
Analyzing the Spread of the Asian Giant Hornet

Summary

Species invasion are always concnred by the people, because it could bring so many severe problems such as breaking the ecological balance and demaging the biodiversity. Here, we discuss a certain insect called Asian Giant Hornet (hornet below), discovered on Canada and propose a model to deal with the relavant problems. To reveal the severity of the hornet invasion, we constructed a model that simulates the spread of the hornet in Washington State without human intervention. Secondly, aiming at the analysis problem of massive reports, we build a report reliability evaluation system to evaluate the reliability of a report.

In Task 1, we first examine the hornet spread over time, we use ADF method to test the data series stationarity, and found that it is not a smooth time series, at the same time, we calculated the Spearman correlation coefficient of propagation index in time, confirm the spread of hornet and time has a moderate positive correlation. In addition, we divided the target area into small rectangles and used cellular automata to simulate the spread of the hornet, which proved that if nothing was done, the hornet would spread throughout the central and western state of Washington.

In Task 2, we set up a report reliability evaluation system to measure the probability of a report being a positive eyewitness report using a reliability index. We used the analytic hierarchy process (AHP) to select four important indicators through manual selection, namely photo, Dection time, Location and Notes. And their weights are determined by using the positive reciprocal matrix. For photos, CNN is used to evaluate the indicators of images; for Detection Time, statistics and quartic polynomial fitting are used to evaluate the indicators; for Location and Note, the idea of K-nearest neighbor is used to complete the evaluation. And finally, by a weighted sum, To calculate the reliability, which in turn reflects the probability of the report being misjudged.

In Task 3 to Task 4, we used our report reliability evaluation system to analyze the existing reports, and through screening and mapping, we found the areas where reports with high reliability were enriched, which were the areas for our limited investigation. Then, we combined computation cost and updating frequency, and discussed a model updating strategy that could continuously improve model accuracy over time without bringing too much computation burden.

In Task 5, we preliminarily discuss the conditions for judging whether an area is free from the threat of the hornet, based on the reported frequency of occurrence and the breeding cycle of the hornet.

key words : Asian giant hornet; Cellular Automata; ADF; AHF; CNN;

MEMO

Dear Officials:

Recently, we discovered that a species called the Asian giant hornet has begun to invade Washington state.

To further analyze the severity of the issue, we modeled its spread. By 2020, the number of reports of suspected Asian giant hornet sightings has increased each year.

Our model suggests that if nothing is done about the Asian giant hornet invasion, the impact could be so dramatic after five years that the ecological and economic losses could be incalculable. We recommend tracking the insect's movements as soon as possible and taking prompt action to limit the damage. It is impossible to carry out a comprehensive investigation of this pest with limited resources and manpower.

Relying on people to report suspected targets is a good means of information collection, but because ordinary people do not have the knowledge and experience to collect information on such insects, most of the information reported is chaotic, and it is difficult to conduct accurate analysis. In addition, it is difficult to process such a huge amount of information with limited manpower. In this regard, we have established a report reliability assessment system to help filter valuable reporting information, to help you quickly determine whether a sighting report is valid or not.

So far, we have analyzed and evaluated the sightings provided, and made a brief analysis of the temporal and spatial distribution of the reported locations. According to the analysis, hornet sightings are more likely to occur in Bellingham, which can be used as a reference when dispatching commissioners to investigate.

Due to the limited amount of data available, the results of our analysis are incomplete. However, we have taken this into account, and as soon as we have new data, we can update the model further to achieve better results. Our model will give the probability index of whether the report is valid or not, and then judge whether to update the model. If the report can be judged by professionals and then input into the model, the

accuracy can be improved. Given the importance of computational cost and model accuracy, we recommend that departments adopt an update strategy: if there are valid reports, update them immediately; otherwise, update the models with new reports at the end of each month.

The following methods can be used to determine whether the Asian giant hornet has been eliminated:

We consider the month to be safe if reports received within a year from a certain month have reliability predictions below 0.571. Of course, this could be influenced by chance, but the analysis of the reports so far does not indicate that there is such a chance for six months. If it is safe for six months in a row, then we can rest assured.

Based on our analysis, we offer some suggestions for controlling the Asian giant hornet:

1. Small survey teams can be sent to investigate areas of high Asian giant hornet density to obtain more detailed and reliable data.
2. On the website receiving the report, it is advocated to use scientific and accurate words for the characteristics of observed insects and briefly introduce the scientific methods of photographing insects. Improving the relevant literacy of the people will help to improve the effectiveness of the report.
3. Timely update the model with time changes to improve the accuracy of the model and track the trend of Asian giant hornet more accurately and timely.
4. A combination of chemical and biological control may be considered. Spraying interferon or other chemicals targeted at the Asian giant hornet in dense nesting areas reduces its population density, and introducing its natural enemies slows its growth.

We believe this model has implications for controlling the effects of the Asian giant hornet. Please contact us for further cooperation.

Sincerely yours

FROM: Team #2111874

Date: Feb. 6, 2021

Contents

1	Introduction: Assumptions	1
2	Data Processing	1
2.1	Repeated Data Verification	1
2.2	Handle Null or Abnormal Values	2
2.3	Remove Useless Data	3
3	Prediction	3
3.1	Examination	3
3.1.1	Stability Test	4
3.1.2	Correlation	5
3.2	Prediction Method	5
3.2.1	Area Division	5
3.2.2	Cellular Automata	6
3.3	Result	7
4	Model Constrction	8
4.1	Overview of Our Model	8
4.2	Subquestions with Submodels	8
4.2.1	Detection Time	8
4.2.2	Notes	11
4.2.3	Location	13
4.2.4	Photographs	13
4.3	the Determination of Weights	13
5	Model Simulation and Problem Solving	14
5.1	Task 1	14
5.2	Task 2	14
5.3	Task 3	15
5.4	Task 4	16
5.5	Task 5	17

6	Model Analysis	18
6.1	Model Sensitivity Test	18
6.2	Strengths	18
6.3	Weaknesses	19
	References	20

1 Introduction: Assumptions

Due to the lack of necessary data, we make the following assumptions to help us perform modeling.

1. Asian hornet begin to appear as early as March 1 (the beginning of spring in the northern hemisphere), and all hornets except the queen hornet die as late as December 1 (the beginning of winter in the northern hemisphere). According to the PDF of the question, the hornet is born every year. In the spring, the queen wakes up and begins to look for a place to build a nest. The population grows slowly and peaks in August.
2. In September, the queen hornet begins to lay eggs, and about 0.35 females are fertilized and then set out to build new nests next year. According to the accompanying PDF, the fertilized females survived and established their hives.
3. The species is not under very strong control. Within the range predicted by the model, the habitat of Asian hornets was not significantly affected. At the same time, as an invasive species, the Asian hornet is free to breed without having to deal with its natural predators.
4. The new nest distance of the newly fertilized female wasp obeys the Poisson distribution.
5. Because the water area is so small, the area containing water is still considered land.

2 Data Processing

2.1 Repeated Data Verification

According to the PDF description of the title, each global-id is unique and does not have the phenomenon of repetition. The correctness of this

information is verified through programming, so there is no need to screen out the duplicate lines.

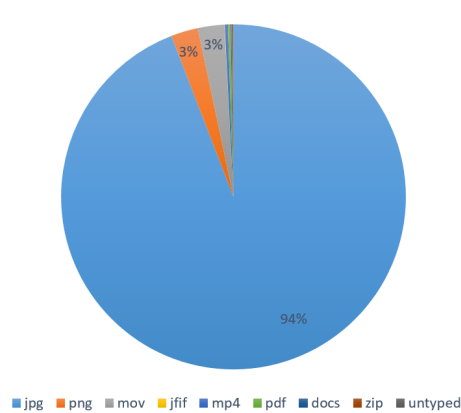


Figure 1: Files

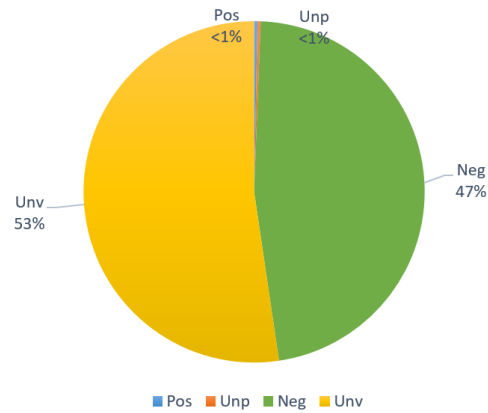


Figure 2: Lab Status

2.2 Handle Null or Abnormal Values

In the 2021mcmproblemc_dataset.xlsx file, you can easily find and fill in blank values using the Excel search function. Three NULL values can be found after sorting from smallest to largest by Detection Date. They should be excluded due to lack of time information.

There are also cases where Notes are empty in the table, with about 500 Notes rows out of 4,000 rows of data. This is because people only upload pictures when they upload information. These items can be further judged through image processing, so they are retained.

After sorting Detection Date, it can be found that the time format of some rows is MM/DD/YY, which is different from the majority of rows in the table. In addition, the time of these rows is before 1900, and they are outliers. Therefore, these rows are excluded in the screening process.

The latitude and longitude provided in the table were detected and found to be within a reasonable range. The latitude and longitude in the map drawing table intuitively showed that the points were all within the range of Washington State.

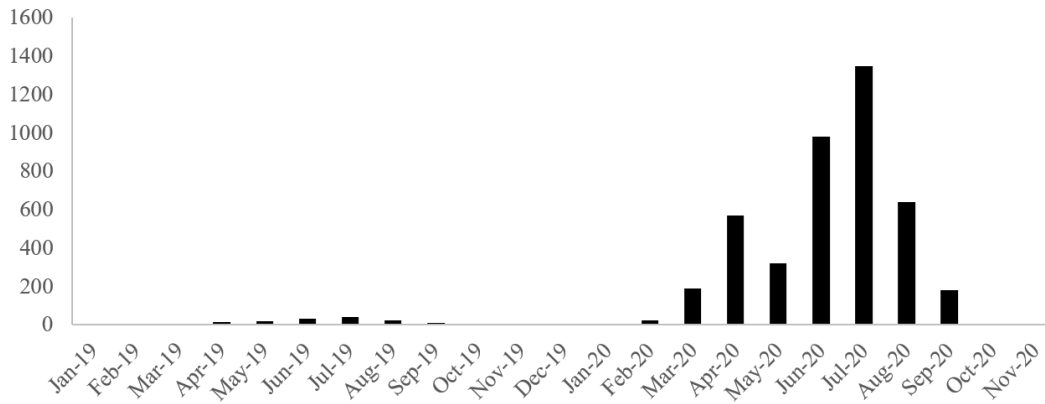


Figure 3: Number of Reports

2.3 Remove Useless Data

By observing the sorted years, it can be found that most of the spots were found between 2019 and 2020, of which the latter accounted for the vast majority. There are only about 50 data before 2019, and these data are not of great analytical value because the years are too scattered, so they are excluded from these years.

According to statistics, among all the 3305 picture documents, there are 3111 JPG files, accounting for the vast majority of all the documents. The other files have various formats, which are not easy to be processed uniformly, and some of them also include text information, so they are directly removed.

Add a new column to 2021mcmproblemc_dataset.xlsx. If a JPG file exists, it will be the file name. If no JPG file exists, the column will be blank.

3 Prediction

3.1 Examination

As the start point for further research, the spread of hornets should be related to time. The first problem here is how to use data to represent the spread of hornets. We use time series analysis to examine this time series is a weak stationary process and this value is relevant to time.

The amount of hornet sighting is adopted to illustrate the increasement of hornets amount, so the amount variation of hornets by time is discussed. However, since the amount of positive samples are 0 is common in each month, it's not suitable to discuss this variation in month. On the other hand, it's still unreachable to see the variation in years since the data only mainly include two years. Under the comprehensive consideration of the Seasonal habits of hornets as the profile said, using quarter (three months) as the time window to have an effective statistics (Figure 8). So divide 2019.01-2020.12 into 8 quarters.

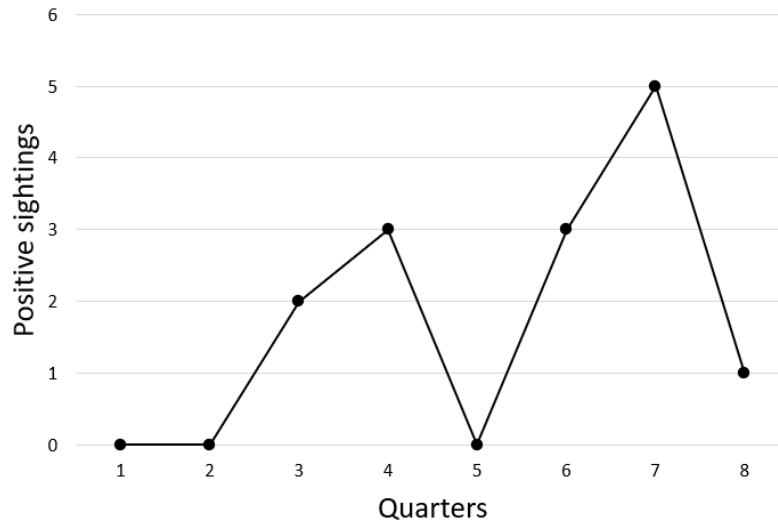


Figure 4: Number of Positive Sighting by Quarters

3.1.1 Stability Test

First, we examine stability of this time series, using augmented Dickey–Fuller test (ADF).[1]

Hypothesis: $H_0 : |\rho| \geq 1, H_1 : |\rho| < 1$; H_0 represents the series is stationary and H_1 represents the series is not stationary.

Test statistic:

$$t = \frac{\hat{\rho} - 1}{\hat{\sigma}_p}$$

After the calculation of test statistic t , and designating the confidence level, the probability of t in the confidence interval (CI) could get by

seeking existed table. Another method is calculate the probability of test statistic t not in the CI, $p - value$. If this value is far less than confidence level C , it is considered H_0 is true with sufficient confidence, or it is considered H_1 is true.

Let confidence level C be 0.05, after calculated the $p - value = 0.846$. Since $p - value$ is far greater than confidence level, the series is related to its own history. In other words, the data of each moment is related to the historical data.

3.1.2 Correlation

The time is regarded as the independent variable and the number of sightings as the dependent variable, and the correlation is directly calculated.[2]

After calculation, $p - value = 0.167$, far greater than 0.05 of confidence level. And the correlation level $C = 0.540$, shows a moderate positive correlation between the two.

3.2 Prediction Method

3.2.1 Area Division

Cellular automata (CA) is mainly used to predict the spread of hornet. To use this method, the map is needed to be divided up to Cellulars before further process. We divide the area in this way:

- The rectangular area of 45° N to 50° N and 120° W to 130° W is taken as the research object.
- Latitude and longitude are drawn every 0.2 degree, dividing the graph into small rectangles.

According to Haversine equation of spherical distance

$$haversin\left(\frac{d}{R}\right) = haversin(|\varphi_2 - \varphi_1|) + \cos(\varphi_1)\cos(\varphi_2)haversin(\Delta\lambda) \quad (1)$$

$$haversin(\theta) = \sin^2\left(\frac{\theta}{2}\right) = \frac{1 - \cos\theta}{2}$$

Here, R is the radius of earth, φ_1 and φ_2 is the latitude of two points and $\Delta\lambda$ is the absolute value of the difference between the longitude of two points.

In this way, the division ensures that the center of one rectangle is about 20~30 km away from the eight adjacent rectangles, which is the maximum distance that a fertilized female can fly.

3.2.2 Cellular Automata

Our cellular automata consists of the following elements:

- Space: Two-dimensional grids on the map.
- State: The state of each grid is defined as $s_{i,j}$, which represents the number of active nests of hornet in each grid.
- Neighbor: The neighbors of a certain cell are defined as its 8 adjacent cells.
- Evolution rules: The rules of evolution can be summarized as

$$s_{i,j}^{t+1} = f(s_{i,j}^t, s_{neighbor_{ij}}^t)$$

Assuming that the flight distance of the female hornet obeys Poisson distribution, the eight orientations of the flight process are equal-probability. [3] Because the probability of the Poisson distribution is maximized around parameter λ , $\lambda = \frac{30+0}{2}$, meaning in the best condition it fly 30 km away, and in the worst condition it only fly hundreds of meters away, which is rounded to 0 km. At the same time, the population limit of hornet within a grid is mainly dependent on latitude and can be simply assumed to be linearly negative.[4] At the same time, suppose that an hornet produces a total of *eggs* during the breeding period, and the probability of each egg successfully hatching is *alive_prob*. Here, according to the history record, let *eggs* = 20000, *alive_prob* = 0.00001 and the probability that a female hornet is fertilized is *fert_prob* = 0.35.[5]

Evolution rules:

- $0 \leq s_{i,j} \leq limit$

- Only the influence of 8 cells around one cell and its own state on the next state is considered. Since the distance between neighbor cells in the south, east, north and west is slightly different from that of neighbor cells in the southeast, northwest, northeast and southwest, their contributions are considered separately. Note that the probability of female hornets flying along the south, east, north and west directions without flying out of the current grid is p_1 , the set of neighbors in these four directions is S_1 , and the probability of female hornets flying along the southeast, southwest, northeast and northwest directions without flying out of the current grid is p_2 , the set of neighbors in these four directions is S_2 .

$$s_{i,j}^{t+1} = fert_prob \cdot eggs \cdot alive_prob \cdot \left(s_{i,j}^t \cdot (p_1 + p_2) + 0.125 \sum_{k=1}^2 (1 - p_k) \cdot \sum_{(i,j) \in S_k} s_{i,j}^t \right) \quad (2)$$

3.3 Result

We use these sightings reported in 2019 as our initial data and assume that each report site has only one active hive at first.

Global ID
{5AC8034E-5B46-4294-85F0-5B13117EBEFE}
{5EAD3364-2CA7-4A39-9A53-7F9DCF5D2041}
{124B9BFA-7F7B-4B8E-8A56-42E067F0F72E}
{F1864CC3-508C-4E60-9098-B158AB413B03}
{7F3B6DB6-2ED4-4415-8DC2-3F03EC88F353}

Table 1: Positive Reports in 2019

It can be seen from the predicted results that the initial population size and the area involved were small, but with the passage of time, the population increased rapidly and the range expanded rapidly, which is in line with the characteristics of rapid spread of invasive species. By 2024, the western part of Washington state is almost completely lost, with a

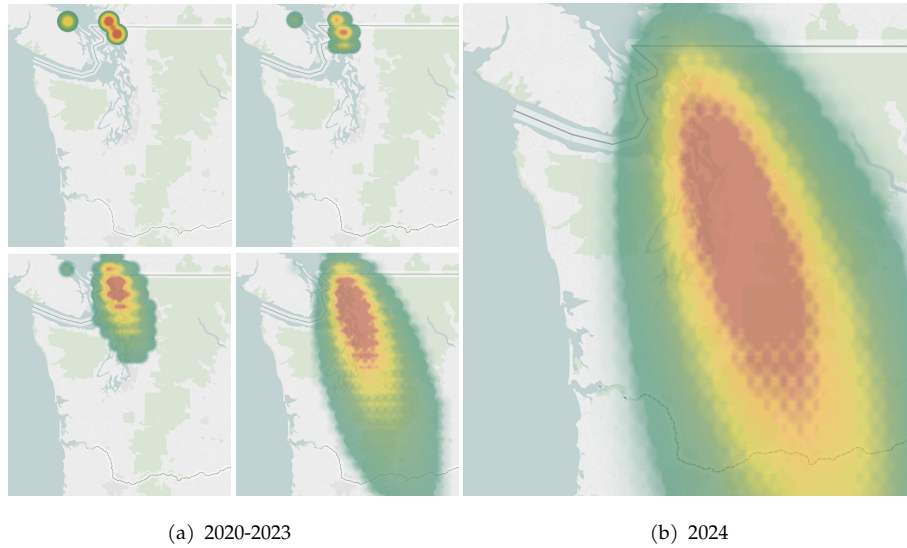


Figure 5: Prediction

trend of rapid growth in the central part. In some areas, population of Asian giant hornets could reach more than 1,000. This is because high latitude largely limits the maximum population in each grid, which makes the population size in high latitude grows more slowly than that in low latitudes, resulting in a situation that the south is fast and the north is slow.

4 Model Constrction

4.1 Overview of Our Model

We selected the most important attributes from the report as our hierarchy (Figure 6). Then The problem is divided into four subquestions.

4.2 Subquestions with Submodels

4.2.1 Detection Time

For dection time, the main consideration is the probability of correct

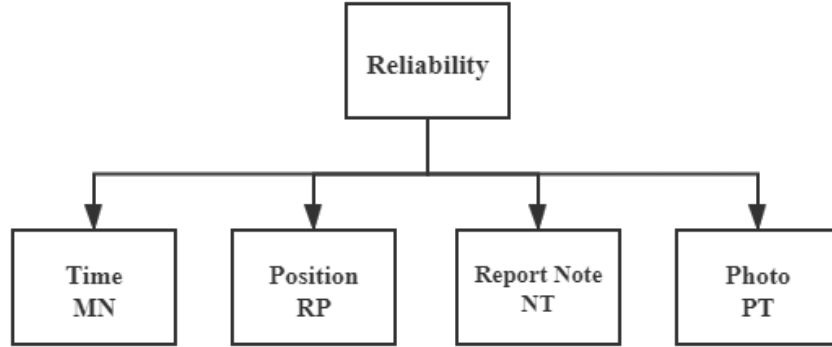


Figure 6: Hierarchy of the Model

sighting reports of each month. Here, starting from the reverse side, calculate the probability of negative sighting reports, P_w , and the probability of being positive, $P_r = 1 - P_w$.

We take negative reports for each of the last two years, with respect to the frequency distribution of the months, because there are too few correct samples. The frequency is used as an estimate of the total error probability. Notice the data in the PDF attached points out that population is at its peak in August. In addition, there is also a case of correct sightings in winter. According to the data, except for the female hornets that survive the winter, the rest are basically dead in the winter.

Considering that there will be 100% false positives in a certain month, it is not accurate to estimate only through the original data. Therefore, data smoothing is needed. In reference to the hornet life cycle, the probability of sightings is lower in winter and higher in other seasons, so there is the following data smoothing strategy:

Note that the number of wrong samples for the j th month within two years is W_j , the number of correct samples is R_j , the total sample number is T_j , and the probability of reporting an error is P_j . Let

$$max_w = max\{W_j\}$$

$$max_r = max\{R_j\}$$

$$min_w = min\{W_j | W_j > 0\}$$

$$min_r = min\{R_j | R_j > 0\}$$

$$W_j = \begin{cases} W_j + \max_w & \{j \in 1, 2, 12\} \\ W_j + \min_w & \{j \in 3, 4, 5, 6, 7, 8, 9, 10, 11\} \end{cases} \quad (3)$$

$$R_j = \begin{cases} R_j + \max_r & \{j \in 1, 2, 12\} \\ R_j + \min_r & \{j \in 3, 4, 5, 6, 7, 8, 9, 10, 11\} \end{cases} \quad (4)$$

$$T_j = W_j + R_j$$

results are as follows:

j	W_j	R_j	T_j	P_j
1	715	1	716	0.9986
2	715	1	716	0.9986
3	6	6	12	0.5
4	43	6	49	0.87755
5	129	8	137	0.94161
6	186	7	193	0.96373
7	539	6	545	0.98899
8	715	7	722	0.9903
9	354	12	366	0.96721
10	88	8	96	0.91667
11	2	7	9	0.22222
12	714	2	716	0.99721

Table 2: Positive Reports in 2019

According to the information provided, we can know that the probability of each negative report error should approximately decrease and then increase with time, which corresponds to the first reproduction of hornet population in its life cycle (As the number of individuals in the population increases, the probability of being found increases), And population extinction (the number of individuals decreases, and the probability of being found decreases). Here, we choose a polynomial function to hit the variation process, using the Least Squares Method. After testing the 1-5 degree polynomial function, we finally decided to use the quartic function to fit the curve of reporting the change of error probability over time.

$$P_w(m) = 0.0011m^4 - 0.0306m^3 + 0.2793m^2 - 0.9373m + 1.7591 \quad (5)$$

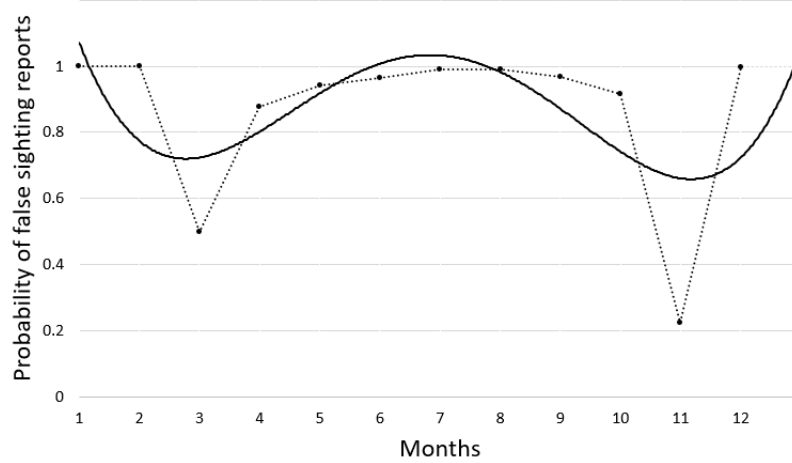


Figure 7: Frequent Words

The two minimum points of the curve are March and November, respectively, which correspond to the time when hornets are awake and when they are out looking for nests, respectively, and tend to produce the most credible reports.

To fit the curve better, an extra abscissa is set, and the corresponding ordinate is the January value. It represents the estimated value on December 31, which should be close to the estimated value on January 1. In addition, it is observed that there are parts of the curve greater than 1, which will be replaced by a probability of 0.99.

4.2.2 Notes

To deal with Note, the first thing is converting each the note sentences to a numerical fixed-length vector. To get these vector, we pay attention to the importance to each words, using TF-IDF (Term Frequency / Inverse Document Frequency) [6] to get calculate each important of words. Finally _

TF represents the frequency of occurrence of word w in sentence D_i . the $count(w)$ represents the the number of word w in sentence D_i and $|D_i|$ represents the number of words in sentence D_i .

4.2.3 Location

Still, we find the closet positive report and use the distance in sphere to calculate the distance x . Considering the America continent is big enough and the data is provided in latitude and longitude, we calculate the distance between two sites in the radian system. The process of calculation is as below.

...

Then we get the closet positive distance x , and put this value in the function below, especially when the distance reach 30 km, this value should be very low since the information says hornets are hardly to reach 30 km away to build the nests. So we use a S shape function to depict this trend.

Can we use another derivative to deal with the baseline function?

attach a photo of only this value here

4.2.4 Photographs

Image information is very important since..

a statistic of importance of information [image]

we could see image is a deterministic information to make sure if this species is hornet.

We use computer methods to evaluate the image part and get a heuristic score. It's worth to say that because the pictures uploaded are various and have a lot of informations to deal with. It's hard to evaluate by computer algorithm completely, so we introduce a human-in-the-loop method combined with CNN to get the value of image.

4.3 the Determination of Weights

Using analytic hierarchy process (AHP) to assign weights, we need to consider the impact of month, feature description, and geographic location of the report. We compare the indicators with each other to build a comparative judgment matrix to represent the relative importance of the indicators, which is represented by A below. After discussion, we think the importance is: photo, report location, month, and feature description.

$$A = \begin{bmatrix} 1 & 2 & 4 & \frac{1}{2} \\ \frac{1}{2} & 1 & 3 & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{3} & 1 & \frac{1}{5} \\ 2 & 3 & 5 & 1 \end{bmatrix}$$

tion.[7]

The larger the value of elements a_{ij} , the more important it is. a_{ij} value greater than 1 means that the attribute represented by the i-th row is more important than the attribute represented by the j-th column.

The above matrix is a positive reciprocal matrix. According to the Perron theorem, it must have a maximum eigenvalue λ_{max} , and the eigenvector \vec{X} corresponding to λ_{max} is a positive vector. There is $AX = \lambda_{max}\vec{X}$, All the components of \vec{X} are unified for normalization, and the vector is obtained as the weight vector \vec{W} . After calculation, get $\lambda_{max} = 4.05111$ and

$$\vec{W} = [0.28438 \quad 0.16990 \quad 0.07286 \quad 0.47286]$$

.

Then, the Reliability R calculation formula is obtained.

$$R = \vec{W} \cdot (RP, MN, NT, PT)^T \quad (9)$$

5 Model Simulation and Problem Solving

5.1 Task 1

Task 1 has already solved in section 2, Prediction.

5.2 Task 2

According to the report reliability evaluation system we established, we used our reliability analysis on the original Report data. It can be seen that most of the positive samples rank in the front after the reliability of

	Global ID	RP	MN	NT	PT	R	Lab Status
{5EAD3364-2CA7-4A39-9A53-7F9DCF5D2041}		1.00	0.98	1.00	0.43	0.73	Positive ID
{49D016FA-358D-4BB0-9678-4F332DCBF3C6}		0.84	0.73	0.64	0.62	0.70	Unverified
{F1864CC3-508C-4E60-9098-B158AB413B03}		1.00	1.00	0.57	0.35	0.66	Positive ID
{BEAC832C-0783-414A-9354-C297F38570AD}		1.00	0.75	1.00	0.37	0.66	Positive ID
{5AC8034E-5B46-4294-85F0-5B13117EBEFE}		1.00	0.75	1.00	0.37	0.66	Positive ID
{DEF5D82B-E326-41A5-9B6C-D46DCD86950C}		1.00	0.72	1.00	0.35	0.65	Positive ID
{A717D86F-23E9-4C8C-9F12-198A71113E93}		1.00	0.22	1.00	0.53	0.64	Positive ID
{AA461F47-1B2B-4EA1-8154-ECF70B55A334}		1.00	0.70	1.00	0.35	0.64	Positive ID
{7F3B6DB6-2ED4-4415-8DC2-3F03EC88F353}		1.00	0.73	1.00	0.34	0.64	Positive ID
{A0909E9E-FBC6-4DAA-9B25-D44C990A7B3B}		0.89	0.93	0.64	0.35	0.63	Unverified

Table 3: Top 10 Results ranked by Raliability

each Report are estimated and ranked by our evaluation system.

It can be seen from observation that a small number of positive samples are not in the list because of the absence of photos and other elements, which are within the range of explanation.

Back to our original question, to measure the possibility that a Report is misclassified, we only need to look at its reliability. If the reliability is low, the greater the probability that it is a negative Report; otherwise, if the reliability is high, the smaller the probability that it is a negative Report.

5.3 Task 3

For this problem, due to the limited manpower and material resources, it is a rational decision to prioritize the areas with high reliability of reports as the survey target. If there are multiple areas with high reliability, we choose the areas with more intensive distribution of reports as our survey target. All we need to do is find those areas where reliable reports are concentrated and prioritize them. We chose the lower quartile of the indicators of the positive sample as the threshold, and reports above this threshold were considered to have high reliability. For all reports whose reliability indicators are above the threshold regardless of their Status, we can see their distribution on the map:

Therefore, give priority to the investigation of 49.0° N 122.7° W area (figure 7). It's near Bellingham city.

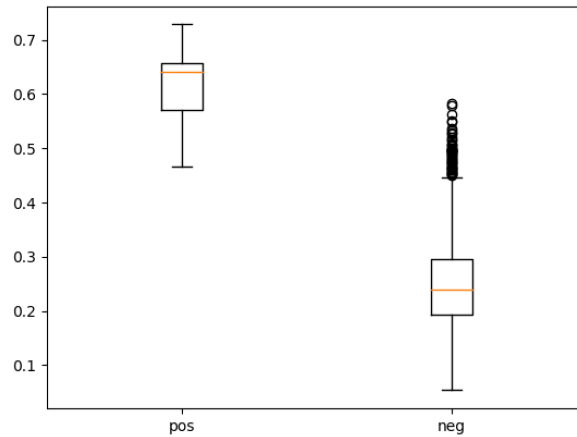


Figure 9: Reliability of Prediction

5.4 Task 4

From the model idea mentioned above, the model we built relies largely on positive report data. However, due to the absence of positive report data (only 14), the data of positive reports had a greater impact on our model than the addition of negative reports. Considering that the more data the more accurate, the model should be updated as frequently as possible. However, too frequent updating will bring unbearable calculation cost. Therefore, a balance between updating frequency and calculation cost should be achieved.

We assume that the new report we get has been confirmed by the laboratory whether it is hornet or not. If not, we use the current model for preliminary confirmation (if the reliability value is greater than the positive report, otherwise it is considered negative report). We adopt the following update strategy:

1. Maintain a queue for a new report, which is initially empty.
2. If no new positive reports appear, then we update our model about once a month. Update model including put the pictures of the new report added to the image recognition of neural network in the training set and the training of neural network further, recount each point to the nearest k report point distance and update the RP value, recount positive report along with the change of in relation to a new curve fitting, and the influence of back calculation note value.

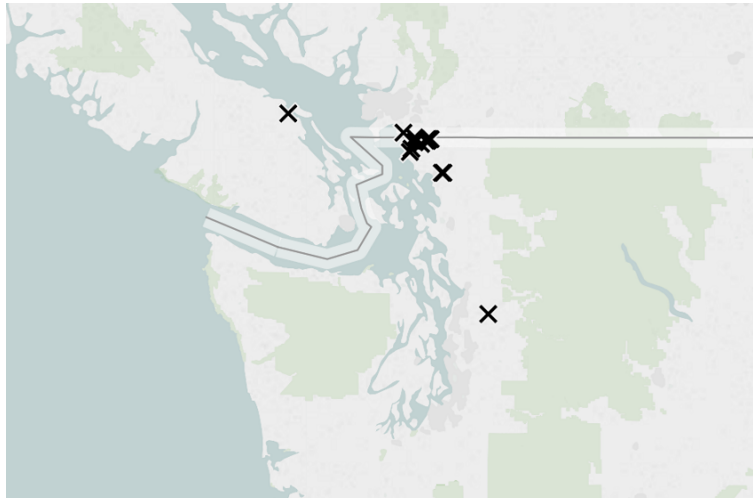


Figure 10: Key Area

3. If new, more positive reports emerge, we immediately update the model.
4. When the update is complete, clear the queue and wait for the next update.

5.5 Task 5

Considering that hornet have a one-year propagation cycle, we chose one year as the time window for observation.

Let the set of all reports received in a certain region within a year starting from a certain month be S_i , The probability of a positive report for the month is measured by p_i , $p_i = \max\{Reliability_r | r \in S_i\}$. As long as $p_i < \alpha$, we think it is unlikely that positive samples will be available this month, and therefore that hornet will not be found, so we call this month "Safe".

But it depends a lot on the month we choose. In some months people send very few reports, and it is very likely that all the reports they receive match the missing photos or missing comments, leading to the false impression of a lower indicator. But through the analysis of the data before it is known that people report number downturn will not more than 6 months, so we just need to continue to examine the possibility of positive

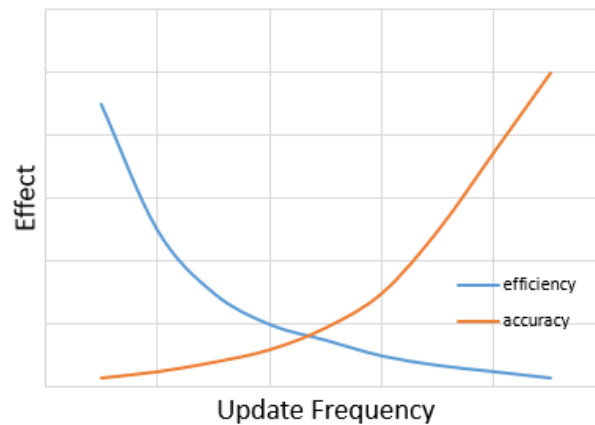


Figure 11: Change Rule

report every month within six months, 6 months as long as it is "safe", so we think the probability of the region found Asian hornet is extremely low, can think Asian hornet has been eliminated in the region.

6 Model Analysis

The sensitivity of our model mainly comes from our determination of the weight of each factor in the index. Therefore, we change the weight of each component respectively to observe its influence on the results of analyzing the topic data of our model.

Therefore, the model is insensitive to time and locations, and relatively sensitive to notes and photos.

6.1 Model Sensitivity Test

The sensitivity of the model mainly comes from our determination of the weight of each factor in the index. Therefore, we changed the weight of each component respectively to observe its influence on the result of analyzing the topic data of our model.

6.2 Strengths

- Rapid prediction

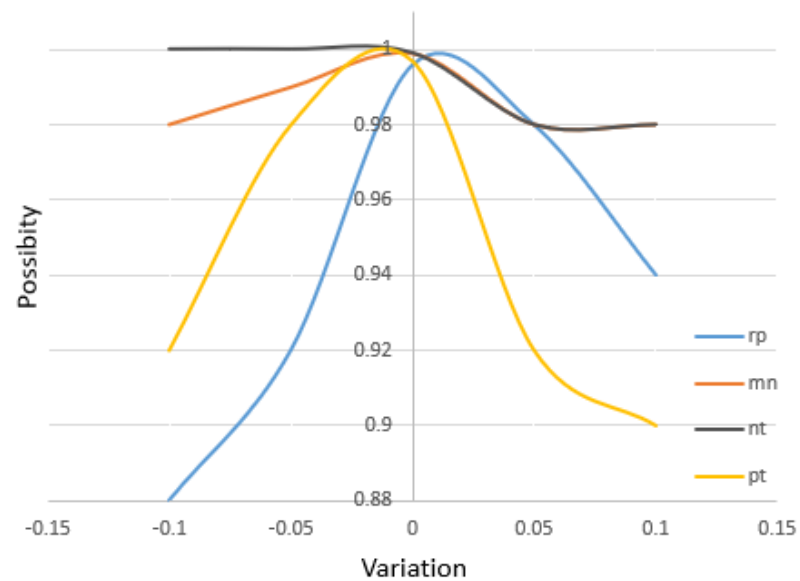


Figure 12: Sensibility

- Fast Iterative Update

6.3 Weaknesses

- Recalculating the neural network is time-consuming

References

- [1] R. Mushtaq, "Augmented dickey fuller test," 2011.
- [2] L. Myers and M. J. Sirois, Spearman Correlation Coefficients, Differences between. American Cancer Society, 2006. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/0471667196.ess5050.pub2>
- [3] S. H. White, A. M. del Rey, and G. R. Sánchez, "Modeling epidemics using cellular automata," *Applied Mathematics and Computation*, vol. 186, no. 1, pp. 193–202, 2007. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0096300306009295>
- [4] D. Franklin, S. Datta, G. E. Budge, G. E. Budge, M. J. Keeling, and M. A. Brown, "Predicting the spread of the asian hornet (*vespa velutina*) following its incursion into great britain," *Scientific Reports*, 2017.
- [5] L. K. Francis L. W. Ratnieks, "Queen control of egg fertilization in the honey bee," *Behavioral Ecology and Sociobiology*, 1998.
- [6] A. A. Hakim, A. Erwin, K. I. Eng, M. Galinium, and W. Muliady, "Automated document classification for news article in bahasa indonesia based on term frequency inverse document frequency (tf-idf) approach," in *2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 2014, pp. 1–4.
- [7] S. Shiraishi, T. Obata, and M. Daigo, "Properties of a positive reciprocal matrix and their application to ahp," vol. 41, no. 3, pp. 404–414, 1998.