A Research of Vespa Mandarinia through Visualization Technology and Convolution Neural Network

Jionghui Duan1,*, Junwei Cheng2, Yitai Cheng3

1School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing 100000
2School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100000
3International School, Beijing University of Posts and Telecommunications, Beijing 102200

*Corresponding author: duanJionghui@bupt.edu.cn

Abstract. After being discovered in September 2019, Vespa mandarinia, a fierce alien invader, quickly got public’s attention. The inhabitants of Washington State flooded to report the suspected cases around them. However, most of them were proved to be false information. How to interpret the provided data and determine the priority of dealing with the problem is of vital importance. Based on this situation, this paper develops a model that robustly and accurately predicts the spread of Vespa mandarinia and the facticity of provided reports. By adopting the idea of mining common behavior patterns, this paper took full use of other wasps’ data to make up for our lacking data on Vespa mandarinia. Though few samples of Vespa mandarinia been provided, this paper took complete utilizing of other data. To evaluate our external model, this paper gave 2019 data of the Vespa mandarinia to propagate then the prediction of 2020 turned to be perfectly accurate. After this, this paper combined the image, location and time information together to predict the sample. This paper got 87.6% accuracy on positive samples and 98.7% accuracy on negative samples. The calculated renewal frequency was proved to be fit for Vespa mandarinia’s actual living habits. And the MSE of our enhanced Logistic function was $10^{-4}$, which can simulate the growth of Vespa mandarinia well and also taken human eradication into consideration.

Keywords: Wasp, Visualization, Convolution neural network, Logical regression model.

1. Introduction

Vespa mandarinia is the largest species of hornet in the world, and the occurrence of the nest was alarming. Additionally, the giant hornet is a predator of European honeybees, invading and destroying their nests. A small number of the hornets are capable of destroying a whole colony of European honeybees in a short time. At the same time, they are voracious predators of other insects that are considered agricultural pests [1]. The life cycle of this hornet is similar to many other wasps. Fertilized queens emerge in the spring and begin a new colony. In the fall, new queens leave the nest and will
spend the winter in the soil waiting for the spring. A new queen has a range estimated at 30km for establishing her nest [2].

Due to the potential severe impact on local honeybee populations, the presence of Vespa mandarinia can cause a good deal of anxiety. The State of Washington has created helplines and a website for people to report sightings of these hornets. Based on these reports from the public [3], the state must decide how to prioritize its limited resources to follow-up with additional investigation. While some reports have been determined to be Vespa mandarinia, many other sightings have turned out to be other types of insects.

2. Analysis of Lab Status, Pictures of Record and Position of sighting of Asian Giant Hornet

2.1. Data Cleaning and Preprocessing

Before data analysis, the availability of data must be guaranteed. No measures, regardless of its value, can provide accurate assessments if based on unreliable data. This paper first remove useless information including submission date [4], on the basis of which this paper can carry out data pre-processing. Besides, this paper figures out that some people related to unique Global ID send more than one pictures of wasps, which may cause confusion when this paper carry out data analysis in the following step of classifying which report sighting information includes image and which does not. Therefore, this paper drop the redundancy Global ID information before the next step [5]. Before data analysis, the availability of data must be guaranteed. No measures, regardless of its value, can provide accurate assessments if based on unreliable data, so this paper use methods below to preprocess the data provided [6].

The competition side provides a lot of data but some of them are not useful. After a preliminary analysis of the data, this paper found that the submission dates are not useful compared with the detection date, so this paper remove the them. At the same time, this paper focused on user-submitted notes and lab comments. After analyzing user-submitted notes by extracting intensity of emotion. This paper find that they are not significantly correlated with positive or negative commits in the submitted cases, so this paper removed them.

2.2. Data Analysis and Visualization

Based on lab status, the data of Asian Giant Hornet can be divided into 4 groups: positive, negative, unverified and unprocessed. By using python program to generate these pie chart, this paper find that the number of valid reports proved positive by laboratory is very small, only taking up 0.3% of total reports. Then this paper split up the data by image property (with or without image). Most of the samples with image are negative, while 98.4% of those without image are unverified and there is no
single data proved positive without image. After classifying distinct images of unverified and unprocessed images, this paper found that 70 of unverified and 5 of unprocessed are all negative. Then this paper removes other unverified cases with unclear images and only comments.

Fig. 2 Classification of Reports of Asian Giant Hornet

3. Classification of Reports of Asian Giant Hornet

3.1. Data Observation and Model Determination
Asian giant hornets typically build their nests underground, usually in abandoned rodent burrows in forests, often in association with pine roots. Nests are sometimes constructed in dead, hollow trunks or roots of trees, but these are never more than 3 to 6 feet above the ground. Aerial nests are rare—of 1,756 nests examined in Japan, only three were constructed above ground. Because of their subterranean nesting habit, locating the nests of Asian giant hornets can be very difficult. After extracting the provided data, this paper found that there're 14 cases of Asian giant hornets reported, 5 in 2019 and 9 in 2020.

Their spread with such a small sample size. At first this paper used the Gray Model to predict the tendency of Vespa mandarinia’s spread and this paper found that the result was not very realistic. After analysing this paper determine that this paper should take full use of the data of other wasps, so this paper assume that all species of wasps have the same essential behavior patterns and the extrinsic mapping is different. So this paper use hidden Markov model to mine the common behavior pattern and combine them with the data of Vespa mandarinia. This paper wants to use this method to get higher accuracy of Vespa mandarinia and the experiment proves that our model performs an excellent result. Finally, this paper utilizes Entropy Weight Method to further solve the over-fitting problem.

3.2. Gray Model
At first gray model is decided to be used to find out the law that plays a role in a certain period and establish the load forecasting model in the case of limited data. GM(1,1) model is one of the most commonly used gray models which is a first-order differential equation model. In this paper, this paper use the enhanced GM(1,1) model (NEED MODIFY) to simulate the spread model to further carry out prediction of the location of Vespa mandarinia overtime. Latitude and longitude this paper predicted by enhanced GM(1,1) model respectively. This paper denotes all the positive latitude in time order as
\[ x^{(0)}(0) = \left( x^{(0)}(1), x^{(0)}(2), x^{(0)}(3), \ldots, x^{(0)}(13) \right) \]  

Then this paper makes one time accumulation for \( x(0) \), we get
\[ x^{(1)}(1) = x^{(0)}(1), x^{(1)}(i) = \sum_{k=1}^{i} x^{(0)}(k)(i = 2, 3, \ldots, 13) \]  

The whitened equation of GM(1,1) is
\[ \frac{dx^{(i)}}{dt} + ax^{(i)} = b \]  

Due to \( x^{(1)}(k) - x^{(1)}(k-1) = x^{(0)}(k) \). Then corresponding gray differential equation for equation (5) is
\[ X \beta = Y \]  

where
\[ X = \begin{pmatrix} z^{(0)}(2) & 1 \\ z^{(0)}(3) & 1 \\ \vdots & \vdots \\ z^{(0)}(13) & 1 \end{pmatrix}, \beta = \begin{pmatrix} x^{(0)}(2) \\ x^{(0)}(3) \\ \vdots \\ x^{(0)}(13) \end{pmatrix} \]

By least square method, we have
\[ x^{(0)}(k+1) = x^{(1)}(k+1) - x^{(1)}(k) = \left( x^{(0)}(1) - \frac{b}{a} \right) e^{-ak} - e^{-a(k-1)} \]  

However, this paper finds that the predicted result is not in accordance with the fact. Because Asian giant hornets typically build their nests underground, usually in abandoned rodent burrows in forests, often in association with pine roots, our prediction shows that they are spreading to the mountains. This paper guess it is because of the small sample size and based on this fact, this paper decided to use hidden Markov model to take full use of other data.

![Fig.3 Predicted distribution of Vespa mandarinia by GM (1,1)](image-url)
4. Hidden Markov Model

In order to generate a predictive result with better fitting effect, this paper introduce Hidden Markov Model (HMM).

The number of positive samples of Vespa mandarinia is too small. Wasp species with similar living and nesting habits this paper selected, and the trend of their prey is taken into account to eliminate the error. Thirteen species of sawflies (Vespula species) occur throughout North America, 10 of which are found in eastern North America. However, queen southern sawflies are sometimes confused for Asian giant hornets when they are active in the spring. Different sawflies species preferentially build their nests in the ground which shares the same will with Vespa mandarinia. These two kinds of creature share the similar nest building habits. Therefore, this paper considers the migrating pattern of sawflies is very similar to Asian giant hornets. And both sawflies and Vespa mandarinia prey on bees. Therefore, this paper use location data of sawflies (more than 500 sets) and bee (more than 200 sets) in the negative region.

![Fig4. Distribution of Sawflies](image1) ![Fig5. Distribution of Bees](image2)

Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (i.e. hidden) states. The distribution of wasp and bee represents the hidden distribution of bee colony, that is, the hidden state sequence representing the state of hidden Markov Chain at certain time. The distribution of Vespa mandarinia is represents the observed distribution of bee colony, that is, the observation sequence in hidden Markov model. And This paper has the following equation

\[ x_{t+1} = Q x_t \]
\[ y_{t+1} = B x_{t+1} \]  

(7)

5. Prediction Results with Discussion of Precision

5.1. Prediction Results

After optimizing our model by EWM, the mean square error is further reduced \(3.7 \times 10^{-4}\), which proves our utilization of EWM is effective and generates more accurate results.

As is known to us, fertilized queens emerge in the spring and begin a new colony and a new queen has a range estimated at 30km for establishing her nest. Based on this fact, this paper modifies the location of predicted points which are out of the circle with the center determined by the last point and a radius of 30km. The way of modification is shown in the following graph.
Then, this paper uses the model to make several predictions, and draw the risk area that Vespa mandarinia might appear in the future.

5.2. Introducing Bilinear Convolution Neural Network
Asian giant hornets are strikingly colored, with yellow heads, a black thorax, and yellow and black or brown striped abdomens. The distance between its forward facing eyes is large.

Given that images provided are all of wasps or bees which share many parts of similarities of body, only doing generic image classification to these images may not generate satisfying precision. Therefore, this paper uses fine-grained image classification which requires the recognition of highly localized features that are independent of posture and position in the image. Compared with (generic)
image classification, fine-grained image classification needs to judge more fine image categories. For example, this paper needs to figure out what type of bird, what type of car, or what type of airplane the target is. Often, the differences between these subclasses are very small. For example, the only visible difference between the exterior of a Boeing 737-300 and a Boeing 737-400 is the number of Windows. Therefore, fine-grained image classification is a more challenging task than (generic) image classification.

Bilinear model is a fine-grained image classification model using two parallel CNN models. This CNN model uses AlexNet or VGGNet without the final full connection layer and Softmax layer, which is used as the feature extractor, and then uses SVM as the final linear classifier.

Bilinear convolution neural network (BCNN) can implement simulation to local pairwise feature interactions in a translation-invariant manner and is suitable for fine-grained classification. Generalize multiple sequentially independent feature descriptors, such as Fisher vector, Vlad, and O2P. The bilinear form simplifies the gradient calculation and enables end-to-end training of the two networks with only image labels.

![Fig. 9 Principle of BCNN](image)

5.3. Image Training and Analysis

This paper put both positive and negative samples into the model for training. In order to predict the likelihood of correct classification, this paper first use the image classification model of BCNN with Focal Loss. Our image classification model achieves an average accuracy of 65.56% on the verification set. This precision is not particularly ideal. Moreover, this paper find that our image classification model had a better recognition rate for negative samples, with a true positive rate of 5.56% and a true negative rate of 98.7%, indicating that our classifier was inclined toward identifying negative samples.

When this paper combines the probability of image classification with time, longitude and latitude, and use logistic regression model for training and verification, the classification effect is improved to a large extent. Although the logistic regression model still has the problem of classification bias. The true positive rate was 15%, the true negative rate was 98.7%, but our average accuracy went up to 75%. This paper believe that the problem of classification bias is inevitable. This is due to the large difference in the number of positive and negative samples. Our logistic regression model has a good accuracy for negative samples after the training of such scale data. In order to predict the likelihood of a mistaken classification, this model is corroborated farely valid.

5.4. Renewal Frequency Selection

As for the selection of update frequency, this paper mainly combines the trend of the number of reports of all cases over time. First, this paper carry out quantitative analysis of reports over time and find the period when the number of reports is the largest and the speed of surge is the fastest. This paper found that June, July, August and September are the months when the possibility of occurrence
of Vespa mandarinia is higher. In addition, according to their living habits, the scale of their nests and number of workers peaked in August. According to Markov’s hypothesis, the current state is determined only by the previous state, so it is reasonable to believe that there is a risk of increased spread of Vespa mandarinia in August and the adjacent months following. This conclusion is consistent with the results of our quantitative above.

Fig. 10 Reported number with time

To sum up, this paper set June, July, August and September as high-risk months every year, and other months as low-risk months. In high-risk months, this paper should update our model as frequent as possible in combination with newly reported cases, and in low-risk months, the model update frequency could be appropriately lower.

6. Sensitivity Analysis
(1) External Factors
This paper removed the prediction part of using other kind of wasps. This paper predicted its spread with only Gray Model and the result was not realistic. This paper can see that they would spread to the mountains which is obviously against its living habits. It is mainly because of the shortage of data and only with more data can me get a more accuracy result. this paper input the images provided to the BCNN classifier. After reducing the number of positive samples to the classifier, this paper find that the accuracy of recognition decreases obviously due to the over fitting of positive samples. In contrast, reducing the number of negative samples also lead to the decrease of accuracy in negative samples.
(2) External Factors
This paper reduces the update frequency of the model between June and September, then this paper found that this paper couldn’t predict the number of Vespa mandarinia as accurate as before. It is due to the habits of Vespa mandarinia, they will increase rapidly from June to September. So the changes of its number can’t be reported precisely.

This paper increases the control level of local government, then this paper figure out that the spread and trend of increase of Vespa mandarinia would be greatly restricted. This could be explained by our enhanced Logistic Function in Model 5. And if the government lower the control level or even do nothing about it, Vespa mandarinia will spread and reproduce at a high speed.

7. Conclusion

The Hidden Markov Model helps us to use other bee populations with similar propagation modes and large amount of data to act as a leading model of the propagation of Vespa mandarinia, which to some extent, alleviate the over-fitting problem caused by too few sample numbers of Vespa mandarinia. •

This paper combines the predicted coordinates obtained by GM (1,1) and the predicted coordinates obtained by HMM through entropy weight method, so as to absorb the environmental factors related to bee colony propagation as much as possible, and at the same time alleviate the over-fitting problem. •

This paper combines the results generated by the image classifier (BCNN+Focal Loss) with the time and space information to participate in the training of logistic regression model, which greatly improved the accuracy of the classification model.

Due to the potential severe impact on local honeybee populations, the presence of Vespa mandarinia can cause a good deal of anxiety. Because of their subterranean nesting habit, locating the nests of Asian giant hornets can be very difficult. Besides, they were only discovered a few months ago, official news about them is scarce, although there will likely be an effort to find and eradicate them from North America before they spread too far.

References

[1] Gengping Zhu, Javier Gutierrez Illan, Chris Looney&David W. Crowde. (2020). Assessing the ecological niche and invasion potential of the Asian giant hornet. PNAS, from https://doi.org/10.1073/pnas.20114411171
[2] Alberto J Alaniz, Javier Gutierrez Illan, Mario A Carvajal& Pablo M Vergara. (2020). Giants are coming? Predicting the potential spread and impacts of the giant Asian hornet (Vespula mandarinia, Hymenoptera:Vespidae) in the USA. Pest Management Science, from https://doi.org/10.1002/ps.6063
[3] L. Rabiner & B. Juang . (1986). An introduction to hidden Markov modelsEEE ASSP Magazine, from doi:10.1109/MASSP.1986.1165342
[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollar . (2017). Focal Loss for Dense Object Detection Proceedings of the IEEE International Conference on Computer Vision (ICCV), from https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html
[5] Joseph Berkson M.D. (2012). Application of the Logistic Function to BioAssayJournal of the American Statistical Association, from https://doi.org/10.1080/01621459.1944.10500699
[6] Tsung-Yu Lin, Aruni RoyChowdhury, Subhransu Maji. (2015). Bilinear CNN Models for Fine-grained Visual Recognition Proceedings of the IEEE International Conference on Computer Vision (ICCV), from https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Lin_Bilinear_CNN_Models_ICCV_2015_paper.html