Vespa Mandarinia Recognition and Prediction Strategy Based on Geographic Location and Image Recognition

Fangzhou Liu*, Hao Zhu, Xianghong Cheng
School of Information, Beijing Forestry University, Beijing,100083

*Corresponding author e-mail: liufangzhou@bjfu.edu.cn

Abstract. The paper conduct data analysis and information mining from the following four aspects: the correlation between geographic location and the spread of Vespa mandarinia, image recognition with the help of convolutional neural networks, species richness prediction and accurate recognition optimization, Vespa mandarinia reproduction law and reproduction characteristics, and then propose a reliable prediction model of Vespa mandarinia spread range and a model to predict the possibility of misclassification. First, consider whether the spread of pests in a period of time can be predicted. We arrange the 14 Positive IDs in chronological order and use a time series model to predict the spread range, quantify the results as longitude and latitude, and compare the results. The optimal degree is calculated. The results show that this model can accurately predict the staged emergence of Vespa mandarinias. Next, we select 1/6 high-resolution images based on the provided data set files and image files, based on the SNP selection algorithm of the $\chi^2$ test, and perform data labeling preprocessing to create a model that can predict the likelihood of a mistaken classification. Further borrow the geographic location of model one and the image recognition of model two, and use both as factors influencing the possibility of whether the newly submitted report is confirmed as a Vespa mandarinia. With the help of the Euclidean distance within the bounds of the rectangular frame, it is compared with the distance between two adjacent points to determine the impact factor 1, which is conducive to predicting the geographical distribution of the Vespa mandarinia through time series analysis. With the help of the image recognition model, the output of the convolutional neural network is used as the second impact factor, which is conducive to high-performance identification of Vespa mandarinias through real pictures. Further analysis shows that the weight ratio of the two impact factors is 3:7. Finally, we analyzed the Vespa mandarinia’s cycle reproduction law, combined with the possible genetic variation and species evolution within the population, and concluded that if there are other new reports, they should be added to the training set to rebuild the model. training. At the same time, the possibility of the elimination of this pest in Washington State is discussed. Through the statistical historical data of the monthly detection report, and further combining with time prediction, until the forecast report stabilizes to 0 within a certain period, or in a certain period If reports are submitted within the cycle, but all non-Vespa mandarinias are judged, it can be deemed that Washington State has eliminated this pest.
Keywords: ARIMA, Faster-RCNN, Euclidean distance, Object detection and image recognition

1. Introduction

According to verification by the Washington State Department of Agriculture (WSDA), this killer bee first appeared in Washington State in December 2019. It is not known how the Vespa mandarinia first arrived in North America, but insects often travel to other countries by international ships and other transportation [1].

So far, the trace of this Vespa mandarinia is limited to the Pacific Northwest. In terms of size, an adult Vespa mandarinia is 2-5 times that of an ordinary bee, with a body length of 4-8 cm. As the largest Vespa mandarinia in the world, it is also one of the most dangerous insects. In addition to its huge size, the Vespa mandarinia is also ferocious. A giant stinger can be easily pierced even in the clothes of a regular beekeeper. The stinger contains a lot of neurotoxins. If you sting repeatedly or have allergic reactions, your life may be threatened. According to scientists According to the introduction, once a group attack is triggered, the toxicity that victims are exposed to is equivalent to that of a poisonous snake [2]. Biologists are even more worried that Vespa mandarinia will destroy local bees, which will have an impact on local agriculture, because within a few hours, Vespa mandarinia can destroy a bee colony, and some Vespa mandarinia will bring some bees back to their nests for feeding Future generations, obviously, if this situation is not controlled, it will have a serious impact on food production in the future.

1) Based on the geographic location, predict the activity range of the Vespa mandarinia from the longitude and latitude, so as to accurately identify the text report data in space. (2) According to the important features of the picture, use the Faster-RCNN [3] model to identify bee species. (3) Analyze the reproduction law and reproduction trend of Vespa mandarinias, and further predict the changes in the number of Vespa mandarinias.

In order to complete the above three tasks, our specific work is as follows: 1) Analyze the geographic location of Vespa mandarinias, use Faster-RCNN modeling for image recognition and prediction, and study the reproduction law and propagation characteristics of this species. 1) Perform in-depth analysis and optimization of the model, and propose a Vespa mandarinia identification strategy [4]. Establish a prediction system based on ARIMA time series and FasterRCNN and choose an appropriate weight ratio to predict the Vespa mandarinia's outburst pattern. Analyze whether the current model has long-term reliability and influence. At the same time, it will explore the possibility of eliminating this pest in Washington State [5]. It will calculate the historical data of the reports detected in the month and predict the time point when the number of reports will be 0.

2. Model 1-Time series forecasting model based on ARIMA model

2.1. Data preprocessing

Firstly, We perform statistics on the 2021MCM_ProblemC_DataSet.xlsx data set according to the type in Lab Status. The following is the statistical graph:
Fig. 1 The count of each category of Lab Status

Statistics found that there are actually only 14 Positive IDs, 2069 Negative IDs, 2342 Unverified and 15 Unprocessed, which means that most of the reports are not Vespa mandarinia. Next, Visualize the latitude and longitude location of each report:

Fig. 2 The Location of The Longitude and Latitude of The Report

Observation found: The sample Latitude range of the Vespa mandarinia that was verified as real: [48.7775,49.1494] Longitude range:[-123.9431,-122.4186]. Therefore, in all the reports that have been submitted, it is proved that the latitude and longitude range of the report of the Vespa mandarinia is very small. Thirdly, According to the first question, because we need to predict the spread of the Vespa mandarinia over a period of time, our group discussed the feasibility of using time series models for modeling. There is a time field in the data set, and the propagation range can be represented by latitude and longitude, so it meets the modeling data needs of the time series model. The last used fields in the 2021MCMProblemC_DataSet data set are Detection Date, Lab Status, Latitude and Longitude. Below is a visualization of our data preprocessing process.
Step1: First of all, we have determined to use Python for data analysis, and first turn the .xlsx format file into a .csv format file for easy operation.

Step2: Use Python's read_csv function to read in the file and generate the DataFrame format by default.

Step3: Take out the row of "Positive ID" in "Lab Status" to generate another DataFrame.

Step4: Remove the rows with missing values.

Step5: Remove the unnecessary columns, leaving only the positive ID data in 14 rows and 8 columns.

2.2. The establishment of model
Since it is based on the time forecast range, the model of the time series is indispensable, and the most classic model is the ARIMA model. The ARIMA (p, d, q) model can be expressed as:

\[
(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d X_t = (1 + \sum_{i=1}^{q} \theta_i L^i)\varepsilon_t
\]

2.3. The process of building model
To predict the propagation range, we use the time series model to predict the latitude and longitude respectively, and then combine the predicted longitude and latitude to calculate the latitude and longitude interval and draw the range. Firstly, according to the results of 3.1 data preprocessing, draw the data into a time series graph:

3. Model verification
Longitude's ADF stationarity test after the difference: The value of Test-Statistic -9.7002<4.3783 is within the critical value, <1%, so the null hypothesis is rejected, and the value of Pvalue 1.07842e-16 is very small, so this The Longitude time series after time difference is stationary, and can be used for
ARIMA order determination and modeling. Thirdly, the order of p and q is determined by ACF and PACF graphs.

From the above figure, we can see the predicted situation and the fitting situation. It is predicted that the latitude and longitude of the Vespa mandarinia in the next two months are \([48.95783361, -122.3731313]\) and \([48.94506292, -122.26846447]\). Finally, summarize the established model: Time series forecasting model based on ARIMA Model.
The possibility of Vespa mandarinia in the geographic location of the new record can be predicted. It may be possible to set the new record point as k, and the two closest to k. The known points are i and j. Take the line connecting i and j as the diagonal to make a rectangle, as shown in the figure below:

### 4. Result analysis

#### 4.1. Model structure

The paper determined to use the ARIMA model as the basic model for modeling, use methods such as ADF, Difference, ACF, and PACF to find the most suitable p, d, and q values. After the model is built, the preprocessed data is substituted into the model for prediction, and draw the fitting and subsequent prediction graphs.

Calculation of prediction accuracy, model-sensitivity analysis: For fourteen sets of Positive data, it may be good to set the original latitude vector as origin, predict the latitude vector through the time series, and use the sum of the squares of the differences between the two vectors as the error standard for analysis:

\[
\text{Error } \_ \text{Latitude} = \sum_{i=1}^{14} (\text{origin}_i - \text{predict}_i)^2
\]  

As shown in the figure below, it is the geographic distribution trend of the Vespa mandarinia in the 30 months after September 2020. Assuming that the predicted latitude and longitude of the i-th month
and (i+1)-th month are marked as shown in the figure, it can be predicted that the location where the Vespa mandarinia appears within this time range is within the green rectangle.

![Figure 11 Predicted spread range](image)

**Fig. 11** Predicted spread range

Year x-Month i represents the predicted spatial distribution position of the Vespa mandarinia in month x year and month i, and similarly (i+1) is the distribution position label for the next month. According to the time series model, the latitude range in the next thirty months is [48.600, 48.970], and the latitude range is [-122.477, -120.070].

In order to improve the accuracy and reliability of the model, we assume that the original sample is composed of a part of non-Vespa mandarinia bee species with more distinctive characteristics. Therefore, in this work, combined with the requirements of the title, we improved the classic correct predictive analysis algorithm, by adding the statistical modeling step of the wrong sample, and using the labeled image/jpg data on the basis of the classic model. In order to reduce the variable space, after strict quality control processing on the image data, we selected 1/6 clear images in the given data set based on the $\chi^2$-like SNP selection algorithm. The results show that this model provides better results than other methods, and the additional misclassification modeling greatly improves the classification accuracy and AUC.

4.2. **Model analysis**

Latitude refers to the latitude of a certain point, and Longitude refers to the longitude of a certain point. (1)<(2) Point k is out of the rectangle (1)=(2) Point k is on the rectangle (1)<(2) Point k is within the rectangle If the condition (1)<=(2) is met, that is, the longitude and latitude of the new record are within the rectangular frame formed by two adjacent points, then the point may exist Vespa mandarinia according to Model One.
If this record is provided with a picture at the same time, the labeling process is input into model 2 for testing, assuming that the positive probability is 2 Positive. Approximate processing by neural network algorithm, with pictures and geographic location as input, and positive and negative possibilities as output. The image and geographic location of the input layer each represent a neuron, each neuron represents a feature, and the number of output layers represents the number of classification labels. Here, two classifications are used, and the softmax classifier is used, and the output layer neurons The number is two, and the number of hidden layers and hidden layer neurons are both set to 2.

**Fig. 13** Ecological timing of Vespa mandarinia showing periods of adult activity outside of the nest by caste and colony cycle throughout the year. Vertical lines in the colony cycle show the first (left-most) and last (right-most) observation dates, with the duration of stages in between.

We believe that when a user uploads a new photo of a Vespa mandarinia or an outdoor camera captures a photo of a new Vespa mandarinia, the photo should be added to our training data set, and a label should be added to this picture, and this picture should be added to the training Set to update the model. From the information provided: Throughout spring and summer, the colony grows slowly, until
it reaches a peak of about 100 worker bees in August. The queen began to produce males and females in September. Males and females leave the nest in October and early November to mate. Under normal circumstances, the life cycle of the giant Vespa mandarinia is about 45-120 days (except for the queen bee). In this topic, we first merge the two data tables according to the Global ID, and perform data cleaning. Then we select high-quality images and label tags to generate xml files to make a VOC-2007 data format suitable for Faster-RCNN training. The training data generates weights, compares the weights, and then selects the weight with the best accuracy. Use this weight to test the pictures in the test set, and finally obtain the average recognition rate of Positive and Negative. The model can be further optimized, especially the increase in the amount of positive image data will also increase its recognition accuracy.

Every fall, the queen bee chooses a suitable place to spend the winter after mating, and most of the members of the original colony will die due to cold, hunger and other reasons. The queen bee wakes up in the spring of the following year, re-establishes the colony, and the bee colony re-emerges again. Due to the poor digestive ability of the Vespa mandarinia, its energy source is mainly the white liquid secreted by the larvae after digesting food. Like other social Vespa mandarinas, the big Vespa mandarinia is a new annual species.

![Fig. 14 Time series plot of the number of reports per month](image)

We statistically organize the historical data of the number of monthly reports from 2020.12 to 2020.12, and substitute the sorted data into the model of the first question to fit the original curve and predict the number of reports for each month in the next year. Generate a prediction curve. When the predicted curve approaches 0, the number of positives is 0 and the time lasts for about a year, we can think that Washington State has eliminated the Vespa mandarinia at this time.

5. Conclusion
Interpret the original data through data visualization technology, and present the results intuitively and concisely. Through the analysis of ARIMA and Faster-RCNN topic models, a plan for predicting the Vespa mandarinia can be given, which is helpful for remote monitoring. Our predictive species category system has wide applicability. As long as the data set is provided for training, the detection and identification of similar species can be realized.
References
[1] Shaoqing Ren* Kaiming He Ross Girshick Jian Sun,” Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, in Microsoft Research, 2018
[2] Sas Institute, “The ARIMA Procedure”, SAS OnlineDoc : Version 8
[3] Brissa-Lizbeth Romero-Rosales1 , Jose-Gerardo Tamez-Pena1 , Humberto Nicolini2 , MariaGuadalupe Moreno-Treviño3 , Victor Trevino,” Improving predictive models for Alzheimer's disease using GW AS data by incorporating misclassified samples modeling”, in PLOS ONE, 2020
[4] U.S. Department of Agriculture (USDA),” New Pest Response Guidelines Vespa mandarinia Asian giant hornet”, United States Department of Agriculture, 2020
[5] Robert H. Lyles and Ji Lin,” Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting”, Statistics in Medicine, 2010.