Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2125555

Into The Asian Giant Hornet:

A Tracking Method of Population Combining Rapid Detection of Related Suspicious Sightings

Summary

In recent years, more attention has been paid to the phenomenons of biological invasions, which harm local ecosystems and cause economic losses to humans. After a colony of Vespa mandarinia (also known as asian giant hornets) was discovered and quickly destroyed on Vancouver Island in British Columbia, Canada, several confirmed sightings have been reported in neighboring Washington State till now, accompanied by a large number of suspicious cases which were either unverified or negative. However, government officials often expend too much effort in working with these datasets. It's necessary to spend limited time and energy on more investigations, which means that a fast and reliable model based on the reported image data is needed to identify the Vespa mandarinia and other types of insects.

In this paper, we independently propose the **RRS** model to analyze the given data and address the corresponding issues. **RRS** stands for Random Forest Model, ResNet-18, and Spatially Distributed Reproduction Model. The researchers have pointed out that surveillance is a crucial step for the state to recognize and detect the amount of Asian Giant Hornets. The **RRS** model consists of a series of models that combine geographic-based, time-based, and image-based measures to predict the spread, prioritize the government's limited resources to follow-up with additional investigation, and analyze possible extinction.

The idea of our paper can be expressed as follows. In subsection 2, we list the main assumptions for the model construction and label all variables that will be used in our **RRS** model. In subsection 3, preliminary processing, filtering and optimization of the dataset are performed and we reached some elementary conclusions. In the following subsection, a machine learning model is introduced which sets different values for the label status to get a single output to answer questions such as how your classification analysis leads to prioritizing investigation of the reports most likely to be positive sightings. In subsection 5, an image processing model is applied to analyze all the image data from public reports. By training and analyzing image data of Asian hornets and other similar hornets, A fast and effective response system can be set up to deal with large quantities of relevant data, from which real Asian giant hornets can be identified in a timely manner. A third spatial distribution model will be introduced in subsection 6 to analyze the reproduction model of Asian giant hornet populations at the spatial level. We will demonstrate the advantage and weakness of our models in subsection 7. Eventually, we conclude in section 8 to summarize all the data we analyzed could be used and all the references and source code are attached in the reference list.

Keywords: Random Forest; ResNet-18; Spatial Distribution

Contents

1 Introduction

1.1 Problem Background

Species in nature are in the dynamic of constant migration and diffusion, whose spread has further been aggravated by human activities, allowing many organisms to break through geographic isolation and expand into outer environments. The rapidly increasing cases of biological invasion have been realized and recognized in modern times. Due to the lack of predators in the new environment or the attribute of the incoming species themselves, there will always be unpredictable consequences until people actually see it happening.

In 2019 the State of Washington received several reports of detection of Vespa mandarinia, also known as Asian Giant Hornets, after the colony of which was discovered and quickly destroyed in the neighboring Vancouver Island in British Columbia, Canada, attracting massive public attention since the occurrence of its nest was alarming showing the sign of invading and destroying nests of European honeybees in the past and having a potential severe impact on local honeybee populations. Another reason raising people's awareness is that agricultural pests will also be in danger with the emergence of this kind of hornets.

1.2 Work Overview

The State of Washington has recently established a website for the public reports on suspicious appearance of related bee species and several confirmed cases have already been uploaded to the dashboard. Provided with the data collection of all public reports, one of the biggest challenges is to interpret the meaning behind the data provided by the public and design certain strategies to prioritize public reports for additional investigation in the future.

Our paper will explore and address the following aspects:

1. Address and discuss whether or not the spread of this pest over time can be predicted,and with what level of precision.

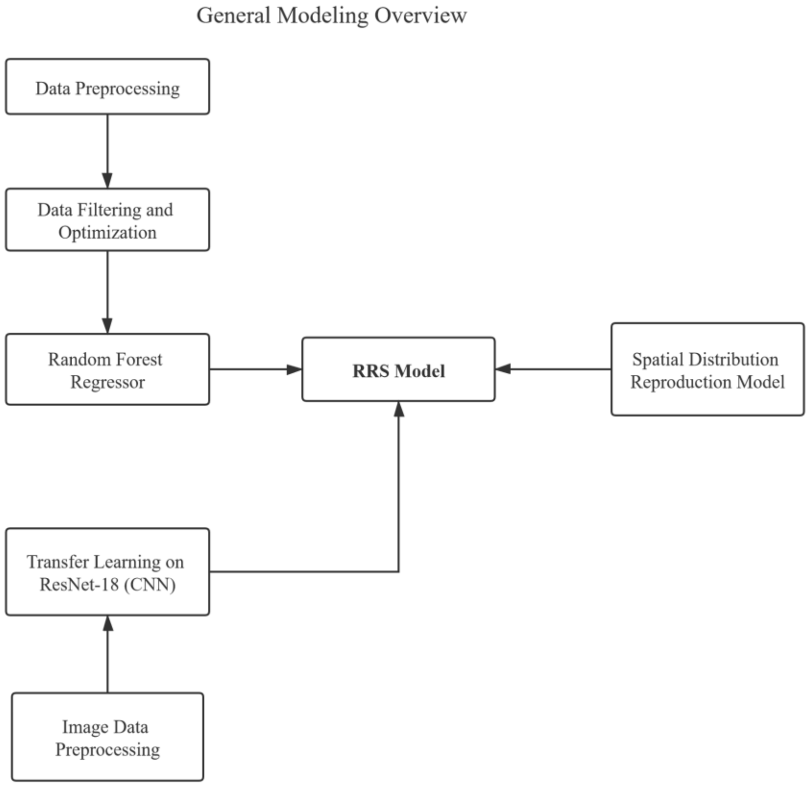
2.Most reported sightings mistake other hornets for the Vespa mandarinia. Use only the data set file provided, and (possibly) the image files provided, to create, analyze, and discuss a model that predicts the likelihood of a mistaken classification.

3.Use your model to discuss how your classification analyses lead to prioritizing investigation of the reports most likely to be positive sightings.

4.Address how you could update your model given additional new reports over time, and how often the updates should occur.

5.Using your model, what would constitute evidence that the pest has been eradicated in Washington State?

The overview of our work is presented in figure 1


General Modeling Overview
Data Preprocessing
Data Filtering and Optimization
Random Forest Regressor
RRS Model
Spatial Distribution Reproduction Model
Transfer Learning on ResNet-18 (CNN)
Image Data Preprocessing

2 Assumption and Nomenclature

2.1 Assumptions

To apply our **RRS** model smoothly to the problem, we make the following assumptions :

(1) The samplings are representative, which is often a trivial assumption by all statistical models.

The following assumptions are necessary to SDRM model:

(1) The biological invasion of the Asian Giant Hornet starts from an original point.
(2) Spatially, the Asian Giant Hornets' nests are spread in spheres and there is a maximum spread rate per year.
(3) The distribution of the likelihood of an Asian Giant Hornet could appear in a certain place is proportional to the φ function of the normal distribution.

2.2 Nomenclature

The input are given in terms of:

x	Longitude of the report
y	Latitude of the report
t_detection	Detection date of the report in terms of days after 1/1/2020
t_submission	Submission date of the report in terms of days after 1/1/2020

The scores generated within the model

cnn_score	The score represents the likelihood of a record to be positive given by the CNN model (Positive or Negative) CNN is trained from all verified data
cnn_score_sampling	The score represents the likelihood of a record to be positive given by the CNN model (Positive or Negative) CNN is trained from all verified data after over sampling
rf_score	The score represents the likelihood of a record to be positive given by the RF model (Real number between 0 and 1) RF is trained from all verified data after over sampling.
sdrm_score	The score represents the likelihood of a record to be positive given by the SDRM model (Real number between 0 and 1)
rrs_score	The score represents the likelihood of a record to be positive given by the CNN model (Real number between 0 and 1)

3 Data Analysis

We first handle and analyze the dataset from all the public reports. We calculated the date by three factors. The data in subsection 3.1 shows the general distribution of all the submitting reports based on the Lab status, the data in subsection 3.2 is the percentage of positive reports in a specific region we chose and among all the reports based on location, and subsection 3.3 will show the frequency of positive cases based on date.

3.1 Distribution Graph

We first made the distribution graph for all the data based on the Lab Status and the result graph as shown in Figure 2 provides us a similar graph compared with the one given by the problem. From the graph we can tell that the actual positive cases are all happening on the border of Washington and British Columbia. Among all the 4441 rows of data in the DataSet, there are 14 reports with Label Status as “Positive” which means the image of hornet in the image is the Asian Giant Hornet and there are also 2069 negative status reports, 2342 unverified reports and 15 unprocessed reports.We get the corresponding information by creating a frequency distribution histogram as shown in Figure 3. As it shows that the small ratio of positive status to other non positive reports implies that the existence of Asian Giant Hornets in Washington is very rare to some extent. From the other side, it also implies that the people are very likely recognizing the wrong sightings.

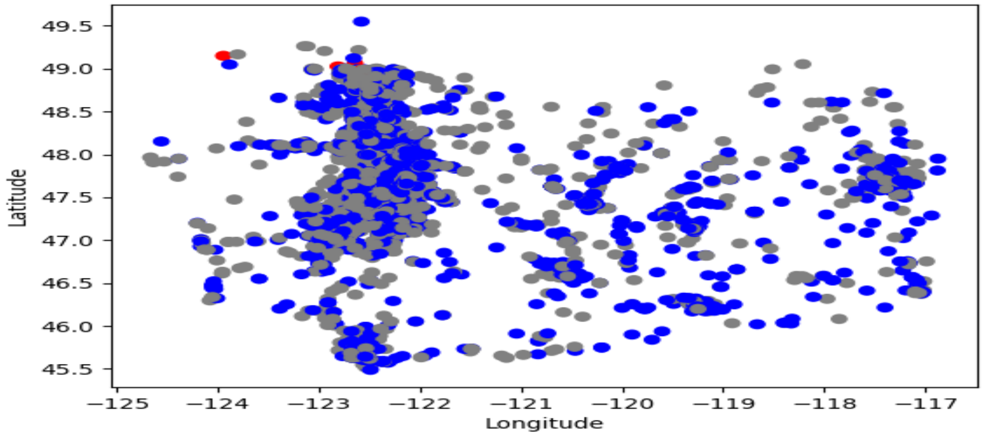


Figure 2: The distribution graph takes the longitude and latitude of images as parameters and labels all the points based on the lab status. The red points denote positive status, blue indicates negative status, and grey indicates all unverified data.

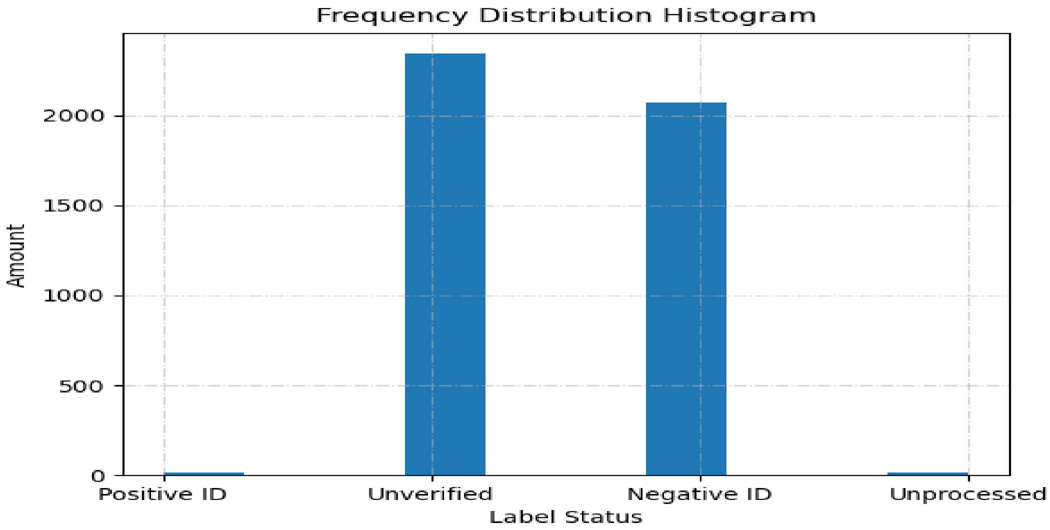


Figure 3: The histogram shows that more than half of the submitting reports are actually unverified images and in the total of 2083 data which has been confirmed of either positive or negative the ratio of positive is 0.6%.

3.2 Percentage of Positive Reports in Different Regions

We then convert the data into a sector diagram for the whole map. In Figure 4, it shows the frequency of positive status among the whole map with a percentage of 0.315%. The percentage of such less cases will affect the final result of our model and therefore further optimization of data is necessary and will be explained further in the later article. Furthermore, in order to get a better understanding of the sector diagram, we have chosen a more specific region to do the research. We have located a range of longitude and latitude which includes all the positive status reporting and in Figure 5 we this time found a better percentage for the positive among all the status in such a region.

	Counts	Percentage Through Whole Map(%)	Percentage Through All Positive Region(%)
Positive	14	0.315	7.071
Negative	2069	46.904	28.788
Unverified	2342	52.781	64.141

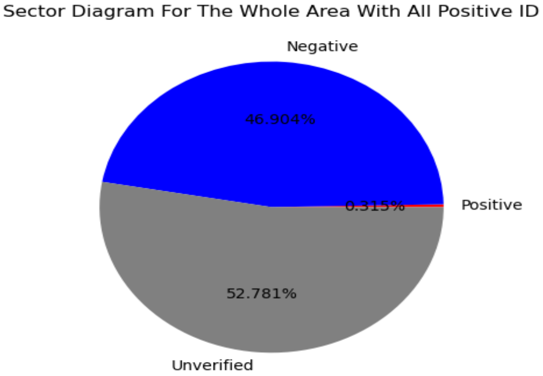


Figure 4: The graph demonstrates the percentage of positive cases among the whole map.

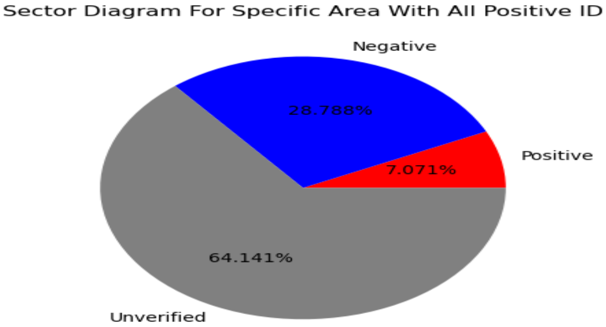


Figure 5: The new histogram shows a new sector diagram for a specific area we decided on the whole map. The interval of longitude is (-123.943134, -122.574726) and the interval of latitude is (48.777534, 49.149394).

3.3 Frequency Graph Based on Detection Date

We have also analyzed the date from the aspect of date. After we processed the date, we noticed that there are some abnormal dates in the file and in the pretreatment step we deleted all the data which looks unreasonable in order to better handle the data. By counting all the positive status, we made Figure 6 which shows the frequency distribution of when the Asian Giant Hornets are found and all of them happened in 2019 and 2020. At the same time, we noticed that they are most found by people in Sep 2020.

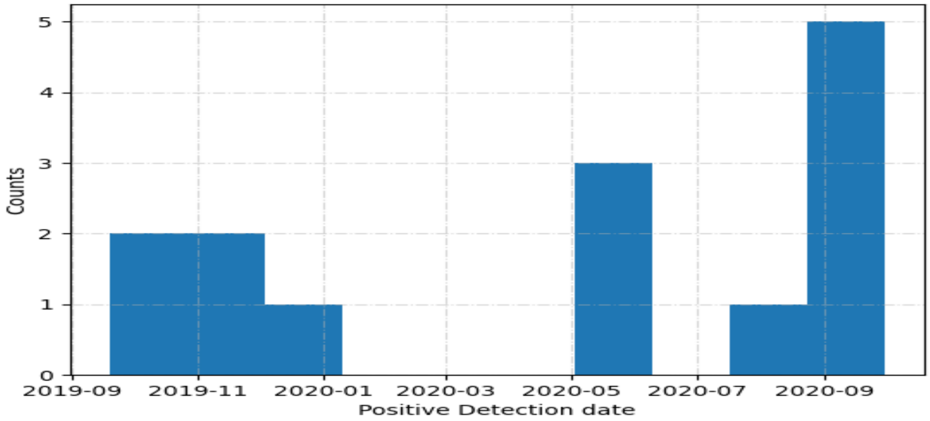


Figure 6: The frequency distribution summarized by date is shown above with an irregular rule through these two years.

4 Machine Learning Model

In this section, we proposed to set up a machine learning model to evaluate a likelihood estimation of finding the Asian giant hornets in the corresponding region as the given location. The key factor in this model is the Lab Status as it is the determining factor for the public to know that Asian giant hornets exist in the related place. We applied a Random Forest Regressor model[1] to predict the possibility of finding such species given its longitude and latitude and also predict the likelihood of a mistaken classification.

4.1 Pretreatment Process

In this section we first process all the data in the file. Since the lab status given are words, we assigned a fixed value for each status so that we can get a better possibility calculated for the map. We decided to assign all “Positive” with value of 1, all “Negative” with value of 0, and all “Unverified” with value of 2. By giving a numerical value for the status, the model can return a more specific output eventually. In addition, instead of having the date as the exact date, we transferred all the dates to a new way to record them by calculating the difference between the detection date and Jan 1st 2020. Appendix has a specific step of how the date transformation is calculated. After doing those steps, a new file is generated which includes all the information provided and new processed data.

4.2 Model Implementation and Results

There are several steps in the model application step to get the final output. After we processed all the data into the way we wanted to calculate, we did one step to optimize the data. As it says that the application of data optimization can find a better solution for a given function [2]. As we figured out in the previous step, there are 14 positive labels with value 1 and 2069 negative labels. The ratio of 0.0068 is way too small for a model to return the result we expect. As says, the result will probably return a negative result no matter what the input is. Even if we submit a graph for a real Asian giant hornet, the model will return a negative output. Therefore, we used the SMOTE method to optimize our data so that we can handle the imbalance situation that is happening in the given dataset. The SMOTE method, this well-known method especially created to handle imbalance data by applying its particular formula[3][4], generated more positive values based on a certain ratio so that the model can better learn when to notify the correct result. The difference of unverified data and the oversampling data can be told from Figure 7 and 8. In addition, we first use a **RRS** model to get a score for all the data. The scores are stored in another new file. At last, by combining the data file that we processed before and the score file into a new model. We will get the final graph of the result as shown in Figure 9. Details of getting such possibilities are given in the Appendix. On the figure, the color is changing from deep to light as we view the graph from top to the bottom. Each value is in the range of (0,1) as 0 indicates 0 possibility of finding the Asian giant hornets and 1 indicates definitely finding the creature.

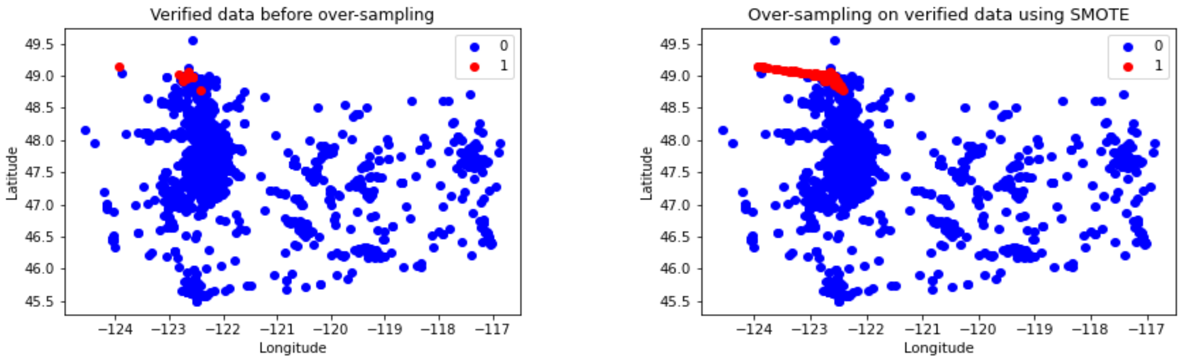


Figure 7, 8 demonstrates the difference between unverified data and oversampling data generated by SMOTE method.

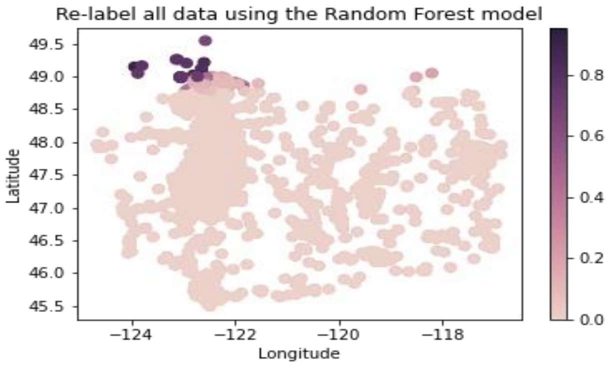


Figure 9: The graph shows the possibility of finding a Asian giant hornet in the future ranging from 0 to 1. As we compare this graph with the distribution graph, we can obviously tell that the place with a positive label has a color way more darker than the rest place on the map.

4.3 Problems That This Model Can Potentially Solve

The final graph generated by the machine learning model looks ideally for us. It is pretty clear to tell which location it is more likely to detect the Asian giant hornets. This model can also be used with other models to predict the spread of the pest in the future as it is likely to spread from Washington or British Columbia if the creature still exists in the next year with all conditions that support the hornets to survive through next year. More precise prediction might be done with more positive reports submitted through the website. In addition, the result from the model can be improved if additional data is given since the current result is produced by data which is produced using a SMOTE method. Although such methods can generate more positive data by a functional calculation, more real positive cases will make the machine learning process work more efficiently.

5 Image Processing Model To Classify Mistaken Sighting

In this section, we construct a new image processing model based on the images provided in the data to mainly analyze the spread of Asian giant hornets in the future and predict the likelihood of mistaken classification. The rest of this section is arranged as follows. In section 5.1, the model is implemented in detail from the beginning. In section 5.2, we analyzed the output from the model and compared it with our prediction based on the given value.

5.1 Image Processing Model

We noticed that there are thousands of images and videos provided as information and it is crucial to analyze the images for a model to determine the uploading image is true Asian giant hornets. The image processing model will be a powerful tool to analyze the data. We first separate all images and videos into two subfolders to do the pretreatment separately. The images are resized to size of 416 * 416 as we decided before we set the input to the model. By resizing the images into the same size could make the model better handle the main project in each image. The code of resizing could be found in Appendix. After handling the images, we make a clip every 150 frames from the video and resize the image into the same size after pretreatment as well. After the steps above, the images will be put into different folders once again based on their lab status so that we can gather all images with positive status. Then we split all the images into a training set and a value set with a ratio of 8:2 and the preparation is done. The model we applied for the image processing Model is called CNN model, which is a mature model for handling images. What the model did is that it will first relebal all the images from the input. During this process, it will have its own scoring system to determine the image that we want, which is the exact image of Asian giant hornet. In addition, what the model did later on is described by “deep learning CNN models to train and test, each input image will pass it through a series of convolution layers with filters (Kernals), Pooling, fully connected layers (FC) and apply Softmax function to classify an object with probabilistic values between 0 and 1.[5] The whole process of applying the model for our data is in Appendix. It processed all the images with different lab status and there will be a single output file. (to be continued)

5.2 Output From the Model

The output is a file with brand new labels for each image. When we plot the data into a graph, how the graph looks is amused but in our expectation. Figure 10 shows that all the points on the graph are blue which indicates that the graph is not a Asian giant hornet. This situation is in our expectation as the ratio of positive images to the whole images provided is too small. Therefore, it is very likely that all the input images will be assumed as negative.

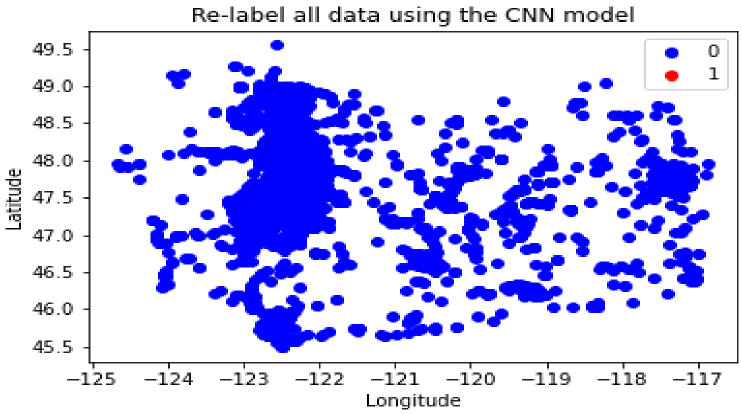
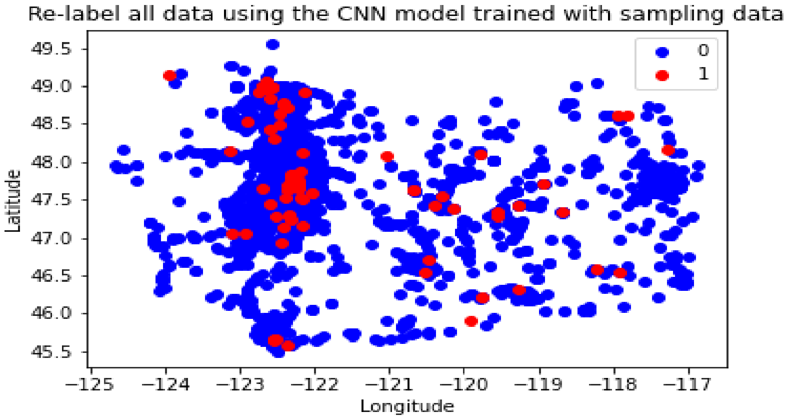


Figure 10: This graph generated by the output of CNN model shows that the current figure information is only able to classify the report image as not as Asian giant

5.3 Improvement and Potential Solving Problems

We further did an oversampling step for this model as we did for the machine learning model in the previous section. In order to oversample the images data, we copied the images with positive status in the dataset and by doing so we reconstructed the ratio of training data to 1:15 which is a much better ratio for the CNN model to classify the correct image. This time, the model provided a graph with much more red points spreading through the map as shown in Figure 11. The new graph looks like it will spread to a high extent while this case happens due to we are analyzing the images only and the geography information is not considered in this circumstance. We believe that the images provided are valuable resources and while the given images are inadequate to practice a model to give feedback with high precisions. The model can be used to predict the likelihood of a mistaken classification but the current model will probably set the image with a negative status even if the image is actually an Asian giant hornet.



6 Model Construction

6.1 Spatial Distribution Reproduction Model (SDRM)

The SDRM model is part of the RRS Model and is used to score each report. Refer to Figure 2, all positive reports are located in the northwest of Washington. So we made the following assumptions:

1. The biological invasion of the Asian Giant Hornet starts from an original point. Because the first verification Positive ID is at (-123.943134, 49.149394) on 9/19/2019(T =-104) with globalID ={124B9BFA-7F7B-4B8E-8A56-42E067F0F72E}. We assume that to be the original point.

2. Spatially, the Asian Giant Hornets' nests are spread in spheres. And the radius of the sphere is the maximum distance a Hornet could fly a year, which is ***max_displacement = 30km /year * t years***, given that a Asian Giant Hornet could move its nest at most 30 km a year.

3. The variable ***ratio*** represents the distance between the original point and where a Hornet might appear.

4. The ratio between distance and max_displacement ∈[0, 1]. Therefore, ***ratio = max (min(distance /max_displacement, 1), 0).***

5. We assume that the ratio between ***distance*** and ***MAX_DISPLACEMENT*** is positively correlated with the possibility that the report is positive in terms of Z-score.

$$\varphi(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$$

Where

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

So we made the SDRM scorer model:

Max_displacement = max_spreading_rate * delta_t
ratio = max(min(distance /max_displacement, 1), 0)
Sdrm_score = 1- (z(ratio) - 0.5) * 2

Where z is the \phi function of standard normal distribution.

6.2 RRS Combined model

RRS is combined by 3 parts, which each offer a score that evaluates the likelihood of a report to be positive。

$$\begin{aligned} \text{RRS score} = \ & \alpha * \text{RF score} \\ & + \beta* \text{CNN score} \\ & + \gamma* \text{SDRM score} \end{aligned}$$

• Since RF is the model with the least assumptions and the best performance, α should be the factor with the largest weight in the RRS score.

• Since CNN is solely based on image recognition and the existing training set is extremely imbalanced, the performance of the model is not ideal. So the weight of β is slightly less than that of α.

• Since SDRM is a simple theoretical model that only satisfies the situation of Washington State, we believe that it is a model with large bias. Therefore, γ should be the one with least weight.

6.3 Update Modeling with newly income reports

In reality, because the concept of Machine Learning is to predict the future by the given experience. Therefore, the two machine learning models inside RSS -- RF and CNN -- should also be re-trained regularly.

In this case, we could retrain the model on a monthly basis. (CNN's image preprocessing has been combined in run_me.sh. For CNN's training, please refer to transform_learning.ipynb, and for RF training, please refer to machine_learning.ipynb for RF training)

7 Model Evaluation

7.1 Strengths

1. The RSS model is a combined model whose three parameters are derived from three different approaches to the model. So RSS is a better fit for the real world.
2. The only assumption of Random Forest is that the samplings are representative, which is often a trivial assumption by all statistical models, which makes our model suitable for other similar situations in the real world.
3. We use transform learning in the CNN model with ResNet-18, which makes the training time of the model drops dramatically.
4. We use SMOTE model to handle the imbalanced data to optimize the performance of the model.

7.2 Weaknesses

1. In SDRM model, we assume that the ratio between ***distance*** and ***MAX_DISPLACEMENT*** is positively correlated with the possibility that the report is positive in terms of Z-score, which could lead to a relatively large error according to the real world.

2. The training set of CNN is too bad to make an ideal performance. But in real world, we could use more suitable training set to training the model in order to obtain a better performance.

8. Conclusion

A surge in biological invasion cases should be noticed and prepared well by the related department since it is unpredictable how the environment will evolve in the future. Just like the case that Asian giant hornets destroyed a whole nest of honey bees, while the truth of there is no predator of this terrifying creature will bring anxiety to people. Under such conditions, building mathematical models to predict the spread of this species and detecting the right appearance has a significant meaning.

In our **RRS** model, we handled the given data from different angles. By analyzing the geography location, images of reporting, and date, we have built several models including Random Forest Model, CNN with image processing, and Spatially Distributed Reproduction Model. By combining them together we can handle different questions such as predict the spread of the Asian giant hornets and to some extent predict the likelihood of mistaken classification. According to the predictions given by our RRS model, it is very much likely that the Asian hornet will eventually become extinct in the region. What’s more, government departments can rely on this model to quickly and accurately classify and process information reported by suspected hornets according to the general scores given by model, resulting in significant time savings, resource savings and office efficiencies.

Memorandum

TO: MCM

FROM: Team 2125555

DATE: February 9th, 2020

RE: Confirming the Buzz About Hornets

The confirmation of cases of Asia giant hornets in the State of Washington and British Columbia has brought both the attention of the United States and Canada. The public concern about the spread of these hornets and the damage it might cause to a new environment without predators are discussed by people more and more frequently. The Asian giant hornets group found in Washington has a queen hornet, which means that after they live through the winter it might breed and have more giant hornets in the United States. And a new queen has a range estimated at 30km for establishing her nest. Therefore, some measures must be taken to make accurate predictions. Besides making predictions on the spread of hornets to other places, another task is to predict the likelihood of mistaken classification. In the given data, the fact of only 14 positive reports existing among more than four thousand of reporting data is happening not only due to the fact that the amount of Asian giant hornets is very small in Washington, but also because of people recognizing the wrong species. Many images of honeybees which have existed in the United States for years are submitted to the website. We applied three models in total to make the prediction. We checked the result by those models individually at first and then we tried to combine the models together to see how it works.

The results of three models come as follows.

- **Geography Machine Learning Model**

One of the results comes from the machine learning model using all geography locations. Since all the information about longitude and latitude for each report are given by the data, our first model calculates the possibility of existing of Asian giant hornets through the whole map. Our first graph based on the given location looks similar to the one given in the problem. This is the graph without any processing. Then, we tried to make some improvements to the given data for better prediction and generating better models. We used the SMOTE method for handling the imbalance value in the given data. Then, by taking the data after pretreatment as input and applying the Random Forest Regression model, we received an output with each location assigned a new value range from 0 to 1. The value now indicates the possibility of each place finding Asian giant hornets in the future. This graph seems reasonable as it is showing a high possibility represented by deeper color at the places where the official department has confirmed the hornets and the place way far away from the border has a value almost equal to 0.

- **Image Processing Model**

Another result is produced by using a completely different model. The 4441 images provided could be used to analyze and predict the possibility of mistaken classification. We first resized all the images into the same size and cut the videos into images. Then a CNN model is applied to learn to recognize the true Asian giant hornets. However, one of the obstacles we encounter through the whole process is that the ratio of positive status images to the whole images are too small so that the model will eventually determine all the input images to a negative output no matter if the image is Asian giant hornets or not. We have tried to optimize the ratio as well. We copied all the positive images and enhanced the ratio to 1:15. Then, the new result shows a scenario of hornets spreading all over the map. This is happening since this time we only consider the image as the parameter in the model and we later make the graph with geography location which is not related to the output. This model can predict the likelihood of mistaken classification as we assumed and by updating the images of positive status in the future can improve the model better.

- **Combination of different models**

Last but not least, we tried to combine those models together to consider both the factor of image and geography location which can optimize the effect of prediction. We believe that the combination of models can handle the questions asked in the problem such as we can predict the spread of hornets. In addition, the key factor for constructing models is to have more confirmation of the appearance of Asian giant hornets or positive images. It is also not necessary for people to be too anxious for the species since according to history, there was one case of Asian hornets arriving in New York many years ago and the result was that there were none of them in New York now. [8] So it can be guessed that the Asian giant hornets are eradicated in the state of Washington.

Appendices

process_images.sh

```
#!/bin/bash
# process_images.sh
# 1. move png, jpg, and jfif files into the folder `images`
# 2. move mov and mp4 files into the folder 'video'
mkdir data
DIR=2021MCM_ProblemC_Files
cd $DIR || exit
mkdir images videos
mv ./*.png ./*.PNG ./*.jpg ./*.JPG ./*.jfif ./*.JFIF images/
mv ./*.mov ./*.MOV ./*.mp4 ./*.MP4 videos/
```

Machine_learning.ipynb

```
from IPython.display import display
import numpy as np
import pandas as pd

#%%

# Load Data
from preprocess import load_and_process

# data: (5612, 9)
# Detection Date: days after 1/1/2020
# Lab Status:
#   0: Negative ID
#   1: Positive ID
#   2: Unverified
data, image_id = load_and_process()

display(data.sort_values(by=['Detection Date']).head(20))
data.to_csv('data.csv')

#%%

# It's too long ago so we drop data before 2010, whose detection date < -4000 days
data = data[data['Detection Date']>-4000]

display(data.head())

#%%
```

```
# Select data used to train the model
X = data.loc[:,['Detection Date', 'Submission Date', 'Latitude', 'Longitude']]
y = data['Lab Status']

# Drop rows that at unverified
X_data = X[y!=2]
y_data = y[y!=2]
X_unverified = X[y==2]
y_unverified = y[y==2]

#%%

# Use SMOTE to handle imbalance of data

from imblearn.over_sampling import SMOTE

oversampling = SMOTE(sampling_strategy=0.1, k_neighbors=3, n_jobs=n_jobs)
X_sampling, y_sampling = oversampling.fit_resample(X_data, y_data)

display(X_sampling.head())
display(y_sampling.head())

print(y_sampling.sum())
#%%
# Make a plot for X_sampling and y_sampling
import matplotlib.pyplot as plt

c_dict = {
    0: "blue",
    1: "red"
}
fig, ax = plt.subplots()
for label in c_dict.keys():
    plt_data = y_sampling==label
    ax.scatter(X_sampling.loc[plt_data, 'Longitude'], X_sampling.loc[plt_data, 'Latitude'],
c=c_dict[label], label=label)
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.title("Over-sampling on verified data using SMOTE")
ax.legend()
plt.savefig('fig/oversampling_smote.png')
plt.show()
#%%
# Make plot for data before oversampling
fig, ax = plt.subplots()
for label in c_dict.keys():
    plt_data = y_data==label
    ax.scatter(X_data.loc[plt_data, 'Longitude'], X_data.loc[plt_data, 'Latitude'], c=c_dict[label],
label=label)
plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.title("Verified data before over-sampling")
ax.legend()
plt.savefig('fig/verified_data.png')
plt.show()
#%%
# Try Random Forest
from sklearn.ensemble import RandomForestClassifier, RandomForestRegressor
clf_rf = RandomForestRegressor(n_estimators=40, max_depth=3, n_jobs=n_jobs)
# print(cross_val_score(clf_rf, X_data, y_data, scoring='f1', cv=5).mean())
# print(cross_val_score(clf_rf, X_sampling, y_sampling, scoring='f1', cv=5).mean())
#%%
# Predict the label of unverified records

clf_rf.fit(X_sampling, y_sampling)

y_pred_on_unverified_sampling_rf = clf_rf.predict(X_unverified)
y_pred_on_unverified_sampling_rf =
pd.DataFrame(y_pred_on_unverified_sampling_rf.reshape((-1,1)), index=X_unverified.index,
columns=['rf_score'])

display(y_pred_on_unverified_sampling_rf.head())

#%%

# Plot the predicted label of unverified records

import seaborn as sns

cmap = sns.cubehelix_palette(as_cmap=True)

f, ax = plt.subplots()
plt.scatter(X_unverified['Longitude'], X_unverified['Latitude'],
c=y_pred_on_unverified_sampling_rf['rf_score'])
points = ax.scatter(X_unverified['Longitude'], X_unverified['Latitude'],
c=y_pred_on_unverified_sampling_rf['rf_score'], s=50, cmap=cmap)
plt.xlabel('Longitude')
plt.ylabel('Latitude')
f.colorbar(points)

#%%
```

```
# Re-label the entire dataset

y_pred_on_sampling_rf = clf_rf.predict(X)
y_pred_on_sampling_rf = pd.DataFrame(y_pred_on_sampling_rf.reshape((-1,1)),
index=X.index, columns=['rf_score'])

display(y_pred_on_sampling_rf.head())

#%%

# Plot the re-labeled dataset

f, ax = plt.subplots()

plt.scatter(X['Longitude'], X['Latitude'], c=y_pred_on_sampling_rf['rf_score'])
points = ax.scatter(X['Longitude'], X['Latitude'], c=y_pred_on_sampling_rf['rf_score'], s=50,
cmap=cmap)

plt.xlabel('Longitude')
plt.ylabel('Latitude')
plt.title("Re-label all data using the Random Forest model")
f.colorbar(points)

plt.savefig('fig/re-labeled_rf.png')
```

Sdrm.ipynb

```
from IPython.display import display

import numpy as np
import pandas as pd

# Load data similar to machine_learning.ipynb
from preprocess import load_and_process

data, image_id = load_and_process()

#%%

# Define SDRM score according to the formula

# Max spreading rate of the queen of the Asian Giant Hornet is 30 km per year
MAX_SPREADING_RATE = 30
```

```
# The first verified Positive ID is at (-123.943134, 49.149394) on 9/19/2019(t=-104)
# with GlobalID={124B9BFA-7F7B-4B8E-8A56-42E067F0F72E}
# which is used as the original point.
X_0 = -123.943134
Y_0 = 49.149394
T_0 = -104

# Constant used to convert coordinate degree into km
KM_PER_DEGREE = 111

#%%

# Define the function to calculate distance between to Coordinate
import numpy as np

def distance(x, y):
    d_x = x - X_0
    d_y = y - Y_0

    d_x_km = d_x * KM_PER_DEGREE
    d_y_km = d_y * np.cos(np.abs(x) / 180 * np.pi)

    return np.sqrt(d_x_km ** 2 + d_y_km ** 2)

#%%

# Calculate the sdrm_score with
# sdrm_scroe
import scipy.stats as st

def sdrm_score(x, y, t):
    max_displacement = MAX_SPREADING_RATE * (t - T_0 + 1e-6) / 365
    ratio = max(min(distance(x, y) / max_displacement, 1), 0)
    score = 1- (st.norm.cdf(ratio) - 0.5) * 2
    return score

#%%

# Re-label all data
n = data.shape[0]

y_pred_sdrm = np.zeros(n)

for i in range(n):
    y_pred_sdrm[i] = sdrm_score(data.loc[data.index[i], 'Longitude'], data.loc[data.index[i], 
'Latitude'], data.loc[data.index[i], 'Detection Date'])
```

```
y_pred_sdrm = pd.DataFrame(y_pred_sdrm.reshape((-1,1)), index=data.index,
columns=['sdrm_score'])
```

Transform_learning.ipynb

(Removed codes copied from
https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html)

```
#%%

# Train a model with oversampling

# *************************************************************
# ***** CAUTION ***********************************************
# Need to exec `oversampling.sh` for over sampling the images
# *************************************************************

data_dir = 'data_oversampling'
image_datasets = {x: datasets.ImageFolder(os.path.join(data_dir, x),
                                          data_transforms[x])
                  for x in ['train', 'val']}
dataloaders = {x: torch.utils.data.DataLoader(image_datasets[x], batch_size=4,
                                             shuffle=True, num_workers=4)
              for x in ['train', 'val']}
dataset_sizes = {x: len(image_datasets[x]) for x in ['train', 'val']}
class_names = image_datasets['train'].classes

#%%

# Make the model
model_conv_on_sampling = torchvision.models.resnet18(pretrained=True)
for param in model_conv_on_sampling.parameters():
    param.requires_grad = False

# Parameters of newly constructed modules have requires_grad=True by default
num_ftrs = model_conv_on_sampling.fc.in_features
model_conv_on_sampling.fc = nn.Linear(num_ftrs, 2)

model_conv_on_sampling = model_conv_on_sampling.to(device)

criterion = nn.CrossEntropyLoss()

# Observe that only parameters of final layer are being optimized as
# opposed to before.
```

```
optimizer_conv = optim.SGD(model_conv_on_sampling.fc.parameters(), lr=0.001,
momentum=0.9)

# Decay LR by a factor of 0.1 every 7 epochs
exp_lr_scheduler = lr_scheduler.StepLR(optimizer_conv, step_size=7, gamma=0.1)

#%%

# Train the model
model_conv_on_sampling = train_model(model_conv_on_sampling, criterion, optimizer_conv,
                       exp_lr_scheduler, num_epochs=25)
```

References

[1]Chakure, A. (2020, November 06). Random forest and its implementation. Retrieved February 08, 2021, from https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f

[2]Beginning scientific computing. (n.d.). Retrieved February 08, 2021, from http://courses.washington.edu/am301/page2/video.html

[3]Brownlee, J. (2021, January 04). SMOTE for imbalanced classification with Python. Retrieved February 08, 2021, from https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[4]Chakrabarty, N. (2020, September 25). Application of synthetic minority over-sampling technique (smote) for imbalanced datasets. Retrieved February 08, 2021, from https://medium.com/towards-artificial-intelligence/application-of-synthetic-minority-over-sampling-technique-smote-for-imbalanced-data-sets-509ab55cfdaf

[5]Prabhu. (2019, November 21). Understanding of convolutional neural NETWORK (CNN) - deep learning. Retrieved February 08, 2021, from https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148

[6] Exotic pests. (n.d.). Retrieved February 08, 2021, from https://beeaware.org.au/archive-pest/asian-hornet/#ad-image-0

[7] Transfer learning for computer vision tutorial¶. (n.d.). Retrieved February 08, 2021, from https://pytorch.org/tutorials/beginner/transfer_learning_tutorial.html

[8] Validate user. (n.d.). Retrieved February 09, 2021, from https://academic.oup.com/aesa/article-abstract/113/6/468/5901552?redirectedFrom=fulltext