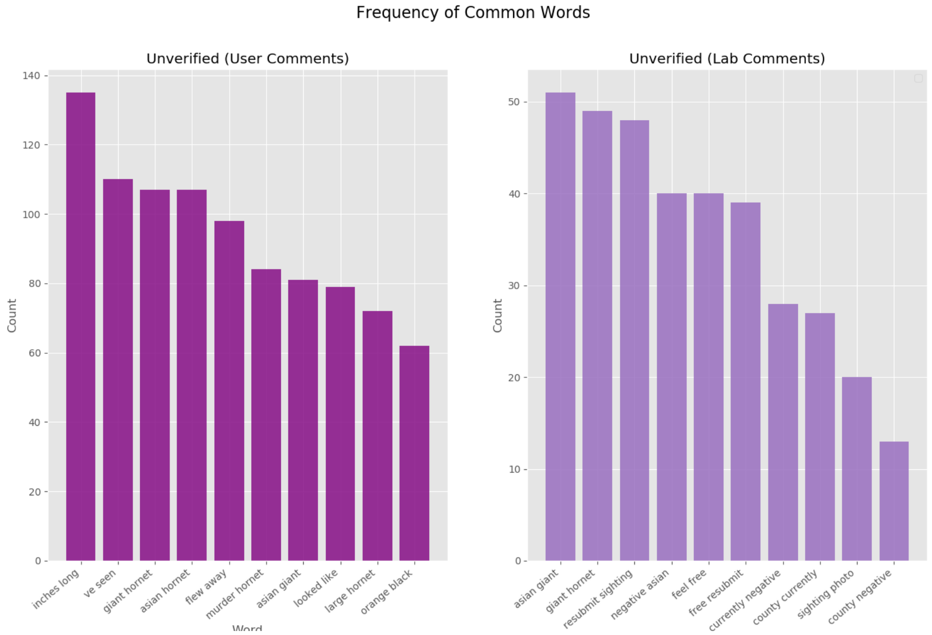

Figure 4: Unverified Cases.

too many resources given the small dataset size. Thus, the defining features in our data are Notes, Latitude, and Longitude.

Here, Latitude and Longitude are numerical values while Notes are strings. Thus, we need to put the notes in a numerical form to work as features. For that reason, we chose to use CountVectorizer[2], a tokenizer that counts the occurrences in a sample of text and encodes the i^{th} word as the i^{th} feature. This is done for each sample to create an $n \times m$ matrix, where n is the number of samples and m is the number of distinct words across all samples. Once this matrix was created, we concatenated it with the other features across the column axis.

After concatenating, we noticed due to sparsity of the word features, we needed to scale our features. This is especially important for distance based models such as KNN or SVM. It is less so for Naive-Bayes but still is common in practice to transform the data. We implemented this pipeline for each of our models.

To address the class imbalance, we utilized *Synthetic Minority Oversampling Technique* (SMOTE) to resample our training data so each class would be given equal weight. In addition, we also penalized our model more when misclassifying the positive cases in order to extract more information. Note that this problem is especially hard for SMOTE as the positive cases represent such a low proportion of our overall data set.

Lastly, to ensure robustness, we ran an exhaustive grid search over possible parameters with a cross-validation fold of five. This helped to confirm our choices for parameters and generally accept the performance of the given model.

[2] sklearn.feature_extraction.text.CountVectorizer

4.2 Direct Supervised Problem

4.2.1 Insect Classification

Before we implemented a full supervised learning model, we wanted to try to see if we could find discernible patterns among users who miss-classified AGH's as "golden digger wasps," "horntails," "sawflies," and "cicada killers." To do this, we iterated through the original dataset to find when lab professionals responded to users suggesting that it was one of these four insects. Once stratified, we labeled each sample to correspond to the mistake it created. This is depicted in Table 1

Table 1: Dataset for Insect Classification

Notes	Insect
slow moving, i have this in the freezer still ...	0
walking in our field observed these bugs hover...	0
i have had four of these i think. they are lar...	0
32nd and gennessee in west seattle 98126. ther...	0
single. on rocks. slow moving.	0
they were coming out of a hole in the ground, ...	0
landed next to us on the magnuson soccer field...	0
next door in a house under construction.	0
⋮	⋮

After doing this for each insect, we combined our data into a single dataset. Then, we used this as a supervised problem and sought activation among certain insects. Since this new dataset contained only the description of the sighting and we wanted to not overfit our data, we decided to use a *Multinomial Naive-Bayes* model introduced below in Eq. 1.

$$p\left(x \mid C_k\right) = \frac{\left(\sum_i x_i\right)!}{\prod_i x_i!} \prod_i p_{ki}^{x_i} \tag{1}$$

where $x = (x_1, x_2, \cdots, x_n)$ represents a feature vector for word i. Multinomial Naive-Bayes was originally structured so that the x_i entry represented the number of occurrences for word i. Among the hyperparameters to be tuned in our program, α represents an additive smoothing parameter, which we will find optimally through a grid search.

4.2.2 Linear SVM

Due to the small dataset, in order for us to have a robust model, we must consider a model that has low complexity but suffices in generalization. We decided to choose a linear *Support Vector Machine* (SVM) due to previous literature accomplishing a similar issue with a small, imbalanced dataset in NLP [6].

We also wanted to test various models to validate the findings in the paper. Doing so, we also tested *logistic regression* and *Gaussian Naive-Bayes*, both of which did not result in better classification. Within the linear SVM, the only hyperparameter that needed to be optimized was C, each in some sense a normalization factor. To find the optimal choice, we ran an exhaustive grid search over the choices of $C = [0.1, 1, 10, 100, 1000]$ and found slight change in results, albeit an optimum.

4.3 Unsupervised Clustering

To determine if there were features common to the negative class, we chose to isolate the negative cases and use a clustering algorithm. After embedding our words as vectors and concatenating them to the other features in our dataset, we had a high dimensional dataset. We then scaled our dataset due to the differing magnitude of each feature and scaled each sample to be unit length with respect to the Euclidean norm.

Using this, we utilized k-means clustering to give predictions for new samples. Additionally, we also explored the features associated with the clustering to provide a high level understanding of what patterns the model extracted. Lastly, as we are operating in high dimensions, we projected our clusters into $\mathbb{R}^2$ with both a linear (PCA) and non-linear (UMAP) model.

4.3.1 k-means Clustering

Using *k-means clustering*, we were able to identify prominent features within the reports that were classified as negative IDs. The k-means algorithm finds groupings of data points with similar characteristics by randomly guessing k centroids and assigning each data point to its nearest centroid. The algorithm then iteratively improves on the locations of the centroids until convergence or the number of maximum iterations is reached.

This particular clustering algorithm was chosen, because it was a simple starting point for finding general groupings of negative reports with similar characteristics. From k-means, we were able to obtain an initial visualization of the clustering of the data, but it had too many drawbacks. The k-means algorithm is limited to finding linear boundaries. Additionally, the number of groups, k, must be predetermined and the algorithm could not learn to determine the optimal number of groups. Using the sum of squared errors (SSE) we determined the optimal k.

Because the results are quantitative, a major drawback to k-means is that the results are often hard to interpret. We found it very difficult to determine what our clusters actually meant or if they were even meaningful at all. Without very many positive test cases to quantify a result that meant 'good', it was even more difficult to answer this question. Was it possible to define what each cluster represented? If a submission was actually a positive ID, would it occur outside of our three clusters that represented data classified as negative? Ultimately, this ambiguity led us to suggest using a *Gaussian Mixture Model* (GMM), a slightly more complex clustering technique with quantitative results.

4.3.2 Visualizing Clusters

To visualize the clusters, we projected our dataset into a lower dimension and labeled each point corresponding to the cluster. To do this, we first used *Principal Component Analysis* (PCA). This is a linear dimensionality reduction technique that takes the 2 (due to $\mathbb{R}^2$) most prominent singular values from a Singular Value Decomposition (SVD). In short, for a matrix A, the decomposition is the following

$$A = U\Sigma V^T \tag{2}$$

where U are the left singular values, Σ is the diagonal matrix of singular values, and V^T are the right singular values.

Notice that as this uses variance as a metric, scaling the data is needed as it would otherwise prioritize features with higher numerical values. Based off the distribution of the data, we witnessed certain outliers. To combat this, we decided to use RobustScaler[3] as it preserves the structure of

[3] sklearn.preprocessing.RobustScaler

the tail ends of the distributions but excludes values that are either above the 75^{th} or below the 25^{th} percentile for each feature. The final result is depicted in Fig. 9.

Next, we chose to use *Uniform Manifold Approximation and Projection* (UMAP) versus t-SNE due to us wanting to preserve global structure. This method is also faster than t-SNE and provides the same added benefit of non-linearity. The result from UMAP is shown in Fig. 10.

4.4 Incorporating Future Cases

Since the model we used to predict the likelihood of negative cases required to be trained, in order for the model to sustain robustness to a degree, the underlying distributions and relationships must remain the same [9]. This applies especially to our target distributions as currently there are not many positive cases. Over time, if AGH's grow in quantity, there will be more positive cases and less of a class imbalance problem, throughout off our model parameters. Thus, we should inspect both the input feature distributions as well as the target.

4.4.1 Comparing Distributions

To quantify if our distributions have changed, we can use statistical tests. Due to the Kolmogorov-Smirnov working with continuous data only, and considering that we have discrete data, we will use the *Chi-Squared* test to compare distributions defined in Eq. 3

$$\chi^2 = \sum_{j=1}^{k} \frac{(f_{b_j} - f_{e_j})^2}{f_{e_j}} \tag{3}$$

where f_{b_j} and f_{e_j} represents the observed and expected cases in bin j, respectively.

Here, we can use this test to compare distributions within the existing distribution of a given feature such as latitude or longitude. For vectorized words, where the occurrences were counted, we will have an $n \times m$ matrix, with n samples and m distinct words.

Given a new observed sample, we can use the vocabulary that was fit on our previous matrix and increment the count on word m if our new sample contains it, and extend our matrix to size $n \times (m+1)$ if it contains a new word. Then we can sum across the rows to compute the histogram, fix the bins, and use our test. We should do this for each new sample and **maintain our original distribution** until enough new samples cause the test to fail.

5 Modeling the Spread

Given some current distribution of positive IDs and/or the likelihood that unknown cases are positive, we present a physical model for determining which cases to prioritize.

5.1 Proposed Model

Consider the loci of the positive sightings of the AGH, on a globe. Because the area of interest is only a small portion of the surface of the globe, let us assume that we may project the longitude and latitude coordinates of sightings to equiplanar points within $\mathbb{R}^2$. Next, consider some rectangular domain, $\mathscr{D} \subset \mathbb{R}^2$. Further, let all reported longitudes and latitudes of sightings map to within this domain. Next, consider some $[n^2 \times 1]$ vector $\mathbf{v}$ which will discretize the continuous domain $\mathscr{D}$. If we grid $\mathscr{D}$,with $n-1$ equispaced lines in the x and then $n-1$ equispaced lines in the y direction, we are left with n grid rows, and n grid columns. Without a rigorous definition, this is a one-to-one

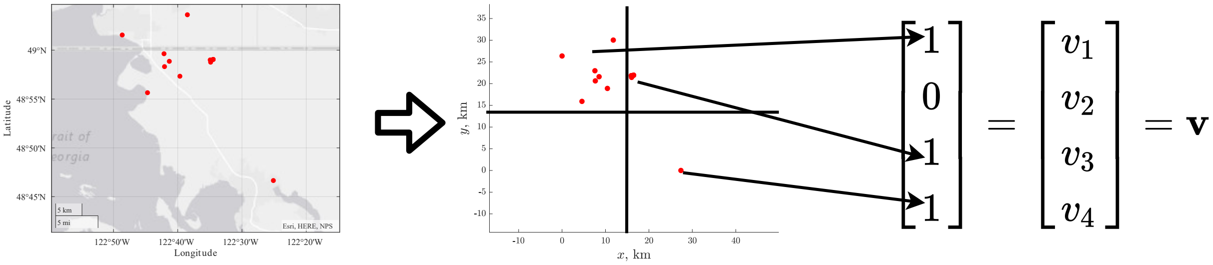

Figure 5: Transfer of a continuous domain to a discrete vector representation. This vector representation also corresponds to a gridded version of D, where 3 of the grid squares would have values of 1 and be displayed as yellow. See Fig. 7 as a higher-resolution example.

and surjective map from the grid squares to $\mathbf{v}$. The entries of $\mathbf{v}$ are given by Eq. 4. For now, think of $\mathbf{v}$ as a representation of where we have already or are likely to find AGHs.

$$v_{i+nj} = \begin{cases} 1 & \text{If the grid square of } \mathscr{D} \text{ at row } i\text{, column } j \text{ contains any positive reports} \\ 0 & \text{Otherwise} \end{cases} \tag{4}$$

Now, we create an undirected graph of nodes and edges, representing the possible movement and growth of AGHs. Every grid square associated with an element in D is a node. These are the locations that AGHs may be located at. Edges will connect neighboring grid squares, as AGHs may physically move across the boundaries joining any two grid squares [10]. Additionally, let each node be connected to itself, which represents the ability of AGHs to reproduce within their given area. This graph has an associated $[n^2 \times n^2]$ adjacency matrix, B, where all the edges are taken to have a weight of one. In summary, if $\mathbf{v}$ represents the current locations of AGHs, then $B\mathbf{v}$ represents the growth and diffusion of AGHs at some future time.

However, at this point, B simply models the potential spread over a small time frame. As defined by Eq. 5, A, a new matrix of the same dimension, more realistically takes the place of B, where a, b, and c are positive constants that we will soon discuss. Note that a and b are integer exponents on the matrix, indicating a matrix multiply operation, rather than an element-wise exponential.

$$A = B^a - cB^b \tag{5}$$

Eq. 5 gives a matrix that corresponds to a diffusion process of some initial state, from a time index t to a time index $t + 1$, as given by Eq. 6. The matrix in Eq. 5 models the spatial population dynamics of the AGH in a few ways. The positive B^a corresponds with AGHs growing and spreading throughout a surrounding area, representing an exponential growth term. Similarly, the negative term, $-cB^b$ captures the same dynamics of a density-dependent growth term. In an area filled with AGHs, we expect for them to exhaust their immediate environment, and to risk being exterminated after being detected. The likelihood of either of these events increases at the more populated epicenter, where this term is also centered. Similarly, AGHs do *not* nest communally [2], so future AGH nests will be in different areas, further justifying the dynamics in 6 of AGHs spreading out from the epicenter. The same dynamics are commonly used in the literature to model insects and invasive species, in the forms of the reaction-diffusion equation and the classic logistic equation [11], [12].

$$\hat{\mathbf{v}}_{\mathbf{t+1}} = A\mathbf{v}_{\mathbf{t}} \tag{6}$$

We will denote $\hat{\mathbf{v}}$ as the predicted spatial probability distribution, rather than the observed probability distribution $\mathbf{v}$. Of course, we wish for our model to accurately represent the observed data, so we are left with an error minimization problem, as defined by Eq. 7, where p denotes the p-norm.

$$\text{Minimize } ||\mathbf{e_t}||_p = ||\mathbf{v_{t+1}} - \hat{\mathbf{v}}_\mathbf{t}||_p = ||\mathbf{v_t} - A\mathbf{v_t}||_p \tag{7}$$

To accomplish this minimization, we must know both $\mathbf{v_t}$ and $\mathbf{v_{t+1}}$, while we tune a, b, c in Eq. 5 to accomplish this. The parameter a corresponds to both the reproduction and the expansion of the AGHs within some area, while b indicates the rate and area impacted by density-dependent reductions. The parameter c scales the relative importance of the exponential growth term (B^a) versus the density-dependent term (B^b).

5.2 Tuning Model Parameters

Let us consider only three sections of time, $t = 0, 1, 2$, each of about 4 months in duration. This division is arbitrary due to the sparsity of the data, but the approach is easily generalized to different length durations. With these time-steps, Eq. 8 gives 3 error estimates using our available data.

$$\begin{aligned} ||\mathbf{e_{0,1}}||_p = ||\mathbf{v_1} - \hat{\mathbf{v}}_\mathbf{0}||_p = ||\mathbf{v_0} - A\mathbf{v_0}||_p \\ ||\mathbf{e_{1,2}}||_p = ||\mathbf{v_2} - \hat{\mathbf{v}}_\mathbf{1}||_p = ||\mathbf{v_1} - A\mathbf{v_1}||_p \\ ||\mathbf{e_{0,2}}||_p = ||\mathbf{v_2} - \hat{\mathbf{v}}_\mathbf{0}||_p = ||\mathbf{v_2} - A^2\mathbf{v_0}||_p \end{aligned} \tag{8}$$

Let us denote these errors as $\mathbf{r} = [r_1, r_2, r_3]$ respectively. We must assign a weight to each of these cases, as described by $\mathbf{w} = [w_1, w_2, w_3]$, and generate an error function $E_p(a, b, c)$ given by Eq. 9. Similarly, these errors are all normalized by some arbitrary case, so they are of the same magnitudes.

$$E_p(a, b, c) = \frac{\mathbf{r} \cdot \mathbf{w}}{||\mathbf{w}||_p} \tag{9}$$

By choosing different weights, and finding the optimal parameters, we may examine how susceptible our model's parameters are to the weight function. We may repeat the same process with different norms, as the data used for tuning essential consists of *only* outliers. A brief study of the weights revealed that for the Euclidean norm $p = 2$, and for a more outlier sensitive norm $p = 3$, $a = 8$, $b = 7$, $c = 0.05$, yields the lowest error, for a range of nonzero weights, over a discretization of $n = 30$. As mentioned in the assumptions, this was done in a subdomain of about 10000 km^2 within Washington State. From this probability distribution, choose to investigate cases that are located in the highest probability areas first.

5.3 Random Model Add-On

In addition to the above Matrix Model for locating those cases likely to be positive IDs, we now propose an additional portion, with the intent of better accounting for outliers.

The Matrix Model provides a reasonable procedure for how to prioritize investigating cases within the areas immediately surrounding past positive IDs. However, with the small number of positive IDs, it is nearly impossible to predict the locations of positive IDs occurring far from previous positive IDs. For instance, the positive ID in Canada is 80km away from any of the positive ID cases in Washington, while the estimated range for a new queen establishing a new nest is only 30 kilometers [13]. Similarly, genetic analysis suggests that these may have been separate introductions of the species. This frightening coincidence hints at the possibility that we do not fully understand the situation, and thus we can not form reasonable simplifying assumptions for the development of a fully deterministic model.

With this in mind, we unironically suggest a portion of the model to consist of **random sampling**. For 15% (or 1/6) of the prioritized cases, we suggest randomly generating a point that is 20 km to 40 km away from the epicenter, and reviewing the case closest to this point. Additionally, for 15% (or 1/6) of the prioritized cases, we suggest repeating this procedure, but investigating cases farther than 40 km away. The first range of distances usually falls outside of the Matrix Model and includes the range that a new queen may travel to establish a new nest. Next, we have already seen the two apparently distinct introductions of the *same non-native species*, in the *same area*, within a couple of years. Thus, the larger random search is a precaution for factors that are beyond the consideration of *any* reasonable, physically-derived model.

Lastly, this method of random selection targets the investigation of cases over a larger area, as opposed to focusing resources in small areas that are more populated and thus more likely to have a higher density of reports. It is important to turn outside of the 40 km range, as if a positive ID case is left undiscovered for a whole reproductive season, the AGH problem may explode in a new area. This would suggest a more thorough investigation of this new area, causing government resources to be spread even more thin than they already are. These beliefs motivate us rounding the values from Table 2 to favor more random sampling.

6 Combining Models

As hinted above, we may use a vector of probabilities as an input into the Matrix Model, thus combining our natural language classification model with our physical model. For any report in the test case category, the natural language classification model gives a probability that the report is a positive ID. These probabilities are added to the vector $\mathbf{v}$ based on their physical location. Then, the model in Eq. 6 combines these spatial probabilities to produce a new probability distribution like Fig. 7, which is relevant for the next couple of months. In the same way as in Sec. 5, this probability distribution may then be converted to a prioritization of certain reports over others. The same approach may be applied to combine the Matrix Model with a model that analyzes the photographs.

7 Results

In this section, we discuss the results of the models presented in Sec. 4 and 5.

7.1 Matrix Model for Positive ID Cases

Next, we compare our model results to our known time-steps. Because this model has a very physical basis, we are not as concerned about it overfitting the inputs used to tune the parameters. Fig. 6 shows the comparisons described for each of the possible error metrics previously mentioned. Fig. 7 gives an updated prediction from our model, for one time-step, including data from each of the time-steps. It is reasonable to test this model on the tuning data. We only adjusted three degrees of freedom, so we pose little risk of being able to fit all 16 degrees of freedom describing the positive IDs (x and y coordinates for 8 points), in addition to the degrees of freedom provided by the *lack* of positive IDs in other locations. Fig. 6 also reveals the prioritization. **Our model prioritize the positive IDs as the 2nd and 3rd reports out of 10 for the first time-step,**

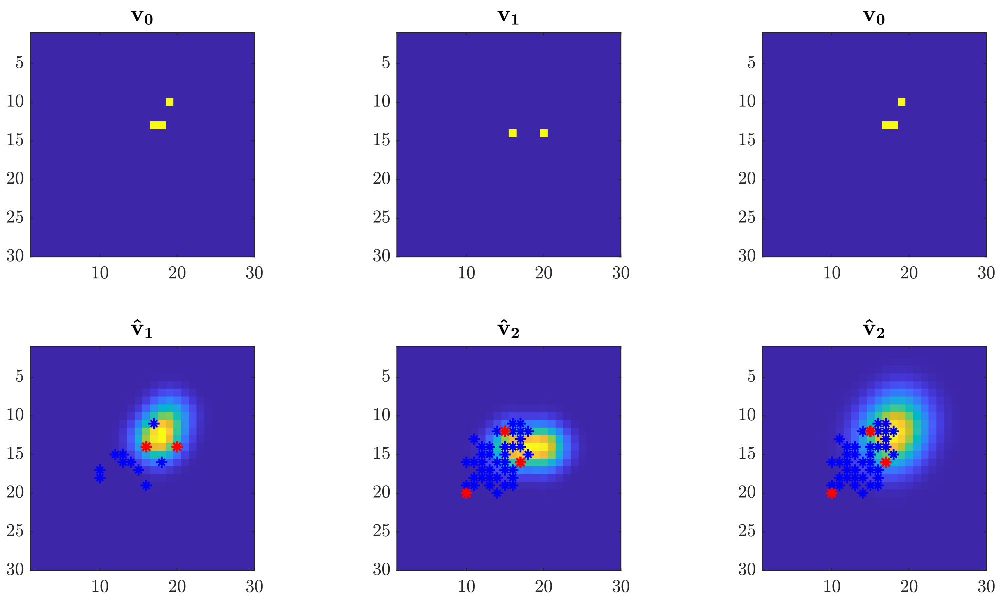

Figure 6: Top row: visualization of vector inputs into the model; Bottom row: predicted probability distribution output. Each column represents an error case, previously defined. Red stars are areas with positive IDs and blue stars are areas with negative IDs.

and 7th and 11th out of 38 for the second time-step, and 6th and 12th out of 38 for the second approach to the second time-step. This is done while omitting the positive ID outlier in the second time-step. Certainly these prioritization orders quantify physical intuition and are significantly better than random prioritization.

Next, we briefly turn to an analysis of the spatial distribution of our results as compared with what is observed. Refer to Table 2. Quantitatively, 85% of positive ID cases are within 20km of another positive ID case, and our model predicts 99.5%, 95%, and 85% of cases to be within this radius, for 1, 2, and 3 time-steps respectively, (4 months, 8 months, 1 year). Further, our model predicts that 98.7% of AGH generally stay within a 30km limit for 3 time-steps (1 year), corresponding to the current understanding of AGH given in [2]. It is worth noting that the distance of positive IDs to other positive IDs has no obvious correlation with time. That is, the distance from a recent positive ID to the immediately previous positive ID is 2.06 km *farther* than the average distance to *all* of the previous positive IDs. This average to all the previous IDs also has an average standard deviation of 10.1 km, which suggests that there is more noise than signal in the overall time-series. Thus, the metrics from the data were calculated without considering the time at which the positive IDs were observed.

Table 2: Distributions of Distances between Positive IDs.

	$<$ 20km	20-40km	$>$ 40km
All Positive IDs	73%	13%	14%
Positive IDs in Washington Only	85%	15%	0%
Probability of Positive IDs from Prediction (4 months)	99.5%	0.5%	0%
Probability of Positive IDs from Prediction (8 months)	95%	5%	0%
Probability of Positive IDs from Prediction (1 year)	85%	15%	0%

Testing the random portion of the model would not provide meaningful results, as it was in fact

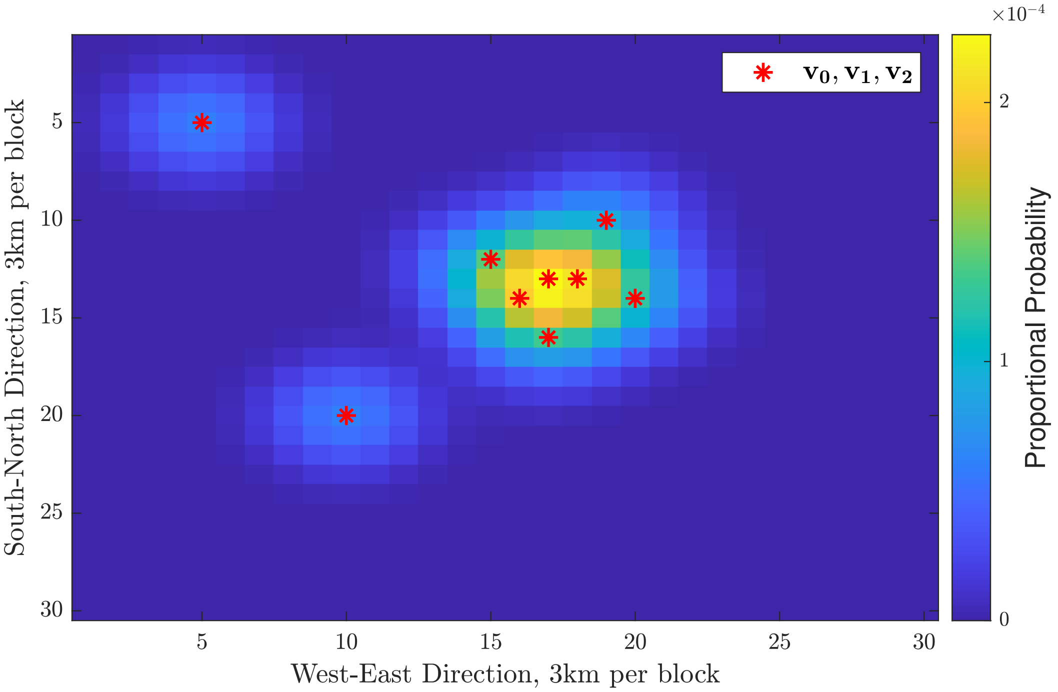

Figure 7: Given the available positive ID, this is the predicted probability distribution of positive IDs, for the next 4 months.

Table 3: Classification Report for Multinomial Naive-Bayes

Insect	precision	recall	f1-score	support
Golden Digger Wasps	0.59	0.77	0.67	56
Horntail	0.12	0.11	0.11	38
Sawfly	0.11	0.10	0.10	42
Cicada Killers	0.47	0.32	0.38	25

developed by considering the proportions given in Table 2. For the development of the model, the proportions from this table were rounded to add extra weight to cases outside the 20 km radius. As previously mentioned, if these cases go undetected, they have the potential to greatly worsen the problem.

7.2 Classification Results

7.2.1 Insect Classification

The results we got showed predictive power for golden digger wasps but not for other classes. This is shown in the classification report from sklearn in Table 3.

This resulted in an overall accuracy of 37.0%, slightly better than random. We also had trouble generalizing the model as our training accuracy was in the range of 60-70%. This demonstrates that our model was still overfitting across all classes despite efforts. Additionally, since this model had only α as a parameter, we tuned it by an exhaustive grid search resulting in a value of $\alpha = 0.0001$.

Nonetheless, since our end goal should not be to classify which common mistake the user created, this result serves as a baseline and validation within each mistake. That is, either use this result to

Table 4: Classification Report for Support Vector Machine

Class	Precision	Recall	F-score	Support
Negative ID	1.00	0.99	0.99	206
Positive ID	0.40	0.67	0.50	3

Table 5: Classification Report for Support Vector Machine

Class	Precision	Recall	F-score	Support
Negative ID	0.99	0.99	0.99	618
Positive ID	0.43	0.43	0.43	7

understand that there might be underlying patterns within the descriptions among golden digger wasps or use it to validate a classification that appears to describe golden digger wasps.

7.2.2 Linear SVM

To predict if the sample is negative or not, we now consider additional features such as latitude and longitude and treat it as a binary classification problem with major class imbalance. Passing it through the same pipeline, we decided to implement a more complex model due to the added information. For that reason, we chose to use a Support Vector Machine with a linear kernel. Doing this, we arrived at the following results in Table 4. These results were formed from doing an exhaustive grid search across the parameter, C, giving us the optimal values of $C = 1000$, respectively. It also achieved an overall accuracy of 98%, which is essentially meaningless due to the class imbalance. However, for the lack of data, the results are still telling as our Recall is also 67%, meaning that if there was a possible case of an AGH, the model would have picked it up. Although, with a support of 3 samples, it is hard to tell how robust the model is at Postive ID cases even if it is a test set.

Using the insect classification as a baseline, we can include the categories that matched descriptions to the four insects as features. This act of feature engineering serves to add complexity to our data so that we can provide more information for our model. Due to not only including the other features such as latitude and longitude but also introducing the insect classes, we can increase the complexity of our model further. Specifically, we tested non-linearity by introducing an 'rbf' kernel. From this, we expected a slight increase in accuracy among positive classes along with better recal; however, we found that the result was not significant and did not provide better accuracy. This is shown in Table 5.

This gave us an overall accuracy of 99%. We validated this by implementing a grid search over $C = [1, 10, 100, 1000]$, $\gamma = [0.01, 0.001, 0.0001, 0.00001]$, and kernel = ['linear', 'rbf']. We also weighted the positive ID class with a ratio of 4:1 and saw no change in accuracy. Due to this result, we suggest using a linear kernel and a C value of 1000.

7.2.3 Clustering

Our goal was to determine if there were any defining similarities among the negative class by using a clustering algorithm. We chose the k-means algorithm, because it was a simple model. First, we vectorized the words so that we could consider latitude, longitudes, and dates as additional features for determining clustering, since our data was very limited. Then, after running k-means for a range of k values, we determined the optimal value to be $k = 4$ using the sum of squared

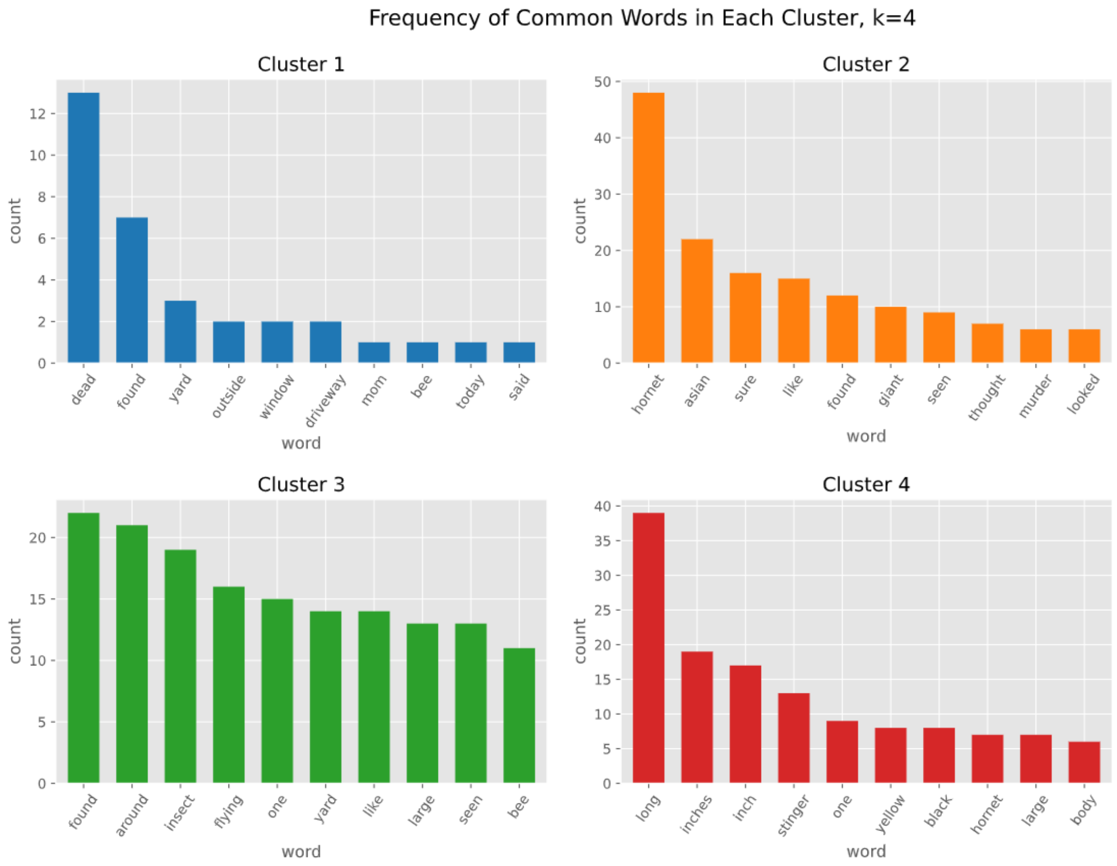

Figure 8: Prominent features in clusters determined by k-means algorithm with $k = 4$

errors (SSE). The word features in each cluster are shown in Figure 8.

When supplemented with dimension reduction techniques, we were able to visualize the clusters produced by k-means. Using PCA, which is linear, we obtain a mapping that has clusters, but appears only partially separable. This is shown in Figure 9. Because PCA did not look idea, we decided to try UMAP, a nonlinear dimensionality reduction technique, to see if we could get better results. We were able to visualize 4 distinct clusters, which suggests nonlinear properties of our data 10. A major drawback to k-means clustering is that results are qualitative and slightly ambiguous. We were unable to implement Gaussian Mixture Modeling (GMM) successfully within the given time-frame, but believe that the more complex clustering technique could offer more insight into the features which define the clusters, as it returns quantitative results.

8 Discussion: Strengths, Weaknesses, and Sensitivity

Next we discuss the strengths, weaknesses, and robustness of each of our models.

8.1 Matrix Model

The Matrix Model for the spread of AGHs is linear, physically-derived, data-informed, and general. However, the model fails to account for the critical presence of outliers, is untested over a long timeframe, and excludes the potentially helpful information of other data classes.

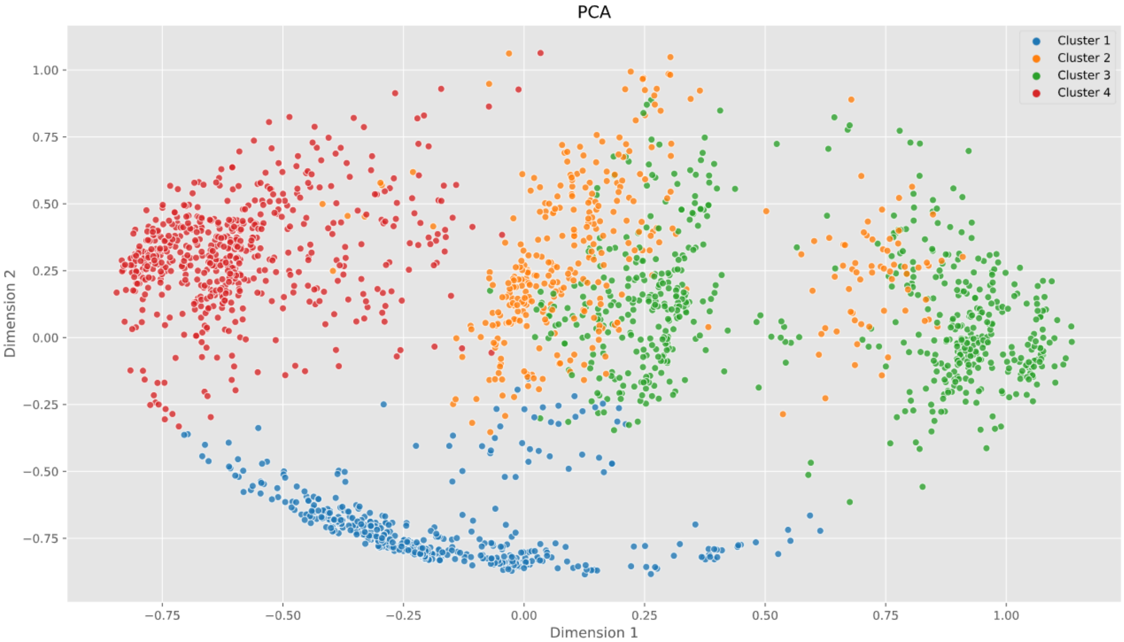

Figure 9: Notice that although there appears to be some pattern of clusters, it does not appear completely separable. This is evidenced by some word overlap in the clusters determined by k-means.

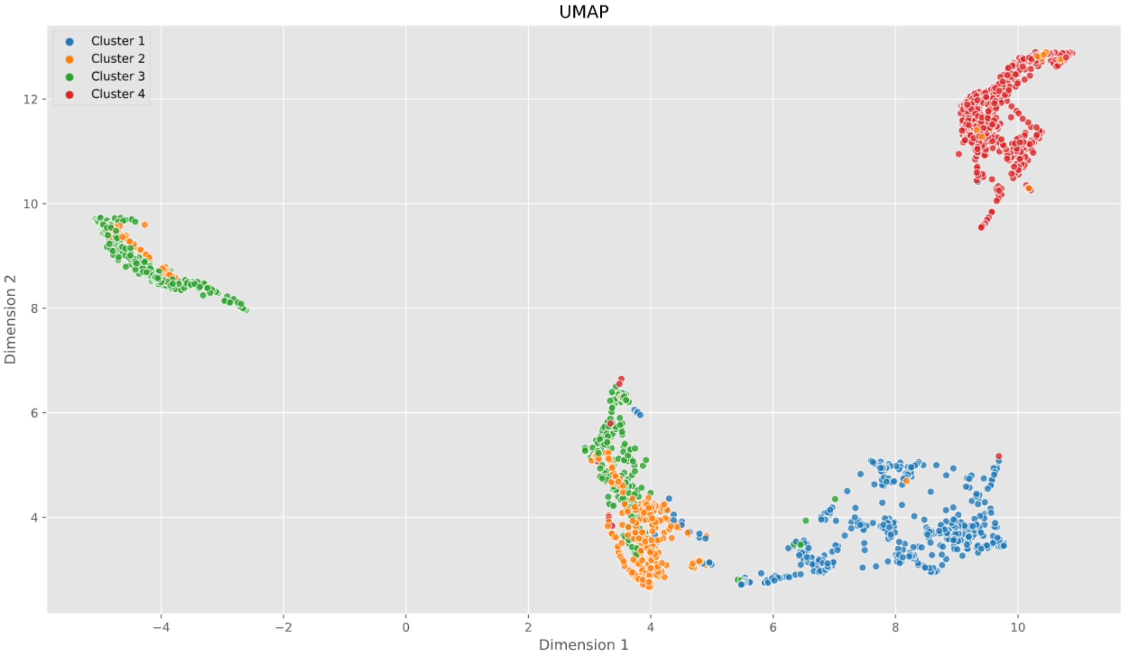

Figure 10: Notice that UMAP was able to distinguish certain clusters from the group, suggesting that non-linearity provides additional information.

8.1.1 Strengths

The model was physically derived to mirror the reaction-diffusion differential equation, which is commonly used to model insect populations *and* invasive species, [11]. This modeling approach *prevents the model from overfitting* our sparse data. Arbitrary parameters are calculated by using the available data, making this a **data-informed, physical model**. Our model provides a heirarchy of locations to view reports from when trying to find new positive IDs. This model provides quantitative justification for what is otherwise "physical intuition", allowing for experts to algorithmically prioritize investigating certain cases.

Additionally, our model is **linear**, allowing for new data to be *easily* added to an existing prediction. A new positive IDs would be converted to a vector as shown in Fig. 5, and Eq. 6 would be applied, with the matrix already constructed. Then a weighted sum of the current and new probability distributions would give an updated distribution. The weighting is simply a ratio between the number of new positive IDs and the number of old positive IDs. Similarly, old cases can be removed from the prediction by negating the original vector input, and following the process to update the prediction. Thus, updates can occur at a high frequency with simple addition operations as updates. Further, the matrix forming our model can be easily updated. The matrix represents a graph based on geographical location, so reproductive ratios, and movement factors can easily be added, and the same tuning process will allow for good parameters to be selected.

Lastly, our model is very general, and the input vector $\mathbf{v}$ can easily be updated to be a probability vector, based on other models. For instance, a machine learning model may assign probabilities that a certain report is a positive ID, based on the image and/or language that is used in the report. This model can use probabilities as inputs, and generate a full distribution of probabilities over a large domain.

8.1.2 Weaknesses

However, our model is also subject to a number of weaknesses, stemming from the lack of available data. First, the model vastly fails to predict the presence of any **outliers**. Similarly, with all data being reported over the past couple of years, the model is *untested over a long time frame*. Similarly, the model fails to capture the developmental stages, reproduction dynamics, genders, and other aspects of the AGH. Developmental stages and reproductive dynamics may be captured in a Leslie matrix [12], which may be integrated with our current model. This is an easy adaption, as our model is equivalent to a Leslie matrix, but over space, rather than stages of development.

Additionally, for extremely sparse sightings of AGH, the model simply predicts that more AGHs are nearby, which matches our physical intuition. With a couple of sightings within some area, our model becomes more helpful in predicting preferred locations to pick. Next, our model is much more of an exponential growth, rather than logistic growth model, so it will quickly fall apart if AGH density begins to approach some carrying capacity of an area.

Lastly, our model fails to consider the data presented by the occurence of *negative IDs* within an area. With more time, another term would be added to decrease the predicted likelihood of a positive ID occuring in an area with many negative IDs. Negative IDs often indicate the presence of other species occupying a similar niche. Similarly, negative IDs occur in high quantities in urban areas, while the AGH is more common in rural areas [2].

8.1.3 Sensitivity

By construction, our matrix in Eq. 5 has a source term, so a nontrivial steady-state prediction does not exist, and can not be examined for stability. For this model, perturbations from the trivial state result in a non-trivial solution for the rest of time, suggesting that eradication is seemingly impossible. To this end, *we suggest following this method, including the random sampling approach, until significantly more data is available to declare AGH eradicated.*

The Matrix Model discretizes a continuous domain, which inherently carries some degree of error. However, this discretization can be chosen arbitrarily, with just a need to adjust the model parameters through the same minimization problem (Eq. 7). Reports are grouped into these discretizations as a method to better fit our assumption that the male AGHs do not reproduce independently and do not stray far from their female. This way, reports of multiple male AGHs within the same area will be treated as just the existence of one female. In this way, discretization actually provides an advantage over an equivalent model on a continuous spatial domain.

The random method add-on trades accuracy to increase the robustness of the overall model. While the randomly generated locations may usually result in negative IDs, there is an added factor of safety by humbly acknowledging that *there is not enough data to create a model that is both accurate and robust for this phenomenon.* From Table 2, at least 30% of the positive ID cases will not be accounted for by the Matrix Model.

The same reasoning behind why we employed the random model beyond 40 km is why the problem of declaring the AGH eradicated is so difficult. Because we do not fully why two separate introductions occurred so coincidentally, there is no reasonable method for when to deem the AGH eradicated. *If there is some explanation about how the separate introductions are related, such as the same shipping company being linked to both of them, then the eradication problem can be considered from a more deterministic perspective.*

8.2 Classification Model

8.2.1 Strengths

For our classification model, we implemented numerous models (both supervised and unsupervised) to test our method. Namely, we first went through the vocabulary of each class and counted the occurrences of words (1 and 2-grams[4]). From this, we were motivated to implement a supervised problem for detecting if we could classify a sample to be one of the four most commonly mistaken insects. Noticing that one of the classes had high activation, we implemented feature engineering and ran a more complicated model, only slightly improving our results. Finally, we decided due to the class imbalance problem to use unsupervised learning techniques to provide insight into common trends of what might constitute a negative sample.

Throughout our paper, we prioritized practical solutions that a “black box” model would be less likely to provide. For instance, we give quantitative numbers on the mistakes users make and show that many native insects are mistaken for them. As such, a solution would be to focus on education among those four common insects. We also note that most unverified submissions lack a photo and so to save resources, a proper solution would be to flag it as unverified immediately unless coming from a verified source.

Another strength of our model are the resources to compute it. Naive Bayes and linear SVM are computationally favorable over say, a Convolutional Neural Network. This provides a benefit not only in dealing with samples that lack pictures or include poor resolution pictures, but also that as time progression, the model will be easy to retrain. One can even implement active learning for SVM so that incremental samples can strengthen the model.

[4] 1-gram and 2-grams are strings that are 1 or 2 words

8.2.2 Weaknesses

Within our model, the main weakness would have to be its automated predicting power. Unfortunately, we could not provide an accurate, robust model that could parse through samples and determine which ones were negative or positive. The model used also would provide probabilities for the linear SVM; however, when used with k-means, no such probabilities exist. In section 4, we suggested using a Gaussian Mixture Model due to the encoding of probabilities in a clustering fashion. However, even that would not provide *the likelihood of being in any one cluster* and so is forced to assign the point to some cluster, even if it is further away from all of them.

Additionally, our model uses the user's notes and location as its main predictor; however, most submission have little writing or no writing at all. Because of this, it is hard to quantify predictors without the use of a spatial model coupling our predictions.

8.2.3 Sensitivity

Regarding sensitivity, our data pipeline provides a robust way to implement new samples over time. In it, we scaled our data mitigating the effect from outliers. We also determined our model through multiple passes and did exhaustive grid searches for our hyperparameters, giving us an optimal value for C to be 1000. With a high C value, our SVM is also more likely to separate classes in a smoothed manner and allow perturbations with samples near the margins. This prevents overfitting within the SVM.

We also made sure to regularize each of our models by altering parameters so that overfitting is less likely. Although in the end our training accuracy was still above the test accuracy, we showed that we tested many combinations of parameters and that our model is indeed tuned. We also used linear models rather than non-linear ones due to the complexity of our dataset. Doing this, provides a model that extracts key information without overfitting. We also noticed that the model was not underfitting as well due to its classifying the positive cases and negative cases above random probability.

9 Conclusion

Penultimately, we review the main points of the preceding contents. Our classification model provides an estimate for the probability that a report is a negative identification, although the data set was too small to confirm robustness with full certainty. Our Matrix Model provides some estimation for the areas most likely to contain future positive AGH identifications based on current positive identification locations and times. Using an integrated model, we are able to first determine the probability that a submission is classified as negative and use the complement of this in our Matrix Model. Additionally, by examining the spatial distribtuions of positive IDs, we determine that 2/3 of the prioritized cases should be chosen using probabilities from the Matrix Model, 1/6 should be selected at a medium distance from our model's predicted epicenter, and the final 1/6 should be selected randomly beyond this area to account for unpredictability in the spread of the AGHs. Without more information, it is unreasonable to establish any criteria for eradication. Finally, we conclude with warnings, suggestions, and insights for the WSDA and citizen scientists in WSDA.

10 Memorandum

To the Washington State Department of Agriculture (WSDA) and other concerned parties:

In 2019, the Asian Giant Hornet (AGH) was discovered in the United States, for the first time. Sums of 1 million and 1.3 million USD have already been allocated for the US Department of Agriculture to aid in the eradication of the AGH and the protection of honey bees respectively. The AGH threatens honeybees, which provide value of nearly $20 billion to U.S. crop production alone [2], [3]. Further, invasive species can cause a cascade of other unforeseen consequences [11], [12]. It is an understatement to state that **the knowledge and eradication of the Asian Giant Hornet from Washington State and from the US, is of utmost ecological and financial importance**.

The current citizen scientist program for reporting possible AGHs is a step in the right direction, although there are changes to make this program more effective. Based on our analysis of the available data, we suggest two overarching objectives for the WSDA.

1. Reiterate critical information to citizen scientists.
2. Prioritize certain reports over others.

10.1 Critical Information for Citizen Scientists

A citizen scientist program is a massively powerful tool for collecting data and performing research. Its efficeincy can be further increased by relaying some simple information to concerned citizens. Thus, we recommend that the WSDA further elaborate and publish the following information, directed towards citizen scientists in Washington State.

10.1.1 Similar Native Species

Citizen scientists should be made aware of insect species that are often mistaken for the AGH. People commonly mistake cicada killers, golden digger wasps, sawflies, and horntails as AGHs, as verified by our natural language processing model. It would be useful to provide photographs of each of these, as well as descriptions of where they may be commonly found, especially emphasizing how this differs from AGHs.

10.1.2 Species Determination Flowchart

In addition to the above, an etymologist should construct a dichotomous key to aid in the determination of the sighted species. For instance, to differentiate an AGH from a European Hornet, the flowchart can ask, "is the abdomen banded yellow, black, and brown or is the thorax black closer to the head, will yellow and rows of black teardrops closer to the stinger?" [2]. Similarly, this flowchart may also include questions prompting other defining traits of the AGH, such as underground nesting and honeybee colony decimation. The report submission portal should be updated to prompt users to enter this information.

This will help gather more useful data, expiate species determination, and better inform citizen scientists.

10.1.3 Photograph Taking

Without a photograph, it is nearly impossible to identify and verify a the species of a reported sighting. Thus, a photograph should always be included. If the citizen can not take a photograph during the sighting, they should be encouraged to pursue the insect in order to get this photograph. One positive identification in a new area can prevent millions of dollars of damage.

Next, as per the USDA guidelines [1], photographs should include a side view of the head, and either a ruler or a coin for scale.

10.2 Report Prioritization

Resources are still spread thin, so there must be some method determining how to prioritize which reports to investigate. This section summarizes the suggested process for determining the order in which an entomologist should look through reports. Corresponding, this order relates to the order in which potential positive cases are further investigated in person.

10.2.1 Content of Reports

By considering a few features within the user submitted reports, our natural language processing model provides some estimate for the probability that the report will end up being a positive ID case. Similarly, reports from unverified sources, without photographs, are never conclusive. Thus, the reporting system should automate a message prompting the user to upload an image, while also marking the case as unverified.

10.2.2 Focus 2/3 of Investigation within a 20 km radius according to our Models

Given some distribution of positive cases, and/or the probabilities from our natural language model, our Matrix Model provides an estimation for the areas most likely to contain AGHs. This model is similar to the reaction-diffusion model [11], and it assumes that the AGHs spread from some epicenter in a predictable manner. This model quantifies and ranks cases based on the intuitive idea that "if we find one AGH, there are probably more nearby". We suggest that our model be used to rank the reports, and then that 2/3 of the prioritized cases are determined by following this ranking. This model is designed to be extremely easy to update with new input data, so it may be updated with each new report.

10.2.3 Choose 1/3 of Investigations to Protect Against Outliers

For the remaining 1/3 of prioritized cases, we suggest that half of the cases are selected between 20 and 40 km of our model's predicted epicenter (or nearby positive IDs), and the remaining half are selected at distances beyond 40 km. In 2019, AGHs were discovered on Vancouver Island, Canada, and experts believe that the AGHs in Washington were from a separate introduction of the species [2]. This is a stunning coincidence, suggesting that there may something deeper going on that we do not understand yet. Further, if an isolated population of AGH goes undetected for a breeding season or more, the problem could dramatically escalate in an entirely new area.

From this basis, we suggest that 1/6 of the prioritized reports are selected based on the nearest report to a random geographical location, at least 40 km from the epicenter. This method should be kept in place until we have a better idea of if something greater is going on. Similarly, we suggest that 1/6 of prioritized reports are chosen in the same way, but between 20 and 40 km from the epicenter. This second set of reports is to protect against rare occurrences of AGHs traveling farther than expected to develop new nests. Thus, we also suggest maintaining this process until significantly more data is available to make some analysis about eradication. Due to the lifecycle of AGHs, the minimum evidence would be the lack of positive cases for *at least* a year.

References

[1] USDA. New pest response guidelines: Asian giant hornet. https://cms.agr.wa.gov/WSDAKentico/Documents/PP/PestProgram/Vespa_mandarinia_NPRG_10Feb2020-(002).pdf, 2020.

[2] Penn State Extension. Asian giant hornets. https://extension.psu.edu/asian-giant-hornets, 2020.

[3] American Bee Keeping Federation. Pollination facts: Honey bees are pollinators. https://www.abfnet.org/page/PollinatorFacts, 2021.

[4] USDA. Usda provides more than $70 million in fiscal year 2021 to protect agriculture and natural resources from plant pests and diseases. https://www.aphis.usda.gov/aphis/newsroom/news/sa_by_date/sa-2021/national-ppa7721, 2021.

[5] WSDA. Detection eradication data. https://agr.wa.gov/hornets, 2021.

[6] Cristian Padurariu and Mihaela Elena Breaban. Dealing with data imbalance in text classification. *Procedia Computer Science*, 159:736–745, 2019. Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 23rd International Conference KES2019.

[7] Guo Haixiang, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

[8] nltk.tokenize package. https://www.nltk.org/api/nltk.tokenize.html, 2020.

[9] Luigi. The ultimate guide to model retraining. https://mlinproduction.com/model-retraining/, 2019.

[10] Stack Overflow. Construct adjacency matrix in matlab. https://stackoverflow.com/questions/3277541/construct-adjacency-matrix-in-matlab, 2010.

[11] John G. Alford. Mathematical models can predict the spread of an invasive species. 2019.

[12] Ellner and Guckenheimer. Dynamic models in biology. Princeton University Press, 2006.

[13] COMAP. 2021 mcm: Problem c: Confirming the buzz about hornets. https://www.comap.com/undergraduate/contests/mcm/contests/2021/problems/, 2021.