Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2102194

Pests in the Nest, Under Arrest: A Hornet Sighting Report Evaluation System

As infamous invaders to the North American ecosystem, Asian giant hornet (*Vespa mandarinia*) is devastating not only to native bee colonies, but also to local apiculture. Sighting reports of the pests have been collected from the public and investigated in Washington State, where positive reports provide clues to where their nests locate. This paper proposes an evaluation system to interpret and prioritize such reports, including the Hornet Spatial Migration Prediction Model (I), the Misreporting Probability Prediction Model (II), and Report Priority Evaluation Model (III).

For Model I, predictability of hornet migration is verified, and a prediction model is hereby proposed. The idea is that, based on the life cycle of hornet colony, the observed distributions should follow a temporal and spatial pattern. As has been verified on positive samples, temporally, hornet sightings mostly occur in September, which is a corollary of their habits, and spatially, the computed distances between sightings are within expected range. The distributions are determine afterwards with a Gaussian kernel.

For Model II, four independent features are extracted from a report: the sighting location, submission date, attached note and images. For location and date, the feature of hornet distribution probability is obtained following Model I; for image, the integrity feature s_{I} is extracted; for textual note, a **characteristics dictionary** for hornets is constructed, with which the similarity feature *w.r.t.* word frequencies q_i is extracted. Given the extreme imbalance in positive and negative reports, **logistic regression** and **weighted cross-entropy** are adopted in and a **gradient descent** method solves for the prediction model of mistaken classification.

The gradient descent method is also applied in **online learning** when additional reports arrive, so that model parameters can be updated. The frequency of such updates are set to *day*s/*week*s during September–November when hornets are most active, and *month*s otherwise, thus balancing the update overhead and response accuracy.

For Model III, due to the fact that reports close to positive others are more interested, the interrelations between reports are quantified as $F_{i,j}$, also combining the misreporting likelihood P_j obtained from Model II, to rank the corrected priorities Z_i of given reports.

The evidence of pest eradication in Washington State is derived using spatial isometric sampling and hypothesis testing. Reports within a colony cycle are sampled as D_n, to which the misreporting likelihood mapping p results in D'_n. The Central Limit Theorem and Student's t-Test are used for hypothetical testing of hornet's absence with confidence level α. Furthermore, absence of consecutive n years is discussed, concluding that $n = 3$ ensures the eradication of Asian giant hornet. Finally, sensitivity analysis is performed on two key parameters in the system. A memorandum for the Washington State Department of Agriculture (WSDA) is attached that briefly introduces the analysis.

Keywords: Iterative Simulation; Logistic regression; Hypothesis Testing; Student's t-Test

Contents

1 Introduction

1.1 Problem Background

Asian giant hornet (*Vespa mandarinia*), known as "Murder Hornet" in Japan, is highly aggressive and its swarm can massacre a bee colony in hours. Recognizing its near-destructive impact on existing honeybee species in North America, it has become vital to detect and predict the distribution of its colonies promptly. Accordingly, Washington State has established hotlines and websites for the public to report hornet[1] sightings, but not all the "sigthings" are true. In fact, only very few sightings have turned out to be *V. mandarinia*. With limited public resources, it is unwise to perform a detailed follow-up investigation of each report. Therefore, it is significant for the government to conduct *a priori* analysis of the text, images and other data submitted by the public to prioritize the reports that are most likely to be positive.

1.2 Problem Restatement

Considering the background information and given datasets as per requirements, we expect to address the following issues:

1. Develop a mathematical model to predict the migration of Asian giant hornet over time, with the prediction accuracy included.
2. According to the dataset with 4,440 reports of sighting reports and the *.rar* file with 3,305 images, build a classification model with analysis and discussion of the likelihood of a mistaken classification.
3. Based on the classification model, discuss how to prioritize the most likely positive reports for investigation.
4. Discuss how to update the model, given additional new report data over time, and how often the update should occur.
5. Construct the evidence using our model that would indicate the pest eradication in Washington State.
6. Write a memorandum including our results for the Washington State Department of Agriculture (WSDA).

1.3 Our Approach

Our work mainly includes the following steps:

- For Task I, we verified the predictability of hornet migration in both temporal and spatial dimensions, thereby constructing a probability model for migration prediction with different precision based on both the Gaussian spatial distribution and the hornet's habits.

[1]Unless otherwise stated, *hornet* in this paper refers to only the species of Asian giant hornet for simplicity, instead of the whole genus.

- For Task II, we first mapped the 4 features of a report, *viz*. the position, detection time, textual note, and image material, to obtain a feature vector that indicates the trustworthiness of the report; then we constructed a Misreporting Probability Prediction (MPP) Model based on logistic regression. For parameter fitting, given the variance of samples, we used modified binary cross-entropy with sample weighting terms as fitting metrics, and solved the model using gradient descent.

- For Task III, we obtained the probability of mistaken classification for each individual report using the MPP Model constructed in Task II. The probability (*i.e.*, the importance of the report), by quantifying the inter-corroboration among reports that are close in space and time, was corrected to obtain the priority ranking of multiple reports.

- For Task IV, upon the arrival of new data, we update the MPP Model built in Task II with the gradient descent method. Based on the habits of hornets, we setup two reasonable updating frequencies for the system: a low-frequency updating scheme on a monthly basis during hibernation, and a high-frequency one on a weekly/daily basis during migration and activity.

- For Task V, we considered no hornets detected for consecutive 3 years as eradication according to WSDA [4]. First, we collect a sample of reports from Washington State using spatially isometric sampling, then obtain the priority for each report using the model of Task IV, and perform hypothesis testing on the sample means. In order to monitor the hornet information in real time, the 36 months are divided into different numbers of time periods to perform hypothesis testing separately. Eradication is considered achieved if no hornet is detected in all the testing results.

2 General Assumptions and Model Overview

- **Assumption 1:** The reported data can approximately reflect the distribution of hornets around Washington State.
  ↪ **Justification:** Given that there were over 4,000 reports in the State within a year, we believe that the observation sample is sufficient to reflect the approximate distribution of hornet nests.
- **Assumption 2:** The growth and migration process of hornets is primarily determined by their habits, and is little affected by human interference.
  ↪ **Justification:** In light of the abundant reports, of which only a few proved to be positive, we consider human interference to have little effect on the growth and migration of hornets, and therefore consider only their habits and migratory patterns.
- **Assumption 3:** The more convincing the fields in a submitted report, the lower the probability of its mistaken classification.
  ↪ **Justification:** The main information in a report includes location, time, text, and possibly attached images, all of which are important evaluative information to determine if the report is mistaken classification, so the assumption is reasonable enough.

In summary, the whole modeling process can be shown as follows:

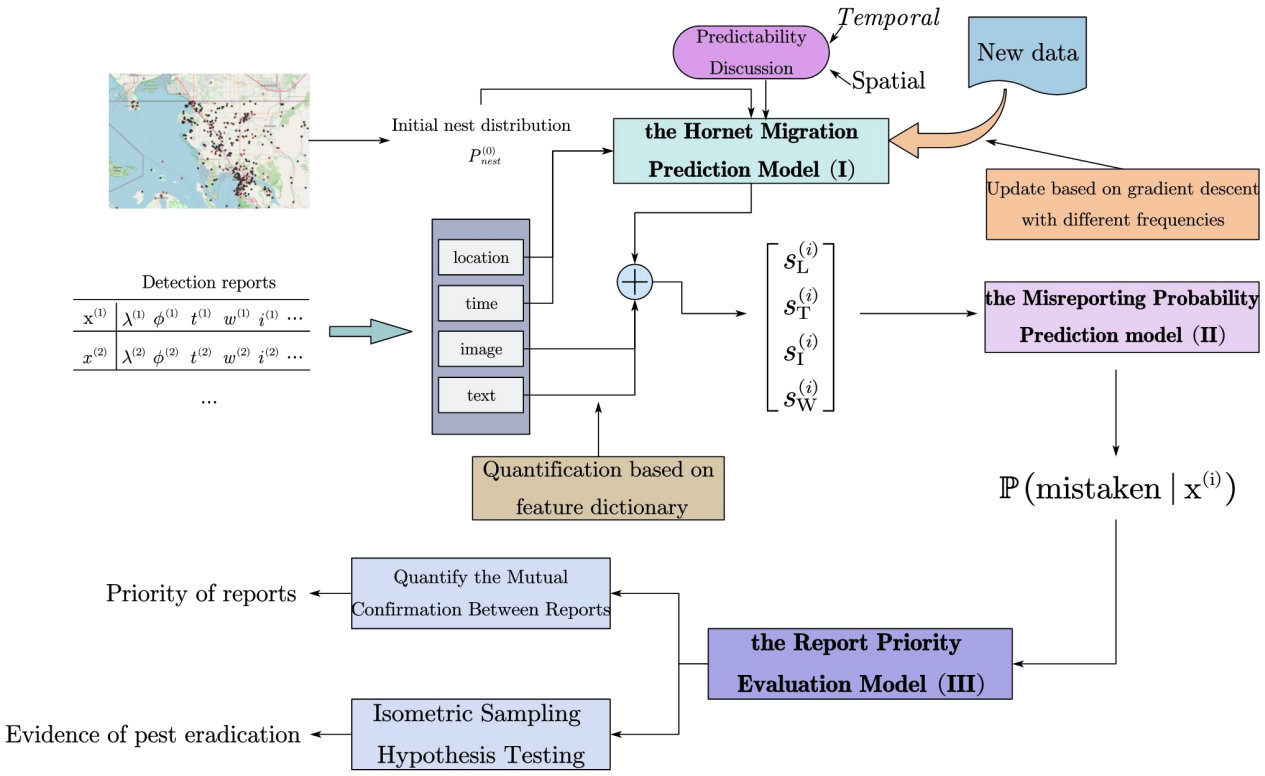


Figure 1: Model Overview

3 Model Preparation

3.1 Notations

Important notations used in this paper are listed in Table 1.

Table 1: Notations

Symbol	Description
$x^{(i)}$	The i-th report submitted by the public
t	Timestamp identifier (in month).
$\mathbf{s}^{(i)}$	The feature mapping of the i-th report.
(λ, ϕ)	The latitude and longitude of a submitted report .
$P^{(t)}_{\text{observe}}(\lambda', \phi')$	The probability of finding hornets at position (λ', ϕ') in month t.
$P^{(t)}_{\text{nest}}(\lambda, \phi)$	The probability of a nest existing at position (λ, ϕ) in month t.
$P^{(i)}_{\text{mistaken}}$	The likelihood of mistaken classification of i-th report.
Z_i	The quantification of the importance (*i.e.,* the priority) of i-th report.
$F_{i,j}$	The mutual influence factor between i-th report and j-th report.
α	The probability determining the confidence interval for complete eradication.
d	The distance.
P_i	The probability that the i-th report is positive.

3.2 Data Processing

In this chapter, we first deal with the outliers in the data, and the characteristics of the overall data are analyzed.

3.2.1 Data Cleaning

Among the 4,400 samples, the main informational features of each report were the coordinates, date, text, (possible) images, and officially given identification label. We excluded the following parts of the data: 15 reports with invalid date, 15 with "unprocessed" label, and 56 prior to 2019. These parts account for only a small fraction, and contribute little to our later study.

3.2.2 Overall Data Characteristics

After cleaning, we got 4,355 pieces of data. We first analyzed them with respect to their label categories. There are three categories of result labels in the cleansed data: positive, negative, and unverified. Their category-count relationship is depicted in Figure 2. Evidently, the number of sample categories is highly unbalanced, with only 14 sightings verified as positive cases, accounting for only 0.3%, and the remaining half of each case being negative and unverified, respectively.

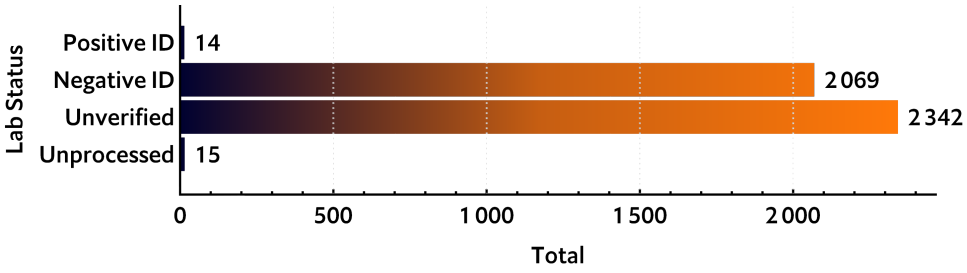


Figure 2: Status Count

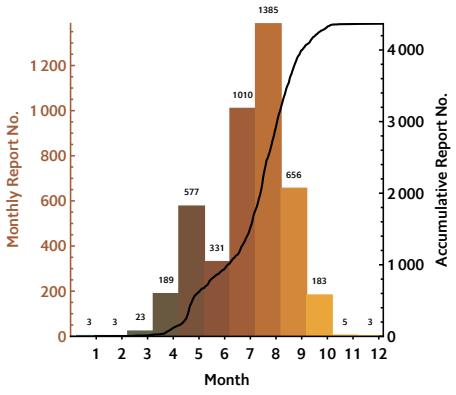


Figure 3: The monthly (brown, left axis) and accumulative (black, right axis) number of reports versus a certain month.

The dates were next analyzed. We found that 92.73% of the reports prior to cleanup were between 2018 and 2020, with 95% of them submitted in 2020. Considering that positive hornet sightings were not observed until recently, *i.e.*, after September 2019, and that the limited historical data did not provide sufficient evidence for prediction, we excluded data reported before 2019.

By analyzing the months in which the reports were submitted, we can obtain the distribution of the number of reports in different months, shown in Figure 3. It is observed that in winter and early spring (from December to March), only few reports are submitted; from July to September, there are a higher number of reports, peaking in August. The observed monthly pattern is consistent with the habits of hornets described in the literature [3]: the time period, in which *worker*s that constitute a large number in a colony actively

forage around, is precisely July and after, reaching a peak in August; after September, the *queen* starts reproducing the next generation of *female*s and *male*s, so fewer numbers are observed.

The textual information was then explored and the distribution of the word count in the report notes was computed, as shown in Figure 4. Over 15% of the reports saw *little to no* note attached (the word count was nearly 0), with 12% being *entirely blank*. Reassuringly, most of the witnesses were pleased to have left at least some descriptive notes, which provides the basis for extracting text features in our later sections. As the note length increases, the report counts show a pattern of *decreasing, followed by increasing, and then decreasing again*, well in accordance with our common sense.

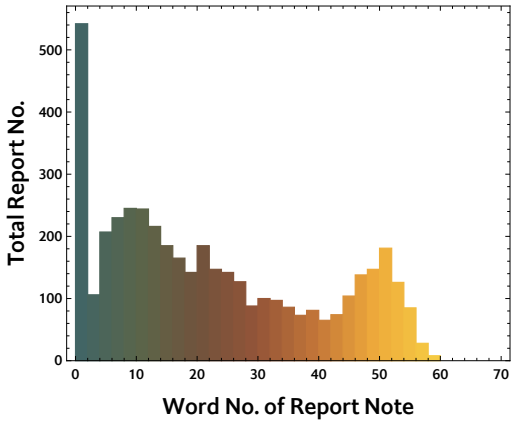


Figure 4: Note Length

Of the 4,355 reports selected, 2,127 (48%) came with associated images, indicating that about half of the witnesses attached relevant images to justify their sightings, among whom approximately 65% submitted only single attachment, with the rest submitting multiple ones. There were several types of the attachments submitted, but we had found that the results of the classification had little to do with their types.

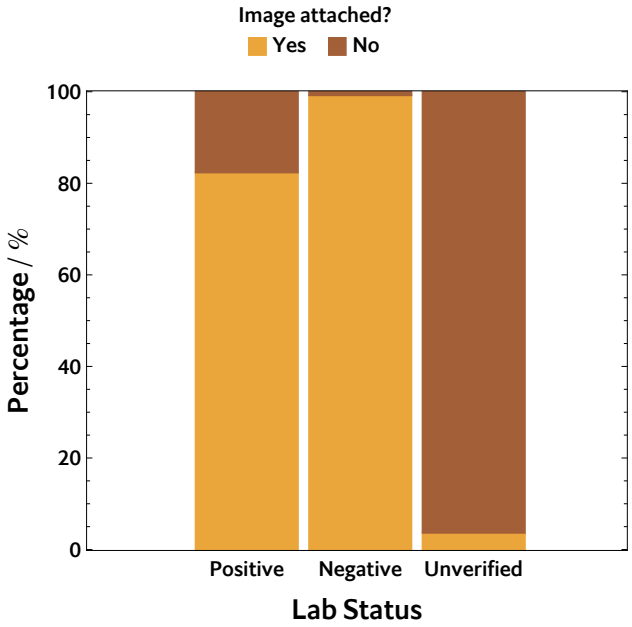


Figure 5: Classifying label versus the percentage of whether or not images were attached.

After grouping the reports by whether or not supporting images were attached, the presence or absence of images turns out to be a key factor in the official decision to head for inspection, as shown in Figure 5: in reports where images were attached, the official mostly chose to verify them; otherwise, the official mostly chose not to. This informs us that the presence or absence of images is an essential metric for recognizing a mistaken classification.

4 Model I: Hornet Spatial Migration Prediction Model

In this chapter, for the predictability of hornet migration, we considered two dimensions: time and space. In time, we analyzed the living habits of hornets and extracted temporal features from the positive reports, so as to determine whether the hornet migration was temporally patterned through comparison analysis and tell the time predictability. In space, spatial features were extracted to determine whether the dimension is predictable. Finally, we also modeled the migration distribution of hornet colonies.

4.1 Discussion on Predictability

In this section, we discuss the patterns of hornet migration in both temporal and spatial dimensions to verify its predictability. We begin by briefly reviewing the habits of *V. mandarinia* from which we inferr some predictable migration patterns in time and space. Then, we verified the existence of such patterns over the data, thus verifying the predictability conjecture.

4.1.1 Habits and migration patterns of Asian giant hornet

Hornets are eusocial creatures with a colony life cycle of one year. Their behavior varies during different months of the year, which affects the spotting probability. There are four types of hornets, namely *queen*s, *worker*s, *male*s, and *female*s, where a fertilized female will become the new queen in the following year. There are certain patterns in the spread of hornets in both time and space.

First of all, hornets' migration temporal patterns. Referring to the literature [2, 3] on their habits, we divided a year into five phases, *i.e.*, December – March, April – May, June – August, September – October, and November, to study the behavioral changes of the nests and pests within. The life cycle of *V. mandarinia* is shown in Figure 6.

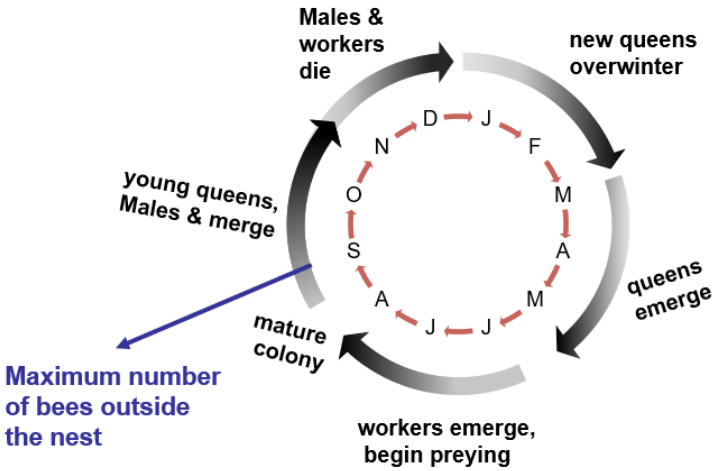


Figure 6: Life Cycle of *Vespa mandarinia*

From the Figure 6, we can see that from December to March, the only surviving individuals, queens, are in hibernation, so few hornets can be found outside the nest. From April to May, queens start to look for suitable places to build nests for themselves. They are then active on the ground, foraging for their worker larvae inside the nest. From June to August, the mature

workers start to go out for food so their queen mother no longer leaves the nest, as the colony fully matures by August. From September to October, the queen begins to reproduce males and females and there are fewer workers, at which time all the males, females and workers go outside to forage, so the number of hornets outside reaches its maximum in September — This is the very reason for the increase in *hornet attacks* in September each year. In late November, the old queen dies, as the females will mate with and possibly get fertilized by a decreasing number of males. As of December, all the hornets die, except the females who hibernate until next April, and the fertilized ones will become the queens for the next cycle.

Secondly, the spatial migration pattern. According to the literature [3], we know that in non-migratory months, workers move within about 8 km radius (average 2 km) from their nest; while in migratory months, a fertilized queen will travel to build her new nest for about 30 km.

To summarize, the patterns of hornets in time and space provide vital and theoretical basis for predicting their migration. Further, we hope to verify the existence of such patterns using given data.

4.1.2 Verification over observation data

(1) Temporal migration pattern verification

We selected the total of 14 officially verified hornet sightings from given data and tallied them by the months of submission (Fig. 7), from which we had the following observations: (a) The highest ratio of hornet sightings occurs in September, which agrees with the conclusion that the maximum number of hornets outside is reached in September; (b) The probability of hornet sightings gets higher in May and October, since a queen appears in May for her nest, and her "sons" and "daughters" are reproduced to leave the nest in October. In other words, there are *relatively* more hornets outside the nest in these months; (c) There were zero positive sightings in January-April, which is exactly the time they hibernate.

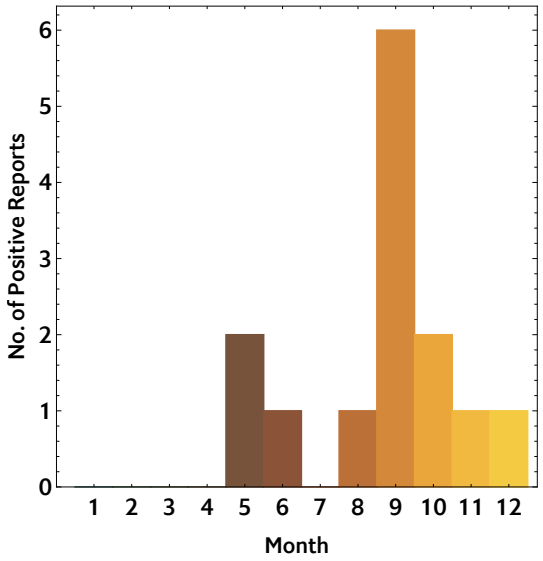


Figure 7: Positive Reports per Month

All the above observations verify the predictability of hornet migration in the dimension of time.

(2) Spatial migration pattern verification

Since the migration of hornets occurs in December-next April, we selected a life cycle of them to plot the relationship of the locations of the 8 positive sightings between April 2019 and March 2020 as Fig. 8. Based on the fact that hornets migrate for a distance of about 30 km in a year, we reason that there were a total of 3 nests initially. For a nest containing several reports, *e.g.*, the nest B in the figure, we suggested the center point to be its location. For each nest, we verified that all the positive reports around it were distributed within 8 km, which is the foraging range of worker hornets, as expected. This proves that hornet migration in space is predictable.

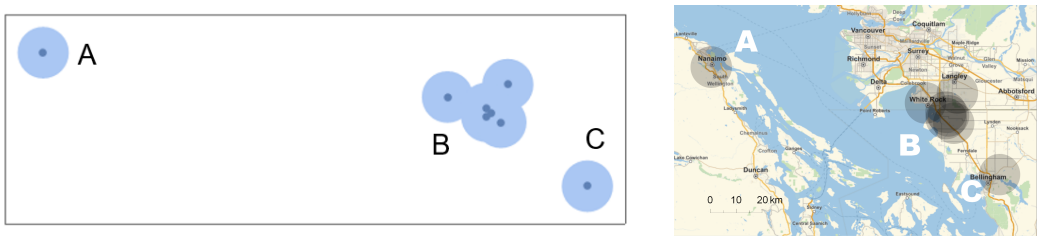


Figure 8: Locations of Positive Sightings

4.2 Model Formulation

In this section, we construct a probability prediction model with different precisions based on the migration patterns of hornets in time and space.

From the previous analysis of hornet's habits, we know that the colony migrates only in specific months. Therefore, given the initial distribution $P_{x,y}^{(0)}$ of the nest in space, we assume that the distance of the new location of nest from the old one follows a Gaussian distribution with an average value of 30 km, so from the updating pattern, defined by Equation (1), of the probability values before and after the migration, we can obtain the new probability values for each location in space thereafter:

$$P_{\text{nest}}^{(t+1)}(\lambda,\phi) = \sum_{(\lambda',\phi')} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(d-30)^2}{2\sigma^2}} P_{\text{nest}}^{(t)}(\lambda',\phi') \tag{1}$$

where (λ,ϕ) and (λ',ϕ') represent the latitude and longitude of two different coordinates in space; σ is a shape parameter for Gaussian distribution, which measures the difference of distances of the newly selected location from the old one, and thus reflects the migration precision. The larger the σ, the wider the range over which the queen builds her new nest; the smaller, the more clustered.

In non-migratory months, we can spot hornets away from near their nest, as its possible location is definite. The further away from the nest, the lower the probability of spotting hornets, and *vice versa*. Formally, according to the nest distribution as defined by Equation (1), we can obtain the probability distribution of spotting hornets in the whole space as:

$$P_{\text{observe}}^{(t)}(\lambda',\phi') = \sum_{(\lambda,\phi)} e^{-\beta_1 d} P_{\text{nest}}^{(t)}(\lambda,\phi) \tag{2}$$

where β_1 is an impact factor that determines the strength of the influence of distance on the probability of spotting a hornet.

4.3 Results

4.3.1 Parameter Estimation

(1) Estimate of Initial Distribution

Given the initial distribution, we can predict the probability of nest migration afterwards at all points in space based on Equation (1). To estimate the initial nests, we assume that the submitted reports could accurately reflect the distribution of hornets and their nests in the State, and that the distribution in 2019 is considered the initial state. We have discussed in Section 4.1.2 that, based on the relationship of hornet's habits and spatial distances, there are three observed nests within the migration cycle April 2019-March 2020, denoted by $O_i, i \in \{1, 2, 3\}$, which are assumed to be the initial distribution in Washington State. Formally, we have:

$$P_{\text{nest}}^{(0)}(\lambda, \phi) = \begin{cases} 1, & \text{if } (\lambda, \phi) \in \{O_1, O_2, O_3\} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

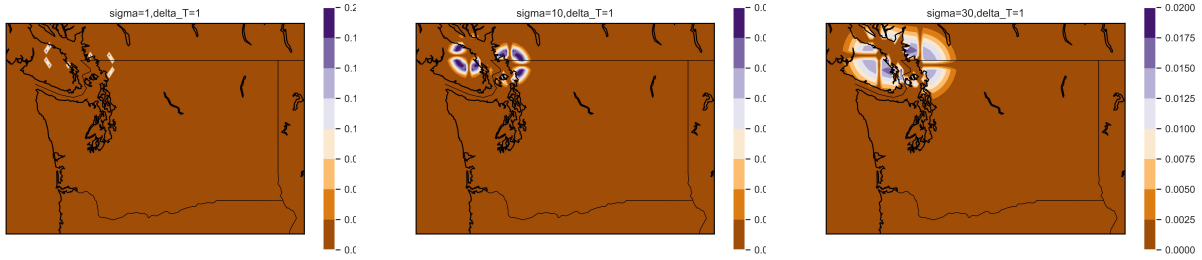
(2) Estimate of β_1

Equation (1) defines the probability of spotting hornets in the whole space given the nest distribution, where β_1 represents the decay factor that decreases the probability as the distance increases. Based on the literature [2], we know that the estimated maximum foraging range of workers $\hat{L}_{\max} = 8$ km. When the distance approaches $\hat{L}_{\max}$, the influence factor in Equation (1) approaches 0, so it's considered that $e^{-\beta_1 \hat{L}_{\max}} \doteq 10^{-2}$, giving the following estimate:

$$\beta_1 \doteq \frac{2 \log 10}{\hat{L}_{\max}} \tag{4}$$

4.3.2 Simulation Results

To verify our migration model, we iteratively simulated the nest migration in Washington State according to Algorithm 1. We estimated the initial distribution $P_{\text{nest}}^{(0)}(\lambda, \phi)$ based on the reported data in 2019, and predicted the distribution probability in 3 years with three different accuracies $\sigma \in \{1, 10, 30\}$. The results are shown in Figure 9.



(a) $\sigma = 1$ (b) $\sigma = 10$ (c) $\sigma = 30$

Figure 9: The probability distribution of the Asian giant hornet in 2020 with different predict precisions.

Algorithm 1: The iterative simulation process of hornet migration spread and public observation prediction.

Input: Initial distribution, $P_{\text{nest}}^{(0)}$;
Prediction accuracy parameter, σ;
Total migratory iteration, T.
Output: Probability distribution after migrations, $P_{\text{nest}}^{(T)}$;
Probability distribution of spatial observations, $P_{\text{observe}}^{(T)}$.

begin
  Initialize the nest probability distribution in Washington State using $P_{\text{nest}}^{(0)}$
  for $t \leftarrow 1$ **to** T **do**
    Determine the prediction accuracy using σ
    Update the probabilities $P_{\text{nest}}^{(t)}$ of nest presence at all points in space after migration using Equation (1)
    Determine the probability distribution value $P_{\text{observe}}^{(t)}$ of hornet observation over the space for t-th iteration using Equation (2)
  end
  return $P_{nest}^{(T)}$, $P_{observe}^{(T)}$ ▷ `Final results of iterations`
end

It can be seen that as σ decreases, the spatial nest distribution predicted by our model becomes more certain and therefore shows a higher probability of local concentration; as σ increases, our prediction is looser and the distribution at a given place becomes more uncertain. This suggests that our model is able to predict migration with varying degrees of accuracy.

We took the total iterations for prediction to be 1 year, with 1 km as the accuracy, to obtain the probability distribution of hornet observations in Washington State according to Algorithm 1. The result is shown in Figure 10.

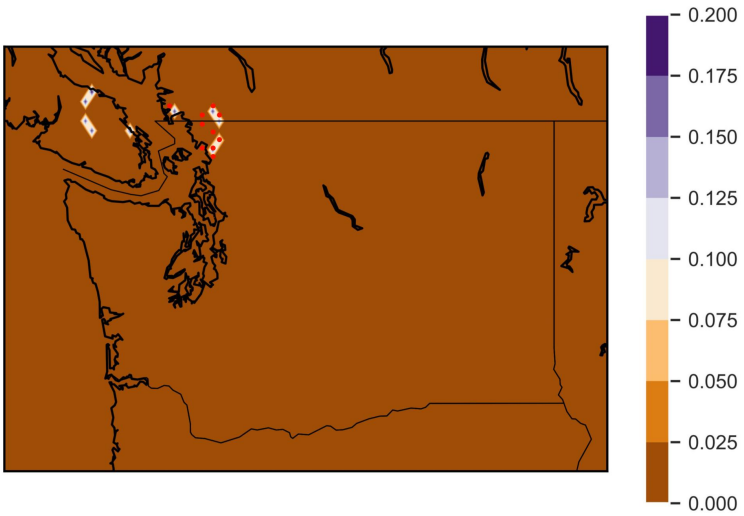


Figure 10: The observation probability distribution of hornets in Washington State

Note that the positive data (*i.e.*, the red point in Figure 10) reported in 2020 basically fall in the high-probability areas of our predictions. However, remarkably, there were no positive reports in 2020 around the Vancouver Island despite the high predicted observation probability. In view of the few reports submitted around this region, we believe it is due to the insufficient number of observation samples.

5 Model II: Misreporting Probability Prediction model

In this section, we map each of the four main features of a report, *i.e.*, the coordinates, date, text, and image, to evaluate its trustworthiness. Then, we construct a Misreporting Probability Prediction model based on *logistic regression*. Considering the arrival of new data and the need for online learning, we adopted a *gradient descent* approach to update our model. Regarding the updating frequency, we gave a hybrid scheme based on the migration patterns of hornets.

5.1 Features Mapping ϕ

Intuitively, a report convincing in all aspects is possibly not misclassified. To quantify the trustworthiness of a report x, we start with 4 critical dimensions for feature mapping ϕ to construct its feature vector $\mathbf{s}$:

$$\mathbf{s}^{(i)} = \phi\left(x^{(i)}\right) = \begin{bmatrix} s_{\mathrm{L}}^{(i)} & s_{\mathrm{T}}^{(i)} & s_{\mathrm{I}}^{(i)} & s_{\mathrm{W}}^{(i)} \end{bmatrix}^{\top} \tag{5}$$

5.1.1 Definition of s_{T} — the Time Factor

From the previous section on hornet's habits, they are more prevalent amid April-December, so the public are more likely to spot them during this time, and the report is more likely to be positive. In December-March, the only surviving individual queens hibernate underground, so the chances of spotting a hornet above the ground is near zero. The period from April-December each year is hence defined as the **active period**, denoted by $\mathbb{T}$. The time of each report was characterized as s_{T} by the following rules:

$$s_{\mathrm{T}} = \begin{cases} c_T, & T \in \mathbb{T} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where c_T is the number of positive reports in month T over the historical data.

5.1.2 Definition of s_{L} — the Location Factor

Based on the predictability of spatial migration patterns discussed in Section 4.1.2, the effect of location on the probability of spotting hornets is not negligible, making it necessary to focus on the latitude and longitude of each report. From the hornet sighting probability prediction equation (1), we define the location factor s_{L} of a report as:

$$s_{\mathrm{L}} = P_{\text{observe}}^{(t)}(\lambda, \phi) \tag{7}$$

5.1.3 Definition of s_{I} — the Image Factor

The data analysis in Section 3.2.2 has shown that, for reports with images, the official almost always choose to validate them, as the feature is also a key factor. From the perspective of

information integrity, reports that provide images are of greater priority, indicating that the witness is confident in his/her spotting the hornet, thus assuring higher accuracy. We therefore define s_{I} to represent the integrity of image information:

$$s_{\mathrm{I}} = \begin{cases} n_{\mathrm{I}}, & \text{image(s) provided} \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where n_{I} denotes the number of image(s) attached to the report.

5.1.4 Definition of s_{W} — the Word Factor

Based on the previous analysis on text length, most of the reports included textual notes. On one hand, the longer the note, the more credible the witness; on the other hand, if the note involves more features closer to *V. mandarinia* than other species, the report is less likely to be mistaken. Therefore, we constructed a hornet feature dictionary (Table 2) from a glossary extracted from attached data, related *Wikipedia* webpages, and literatures. For each report, we are then able to calculate the word frequencies of all the words it contains with respect to this dictionary and obtain its word frequency feature.

Table 2: The dictionary of key characteristics of *V. mandarinia*

Features dictionary	Asian giant hornet	Other confusing hornets
Nest Location	"Underground","forests", "burrows","roots","trunks", $\cdots$	"Time limbs", "house eaves", "exposed", "lawns", $\cdots$
Body Appearance	"Yellow heads", "black thorax", "striped abdomens", "giant", $\cdots$	"Small", "black and white colored", $\cdots$

Combining the two parts, we define s_{W} as the word factor to evaluate a note:

$$s_{\mathrm{W}} = \frac{1}{n_{\mathrm{W}}} \sum_{i=1}^{n_{\mathrm{W}}} (q_i - k_{\mathrm{i}}) + \beta_2 \log\left(n_{\mathrm{W}} + 1\right) \tag{9}$$

where n_{W} is the number of words in the report note, $q_i(k_i)$ is the word frequency of the i-th word in the note *w.r.t.* our (non-)*V.-mandarinia* feature dictionary, and β_2 is a parameter about the weight that balances the two aspects of the text.

5.2 Report Misclassification Probability Regression

In previous sections, we obtained the rating vector of the report sample after feature mapping, which quantifies the trustworthiness of a report. On this basis, we use a *logistic regression* model to predict its misclassification probability.

Logistic regression is a generalized linear model. We define the categorical variables y as:

$$y = \begin{cases} 1, & \text{for mistaken classification} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

Then, we have the misclassifying probability $P_{\text{mistaken}}^{(i)}$.

$$P_{\text{mistaken}}^{(i)} = \mathbb{P}\left(y = 1 \mid x^{(i)}\right) = \frac{1}{1 + e^{-\theta^T \phi\left(x^{(i)}\right)}} \tag{11}$$

where $x^{(i)}$ is the interested report, $\phi(\cdot)$ is the feature mapping for reports (yielding $\mathbf{s}^{(i)}$), and θ is the parameter of the misclassification probability model.

The loss of a sample prediction is usually represented by a *binary cross-entropy function* $H(\cdot)$. However, given the extreme imbalance of reports in two types, we modified the binary cross-entropy function, where, for the correctly classified fraction, *viz.* when $y = 0$, if our model predicts it as misclassification, we scale up its loss by a degree of τ to ensure that our model would not predict all samples as misclassifications:

$$H(p, y) = \tau y \log(p) + (1 - y) \log(1 - p) \tag{12}$$

where τ is the weighting coefficient of the sample status, set as the reciprocal of the percentage of non-misclassified reports. On this basis, we can obtain the overall loss function $J(\cdot)$ of the given dataset $\mathbf{x}$:

$$\min J(\theta) = \frac{1}{n} \sum_{i=1}^{n} H\left(h_\theta\left(x^{(i)}\right), y^{(i)}\right) + \frac{\beta_3}{2} \|\theta\|_2^2 \tag{13}$$

where h_θ is the logistic regression function with parameter θ, and β_3 is the balancing coefficient for regular terms.

It has been proved that the objective function is convex, so in essence, this is a convex optimization problem with global minimum, and its stagnation point is our target optimal estimation parameter. The *gradient descent* method is an optimization algorithm, which can tackle the problem of finding a local minimum of a differentiable function, and thus, we can use this method to minimize $J(\theta)$.

5.3 Online Learning and Updating Frequency

When new data arrives, we use the gradient descent online learning method to adjust the parameters of our classification model to ensure that the model can be updated in time.

$$\theta' \leftarrow \theta - \eta \frac{\partial J(\theta)}{\partial \theta} \tag{14}$$

where η is the step of parameter update.

In order to balance the computing overhead and the model accuracy, we decide the update frequency according to the hornet's habits. In months when hornets are out frequently, setting the update frequency to days or weeks is helpful to capture the latest status, while in other months under hibernation or when fewer hornets forage, we believe it is sufficient to set the frequency to a monthly level.

Specifically, based on hornet's habits, September is the month with the most hornets outside their nest, so it is the easiest month for the public to spot hornets, thus the official receiving the most reports. Therefore, the update frequency for September is set to *day*s. April-August and October-December are active months, but not as active as September, so the frequency is set to *week*s. December-March is the hibernation period for hornets, so there will be a minimum number of sighting reports received, during which the frequency is set to *month*s.

5.4 Model Evaluation

Given much fewer positive samples than negative ones in the data, we selected merely a small number of positive and negative cases as the validation dataset of our model, where all the 14 positive reports were selected, and 14 negative ones were then selected with uniform randomness in space and time from 2019 to 2020. A 3-fold cross-validation was next performed on the data of the two statuses, where the cross-entropy weighting coefficient τ was set to 1 to obtain the following fitting metric for our Misreporting Prediction model.

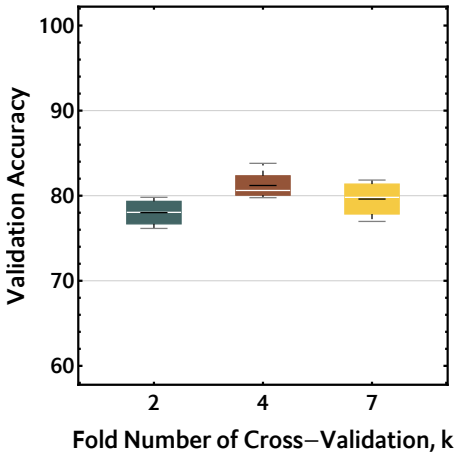


Figure 11: The accuracy of k-fold cross-validation on a small number of samples with balanced statuses.

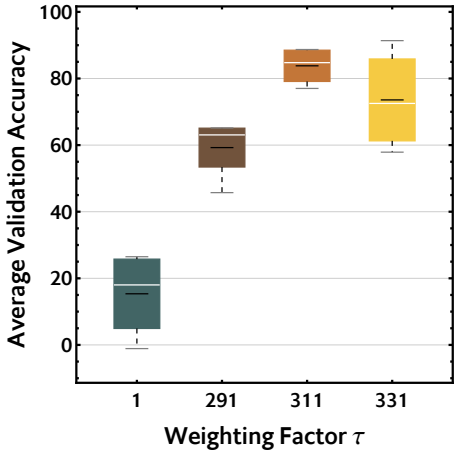


Figure 12: Under different scaling parameter τ settings, the cross-validation accuracies on report samples of imbalanced statuses.

Our model fits reasonably well despite only few negative samples it makes use of, which indicates the reliability of the features we extracted as described in Section 5.1.

Furthermore, to verify the effect of the status weighting coefficient in Equation (12), we collected all the positive and negative reports from 2019 to 2020, regarded τ as a model parameter, and obtain the training and test sets by 5-fold split, meanwhile ensuring the inclusion of samples of both statuses in each set. We performed grid search for $\tau \in \{\tau_0 - \Delta\tau, \tau_0, \tau_0 + \Delta\tau\}$, with a 5-fold cross-validation for each parameter (still ensuring both inclusions). Since the reports were mostly negative, the naïve use of accuracy would not evaluate well the prediction of our model, so we took its average prediction accuracy on positive reports as the metric.

After searching for an optimal parameter setting, a prediction accuracy of 75% is achieved on only few samples. Considering the prediction difficulty due to the extreme imbalance of the samples (with positive/negative $\approx 1/400$), the model performs well enough. Also, comparing results with and without setting the weighting parameter ($\tau = 1$ for the original binary cross-entropy function), we find that the weighting correction parameter contributes much to improved performance.

6 Model III: Report Priority Evaluation Model

In order to interpret the reports submitted by the public, we present the Report Priority Evaluation Model in this section. Based on the Misreporting Probability Prediction model in Section

5, we further corrects the probability with consideration of the inter-corroboration among reports, thus determining the priority ranking of given reports. Finally, based on *spatially isometric sampling* and *hypothesis testing*, we constructed evidence for complete pest eradication in Washington State.

6.1 Priority Ranking of Reports' Importance

To maximize the use of limited government resources and increase productivity, we need to rank the order of submitted reports to guide the investigators. Intuitively, the more likely that a report is positive, the higher its priority, *i.e.*, the sooner it gets further investigated.

We define P_i as the that the i-th report is positive. From the previous model for predicting the likelihood of report misclassification, we can compute the probability $P^{(i)}_{\text{mistaken}}$ that the i-th report was misclassified (*i.e.*, false positive). Then $P_i = 1 - P^{(i)}_{\text{mistaken}}$.

However, P_i considers only one report and does not consider its inter-corroboration with reports that are similar to it in space and time. If the reports are in the same migration cycle, within which the hornets do not migrate, we are interested in whether the reports lead to the same nest; if different, we consider whether the nests that the reports lead to are related. For this, we defined the mutual influence factor $F_{i,j}$ between reports x_i, x_j as:

$$F_{i,j} = \begin{cases} e^{-\lambda d_{ij}}, & \text{if } t_i = t_j \\ \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}(d_{ij} - 30\Delta t_{i,j})^2}, & \text{otherwise} \end{cases} \tag{15}$$

where $\Delta t_{i,j}$ denotes the difference between the migration cycles of i-th and j-th report. As in Section 4.1.1, a migration cycle spans from April to next March.

Based on $F_{i,j}$, we construct the priority evaluation Z of a report:

$$Z_i = \sum_{j=1}^{n} F_{i,j} P_j \tag{16}$$

where n denotes the total number of reports involved in ranking. According to Equations (15) and (5.2), the priority Z_i of each report can be obtained. The greater the Z-value of a report, the more likely it is to be positive, and thus the sooner it should be investigated.

6.2 Evidence of Pest Eradication

6.2.1 Hypothesis Testing of Hornet's Absence

Over a given time period, it is assumed that we are able to collect reports that are sufficiently dense in time and space in Washington State. To estimate the probability P of spotting hornets in Washington State, we obtained the typical sample set D based on a spatio-temporal isometric sampling method with intervals of 1 day and 4 km^2 as:

$$D_n = \{x_1, x_2, \cdots, x_n\} \tag{17}$$

Based on the Misclassification Prediction model constructed in Section 5, the probability p_i of sighting in each report i can be estimated as:

$$p_i \doteq p_{\text{positive}}\left(x_{\text{i}}\right), i \in \{1, 2, \cdots, n\} \tag{18}$$

and we then obtain the probability sample set as:

$$D'_n = \{P_1, P_2, \cdots, P_{\mathrm{n}}\} \tag{19}$$

Define the probability of spotting hornets in Washington State as P, which indicates their overall distribution. We take the mean value $\bar{P}$ of the samples as an estimate of P. We argue that, if the hornets in the state have been eradicated over some period of time, most of the reports submitted should be misclassified, *i.e.*, the positive rate of these samples is expected to approach zero, $\mathbb{E}[P] \to 0$. Based on the idea, we used hypothesis testing in this section to construct our evidence.

Let the mean value of P be μ and its variance be σ^2 . According to the Law of Large Numbers and the Central Limit Theorem [1], the following statistic approximately obeys the Gaussian distribution:

$$\frac{\bar{P} - \mu}{\sigma/\sqrt{n}} \dot{\sim} \mathcal{N}(0, 1) \tag{20}$$

This establishes the following hypothesis test: $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. Since σ^2 is unknown, the corrected sample variance $S_n^* = \frac{1}{n-1}\sum_{i=1}^{n}(p_i - P)^2$ is used instead. Then, the following t-test statistic is obtained:

$$T = \frac{\bar{P} - \mu}{S_n^*/\sqrt{n}} \dot{\sim} t(n-1) \tag{21}$$

Taking the confidence level $\alpha = 0.01$, we accept H_0 if Equation (22) holds, otherwise we reject it. Accepting H_0 indicates that we are 99%-sure that no hornets appear in that period of time.

$$\left| \frac{\overline{P} - 0}{S_n^*/\sqrt{n}} \right| \leqslant t_{1-\alpha}(n-1) \tag{22}$$

6.2.2 Evidence Formulation

Based on our hypothesis testing solution in Section 6.2.1, we can determine the existence of hornets in Washington State within any given time period. Particularly, we can tell their presence in a particular year. Considering the contingency and the special case of an undetected queen, detection for one year is insufficient to justify their complete eradication. The probability of the event where a hornet nest is actually present but not detected in a one-year sample is assumed to be $p_i(t)$. Then, the probability of such an event occurring for n consecutive years under continuous observation is $P = \prod_{i=1}^{n} p_i(t)$. It is clear that $P \to 0$ as $n \to \infty$ since $p_i < 1$. Therefore, to ascertain eradication, multi-year observation would be more convincing. n is thus set to 3 following WSDA's advice [4].

6.2.3 Detecting Frequency in Eradication Monitoring

Section 6.2.2 discusses that we should observe for at least 3 years to ensure pest eradication. During this time, if there's a need to have real-time information of the eradication, we can follow such a detection frequency scheme in different months. Let the time-span of the updated samples be ΔT. For April-December when hornets are active, we choose $\Delta T = 1$ week; for January-March when they are mostly in hibernation, we choose $\Delta T = 1$ month.

6.3 Model Evaluation

Generally speaking, the higher a reporting priority model places the positive reports, the more credible it should be. Therefore, to test the validity of our model, we took the priority metric based on Model II as a baseline, and compared it with the metric of Model III to explore how they place positive reports in the observation set.

Following Section 5.4, we can construct a small sample set with balanced statuses, including 28 reports, of which half are positive. Specifically, the dataset was split as 1:1, with 14 fitting and 14 validating pieces. Then, the classification model was fitted to obtain the priority ranking of positive samples on the validating set.

The results show that our method, on average, places a positive sample at a higher priority (Figure 13). This shows its validity, which, accounting for spatial inter-corroboration, better reflects the how reports relate to each other, rather than focusing on one report alone, thus more accurately determining the priority of a report.

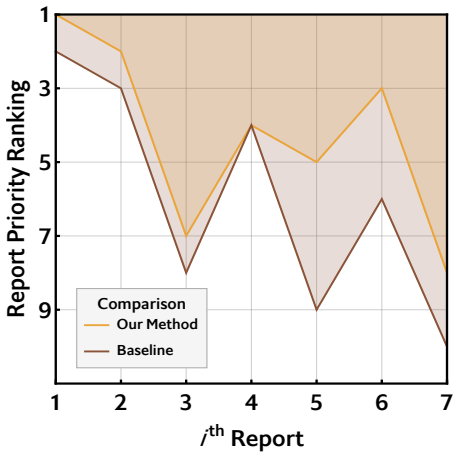


Figure 13: The priority rankings of the 7 positive samples in the validation set under different ranking models.

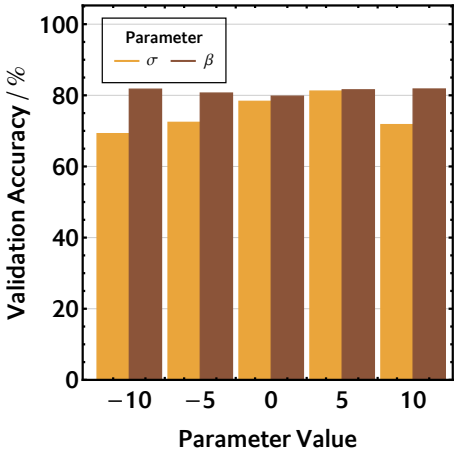


Figure 14: Sensitivity analysis results of two main parameters, σ and β_1.

7 Sensitivity Analysis

We performed sensitivity analysis on two important parameters, σ(Equation 1) and β_1(Equation 2) with the samesplitting method of dataset mentioned in the Section 5.4. First, we varied σ from -10% to $+10\%$ in steps of 5%. The results suggest that the validation accuracy of our Classification Model generally shows a trend of increasing and then decreasing. This is reasonable since σ affects the accuracy of the estimated nest distance before and after migration in Model I, which in turn determines the location feature. Migration predictions that are either too broad or too certain will fail to make the report an accurate reflection of hornet sighting, resulting in a location factor s_{L} introduced in Section 5.1.2. Therefore, σ is a very important parameter that needs to be decided carefully. Then, similarly, we varied the value of β_1 between -10% and $+10\%$ and evaluated its sensitivity. As shown in the Figure 14, the results show that this parameter is insensitive. To summarize, one of our two important parameters, σ, was tested to be more sensitive. Overall, our model has a moderate level of robustness.

8 Strengths & Improvements

8.1 Strengths

- To tackle the imbalance of samples, a weighted correction is performed. The verification accuracy rate on the historical positive data reaches 75%, which indicates the credibility of our model. Further considering such extreme imbalance, the system remains simple and does not involve complicated image analysis.
- The sensitivity analysis of the model demonstrates the validity under the parameter β_1 and the robustness of the model.
- To assess the complete eradication, we use the CLT and Student's t-test instead of u-test, making our model more reasonable.

8.2 Improvements

- When performing feature extraction from the reports, we consider only the 4 feature values in them. We can also mine further into the information given.
- Since the data is incomplete, there will be errors inevitably when estimating parameters. If more data is provided, our model will be further improved.

References

[1] S. G. KWAK AND J. H. KIM, *Central limit theorem: the cornerstone of modern statistics*, Korean Journal of Anesthesiology, 70 (2017), p. 144.

[2] M. MATSUURA AND S. F. SAKAGAMI, *A bionomic sketch of the giant hornet, Vespa mandarinia, a serious pest for Japanese apiculture (With 12 Text-figures and 5 Tables)*, Summary of the Faculty of Science, Hokkaido University, 19 (1973), pp. 125–162.

[3] T. STANKUS, *Reviews of Science for Science Librarians: "Murder Hornets:" Vespa Mandarinia Japonica*, Science & Technology Libraries, 39 (2020), pp. 244–252.

[4] WASHINGTON STATE DEPARTMENT OF AGRICULTURE, *Asian giant hornet: Frequently Asked Questions*. https://agr.wa.gov/hornets. Retrieved online on February 6, 2021.

Appendices

Appendix A Tools and software

This paper is written and generated via LaTeX, free distribution.

Graphs and calculation are generated using the combination of *Python* 3, WOLFRAM *Mathematica*, and MATLAB.

Appendix B The Codes

Here are the programs we used for our model.

```
import pandas as pd
data = pd.read_csv('2021MCMProblemC_DataSet.csv',encoding = 'gb18030')
data_clean = data[data['Lab Status']!='Unprocessed']
data_clean['Notes'].apply(lambda x: len(x.split(' ')))
def is_valid_date(strdate):
    if '-' in strdate:
        return True
    return False
data_clean = data_clean[data_clean['Detection Date'].apply(lambda x:is_valid_date(x)
    )]
data_clean['Detection Date'] = pd.to_datetime(data_clean['Detection Date'])
data_clean['month'] = data_clean['Detection Date'].dt.month
data_clean['year'] = data_clean['Detection Date'].dt.year
data_clean = data_clean[data_clean['year'] >=2019]
def trans(x):
    if (x in set(image_map['GlobalID'])):
        return 'yes'
    else:
        return 'no'
data_clean['submit_image'] = data_clean['GlobalID'].apply(lambda x: trans(x))
image_map = pd.read_csv('image_map.csv',encoding = 'gb18030')
P = np.zeros((len(loc_lon),len(loc_lat)))
lat = df2019[data_clean['Lab Status']=='Positive ID']['Latitude'].tolist()
lon = df2019[data_clean['Lab Status']=='Positive ID']['Longitude'].tolist()
for i in range(len(lat)):
    P[map2grid(lon[i],lat[i])] = 1
sigma=10
mig=30
imin,jmin = map2grid(llon,llat)
imax,jmax = map2grid(ulon,ulat)
from tqdm import tqdm
P1 = np.zeros((len(loc_lon),len(loc_lat)))
for i in tqdm(range(len(loc_lon))):
    if i >= imin and i<= imax:
        for j in range(len(loc_lat)):
            if j >= jmin and j<= jmax and globe.is_land(grid2map(i,j)[0],grid2map(i,
                j)[1]):
                s= 0
                for i1 in range(len(loc_lon)):
                    if i1 >= imin and i1<= imax and i1!=i and j1 != j :
                            for j1 in range(len(loc_lat)):
                                if  j1 >= jmin and j1<= jmax and i1!=i and j1 != j
                                    and P[i1,j1]!=0 and globe.is_land(grid2map(i1,j1
                                    )[0],grid2map(i1,j1)[1]):
                                        d=geodesic(grid2map(i,j),grid2map(i1,j1)).km
                                        s += 1/(sigma*np.sqrt(2*3.14))*np.exp(-(d-
                                            mig)**2/(2*sigma**2))*P[i1,j1]
                P1[i,j] = s
```

APPENDIX C

MCM #2102194 Memorandum on Hornet Sighting Reports

M. C. M.

#2102194

Registration No.

Mandarinia Colony Master

Your reliable advisor on potential *Vespa mandarinia* colonies based on available information.

MEMORANDUM

Date: February 8, 2021

To: Washington State Department of Agriculture

From: M.C.M. #2102194

Re: Confirming the Buzz about Hornets

1. INTRODUCTION

The purpose of this memorandum is to present the latest analysis and related advice of MCM Advisory Team on recent pest crisis as per the requests from the Washington State Department of Agriculture (WSDA). The recent spread of Asian giant hornets (*Vespa mandarinia*) as ecological invaders in America and Canada has posed hazardous threats to local apiculture and provoked panic among the public, from whom increasingly extensive reports of their possible presence had been submitted via Department's hotlines and websites. The information pools require prioritization and interpretation.

2. RESULTS

The predictability of Asian giant hornet colony migration is verified temporally and spatially, under which basis a mathematical model for predicting the migratory distribution and observational probability is proposed. The model adopts a specifiable precision scheme.

Figure 1 shows the result of distribution from the proposed model with specified precision of 1 km.

Pursuant to the attached dataset of lab-verified hornet sighting reports from the public, another model of the likelihood of its mistaken classification is proposed. The model notably takes into account the extreme imbalance in positive and negative samples given, and supports update upon the arrival of new reports, while assuring the overhead balance. The prediction accuracy on given positive samples reaches about 75%.

The limited resources further require yet another model for prioritizing the reports submitted, which is proposed for State's investigators based on the interrelations and the forementioned model with the goal of optimal allocation of resources.

Evidence for hornet eradication in Washington State is hereby derived, being no detection reported of hornet under certain statistical levels for at least three consecutive years.

3. TECHNICAL DETAILS

The predictability results from the life cycle of hornet colony and public reports. Model 1 is further based on a method of Gaussian iterative simulation.

Model 2 adopts logistic regression and feature extraction, involving text processing. The imbalance is dissolved by fitting the parameters using a binary cross-entropy loss function with weight correction and k-fold cross-validation, whereas the update is accomplished by a gradient descent method and variable frequencies, establishing an online learning scheme.

Model 3 proposes an influence factor between reports. The evidence is obtained by t-test statistics and isometric sampling.

Due to space limitation, please refer to the attached research paper.

4. ACKNOWLEDGEMENT

MCM is greatly honored to join efforts in addressing the pest crisis in Washington State.