Research on the Migration Law of Hornet Based on Machine Learning

Yunkai Li*
International School, Beijing University of Posts and Telecommunications, Beijing, China

Abstract: Bumblebees prey on European honeybees, invade and destroy their nests, severely disrupting the local ecological balance and economic stability. In order to study the specific impact of the bumblebee, this paper establishes a gray Gaussian prediction verification model based on machine learning methods, which is used to predict the flight direction of the wasp through gray correlation analysis, and the Gaussian process regression model is used to determine the flight range; Through the discussion of negative ID, the Yolo model is used for feature extraction, the resnet101 model is used for classification prediction, and then a more summarized dimension is selected to find the corresponding bee species.

1. Introduction

The Chinese hornet is the largest wasp in China. In autumn, in order to prepare food for hibernation, Chinese tiger-headed wasps often go out on a large scale. It is very radical and insists on the principle of "If you don't provoke me, I won't sting you". An Asian hornet colony was discovered near Vancouver Island in September 2019, and many local beehives were destroyed. This hornet preys on European honeybees, invades and destroys their nests, severely disrupting the local ecological balance and economic stability. People have taken pictures and reported suspected wasps to the government, but the data collected is mixed. To this end, we need to analyze and compare data and carry out actual research on the specific migration of bumblebees.

Ding Wei guang[1] proposed the automatic detection technology of pests and diseases based on deep learning. They realized the purpose of real-time monitoring of the number of pests in the field by identifying and calculating the pest images captured in the field. Once the number reaches the threshold, the system will issue an early warning in time threshold and issue an early warning in time. Zhang Wen yi[2] studied the effects of three meteorological factors, total evaporation, average temperature and minimum temperature, on the growth of pine caterpillars. They selected three models, namely, multi-layer feedforward network, generalized regression network and support vector machine. It is used to predict the occurrence of pine caterpillars, and SVM performs best. This article provides this article with a way of using deep learning to predict crop diseases and insect pests. Theis and Bethge[3] proposed a multi-dimensional LSTM model that can be used for image modeling. This model can be extended to images of any size and is superior to traditional image processing models in terms of performance and computational complexity. Gao Haoyuan et al. [4] proposed an mQA model in their article, which can answer questions about image content.

Based on this, this article uses machine learning methods to study the migration of wasps. Through this method, we can better predict and analyze the migration of wasps.

2. Grey Gaussian Prediction and Verification Model of Wasp Moving Direction

According to the classification and processing of the photographic information of the bees taken, this
article has launched the research on the movement law of the wasps. That is, whether the pest will spread over time. In this paper, the gray Gaussian regression model combining the gray correlation analysis model and the Gaussian regression model is used to verify the actual direction of the wasp.

2.1. Data Processing
According to the given data set, there are more than 3000 pictures and videos taken by witnesses, as well as two excel tables. In this part, our primary work is to classify and extract all kinds of data, which will bring great help for solving the following problems.

By traversing the bees photographed by eyewitnesses given in the title, we found that some eyewitnesses provided videos, and some provided PDF files. There are three docx files, one JFIF file, 82 MOV files, 10 MP4 files, 8 PDF files and 84 PNG files. Therefore, firstly we need to transform the format of these data that does not meet the image format, and finally, we will unify all the data into JPG format.

Firstly, for MOV and MP4 files, we use OpenCV in Python for screenshots. The first step is to store the area needed in the captured image into the matrix. The second step is to convert the data just stored in the matrix into an image. Only for the image header line, no data storage space is allocated, and the stored image pointer is returned. For others, just simply transform the format.

Secondly, for the data in the excel tables, we plan to use several methods of data screening to accomplish the tasks. Observing the data in the tables, it is obvious that a lot of data are missing and untrue in the whole table. For example, the date in some data is wrong, and some data does not write the date. As a result, list deletion, average replacement and multiple interpolation are needed. Multiple imputation makes up for the defect of single imputation. It does not try to estimate each missing value by simulation value, but proposes a random sample of the missing data value (these samples can be a combination of different model fitting results). Implementing this program properly reflects the uncertainty caused by missing values, which makes statistical inference effective. Multiple imputation can be divided into the following three steps: generating a set of possible imputation values for each missing value, which reflects the uncertainty of the unresponsive model; each interpolation data set is analyzed by the statistical method for the complete data set; the results from each interpolation data set are selected according to the scoring function to generate the final interpolation value.

After data processing, we obtain more than 3000 JPG pictures and complete data tables.

2.2. Grey Relation Analysis
As for the factors between two systems, the measurement of their correlation varying with time or different objects is called correlation degree. In the process of system development, if the changing trend of the two factors is consistent, that is, the higher the degree of simultaneous change, the higher the correlation between the two factors; on the contrary, it is lower. Therefore, the grey correlation analysis method is based on the similarity or dissimilarity degree of the development trend between factors, namely "grey correlation degree", as a method to measure the correlation degree between factors.

The Grey System Theory puts forward the concept of grey correlation analysis for each subsystem, which aims to seek the numerical relationship between each subsystem (or factor) in the system through certain methods. Therefore, the grey relational analysis provides a quantitative measure for the development and change of a system, which is very suitable for dynamic process analysis.

This problem just meets the requirements of grey correlation analysis. We are required to predict whether this hornet will spread over time. Therefore, according to the given data set, 14 bees photographed by the eyewitness are identified as positive, that is to say, these 14 bees are what we want to analyze. Therefore, we first extracted the specific information of the 14 bees, including the specific time and place where they were found.

Step 1: determine the analysis sequence.
The reference sequence reflecting the system behavior characteristics and the comparison sequence
influencing the system behavior is determined. The data sequence reflecting the behavior characteristics of the system is called reference sequence. The data sequence composed of factors that affect the system behavior is called comparison sequence.

The reference sequence is:

\[ Y = Y(k) \mid k = 1, 2, \ldots, n \]  

The comparison sequence (also called subsequence) is:

\[ X_i = X_i(k) \mid k = 1, 2, \ldots, n, i = 1, 2, \ldots, m \]  

Step 2: dimensionless variable.

Because the data in each factor column of the system may be different in dimension, it is not easy to compare, or it is not easy to get the correct conclusion. Therefore, in the gray correlation analysis, the data are generally dimensionless. There are mainly two methods:

Initial value processing:

\[ X_i(k) =, k = 1, 2, \ldots, n; i = 0, 1, 2, \ldots, m \]  

Average processing:

\[ X_i(k) = \frac{X_i(k)}{X_i}, k = 1, 2, \ldots, n; i = 0, 1, 2, \ldots, m \]  

Where \( k \) corresponds to the time period, \( i \) corresponds to a row (i.e. a feature) in the comparison sequence.

Step 3: Calculation of correlation coefficient

\[ \xi(k) = \frac{\min_{i,j} \min \{y(k) - x_i(k)\} + \max_{i,j} \max \{y(k) - x_i(k)\}}{\max_{i,j} \max \{y(k) - x_i(k)\} - \min_{i,j} \min \{y(k) - x_i(k)\}} \]  

\[ \tilde{\xi}(k) = \frac{\min_{i,j} \Delta_i(k) + \max_{i,j} \max \Delta_i(k)}{\Delta_i(k) + \max_{i,j} \max \Delta_i(k)} \]  

where \( \rho \in (0, \infty) \), it is called resolution coefficient.

Generally, the value range of \( \rho \) is \((0, 1)\), and the specific value depends on the situation. When \( \rho \leq 0.5463 \), the resolution is the best, usually \( \rho = 0.5 \).

Step 4: calculation of the correlation degree

As the correlation coefficient is the correlation degree value of the comparison sequence and the reference sequence at each time (that is, each point in the curve), it has more than one number, and the information is too scattered to facilitate the overall comparison. Therefore, it is necessary to concentrate the correlation coefficient of each time (i.e. each point in the curve) into one value, that is, to find its average value, which is used as the quantitative expression of the correlation degree between the comparison sequence and the reference sequence.

\[ r_i = n \sum_{k=1}^{n} \xi_i(k), k = 1, 2, \Lambda, n \]  

where \( \rho \in (0, \infty) \), it is called resolution coefficient.

Generally, the value range of \( \rho \) is \((0, 1)\), and the specific value depends on the situation. When \( \rho \leq 0.5463 \), the resolution is the best, usually \( \rho = 0.5 \).

Step 4: calculation of the correlation degree

As the correlation coefficient is the correlation degree value of the comparison sequence and the reference sequence at each time (that is, each point in the curve), it has more than one number, and the information is too scattered to facilitate the overall comparison. Therefore, it is necessary to concentrate the correlation coefficient of each time (i.e. each point in the curve) into one value, that is, to find its average value, which is used as the quantitative expression of the correlation degree between the comparison sequence and the reference sequence.

\[ r_i = n \sum_{k=1}^{n} \xi_i(k), k = 1, 2, \Lambda, n \]  

Step 5: the ranking of relevance degree

If \( r_1 < r_2 \), then the reference sequence \( y \) is more similar to the comparison sequence \( x_2 \). After calculating the correlation coefficient between \( X_i(k) \) and \( Y(k) \), the average value of various correlation coefficients is calculated. The average value is called the correlation degree between \( Y(k) \) and \( X_i(k) \). Finally, according to the data, we use Python to get the visual graph as shown in the following figure.
Fig. 1 Analysis and forecasting the location of hornets in GM

The blue dots in the figure indicate the location and time of the hornet. In the figure, the exact time and location of the 10 hornets are marked. According to the grey correlation analysis model, we can get the remaining blue dots. From the graph, we can see that with the passage of time, the hornet will move towards the place where the latitude and longitude increase, that is to say, the hornets will move towards the northeast.

2.3. Gaussian process regression model

Gaussian process regression is a non-parametric model that uses the Gaussian process before regression analysis of data. The model assumption of GPR includes regression residual and Gaussian process prior, and its solution is based on Bayesian inference. If kernel function is not restricted, GPR is the universal approximation of any continuous function in compact space theoretically. In addition, GPR can provide a posteriori of prediction results, and the posteriori has an analytical form when the likelihood is a normal distribution. Therefore, GPR is a universal and analytic probability model.

In this part, according to the previous step of grey correlation analysis, we get the hypothesis that the hornet moves to the northeast. Then, according to the Gaussian regression model, we verify the hypothesis.

Through Gaussian regression analysis, we will get a prediction range. If the scope just includes the results obtained by grey correlation analysis, the hypothesis can be verified.

Let the random vector

\[ X = [x_1, x_2, \ldots, x_n] \]  

obey the multivariate Gaussian distribution

\[ X \sim N(\mu, \Sigma) \]  

where

\[ X_1 = [x_1, x_2, \ldots, x_m] \]  

is the observed variable and

\[ X_2 = [x_{m+1}, x_{m+2}, \ldots, x_n] \]  

is the unknown variable, then:
\[ X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \] (13)

So there is:

\[ \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \] (14)

\[ \Sigma = \begin{bmatrix} \sum_{11} & \sum_{12} \\ \sum_{21} & \sum_{22} \end{bmatrix} \] (15)

Given \( x_1 \), find the posterior distribution of \( x_2 \)

\[ \mu_{2|1} = \mu_2 + \sum_{21} \sum_{11}^{-1} (X_1 - \mu_1) \] (16)

\[ \sum_{2|1} = \sum_{22}^{-1} - \sum_{21} \sum_{11}^{-1} \sum_{12} \] (17)

When we get a batch of observation samples on some indices, this batch of observation samples will help us estimate the distribution of samples on other indices (a posteriori). We set the observed index set as \( X_1 \) and the unobserved index set as \( X_2 \). Next, we can use the method in the second section to obtain the posterior probability parameters \( \mu_{2|1} \) and \( \Sigma_{2|1} \) of the sample distribution on \( X_2 \), and finally resample the random variables on \( X_2 \).

The following is the result of Gaussian regression.

---

**Fig.2 Gaussian Process Regression**

By observing the resulting graph of Gaussian process regression from different angles, we can see two green planes and one blue plane. Two green planes are the theoretical range, and the blue plane is the actual range. The blue plane falls between the green planes, which shows that the previous hypothesis is true.

In conclusion, over time, hornets moved northeast (the direction of inland). The final plot is the following one.

**3. YOLO and Resnet101 Model**

When we analyze the data, it is easy to find that only 14 of the pictures taken by all the witnesses are identified as real hornets, and a large number of pictures are identified as negative, which also means
that many wrong eyewitness events have occurred. Many witnesses identified some other insects that looked like hornets as them. Therefore, in this section, we are needed to discuss a model that predicts the likelihood of mistaken classification.

3.1. YOLO Model
The YOLO network structure consists of five parts: input layer, convolution layer, pooling layer, fully connected layer and output layer. The output layer of YOLO is Tensor, and the YOLO recognition network consists of 24 convolutional layers and two fully connected layers. As a typical layer of the neural network, the convolutional layer is used to perform convolution operations on the input fastener photos, and convolve the photo information and the convolution kernel to achieve the function of feature extraction[5].

In the experiment, the pascal voc dataset image is used as the YOLO training sample because YOLO is an end-to-end network, and the entire image can be directly used for training. The whole image is divided into SxS grids. When the center of an object falls in each grid, B detection bounding boxes and the confidence of each box will be predicted. Each detection bounding box contains the width and height of the bounding box, the position of the center of the bounding box relative to its parent network, C category probabilities, and the final output layer outputs a tensor of S×S (B×5+C) dimensions.

The specific calculation process of the YOLO Model is as follows.
  a. Image segmentation
  We divide the input image into 7×7 grids.
  b. Bounding Boxes Prediction
  The segmented image is processed in two ways,
  1)First, look at bounding boxes + confidence. In this step, Yolo gives two prediction boxes for each grid, which is a bit like an anchor of fast RCNN, but not exactly the same. The prediction box given by Yolo is based on the center point of the grid, and its size is customized. Each grid predicts B bounding boxes, and each bounding box has four coordinates and confidence, so the final prediction result is S×S×(B×5+C) vectors. In the original text, B = 2 means there are two pre-selection boxes, C = 20, which means there are 20 categories, S=7.
  2)Let's look at the second-class probability map. In fact, this work is carried out together with the previous one. It is responsible for the categories of grids, and the predicted results are also included in the final 7×7×30 results.
  c. Image extraction
  We extract the images with confidence greater than 0.5 and intercept them as shown in the figure below(a)(b). Then we take out the part of bees. The purpose is to carry out a data cleaning, so that the bees are not obvious picture is cleaned. Then the screenshot, so that the characteristics are more distinctive, the target bees become larger, remove irrelevant information for training. As some pictures(c) are too vague to process, we decided to abandon them.

  ![Fig.3 Process of images selection](image)

3.2. Resnet Classification Network and Fine-tune
  a. Loss function
  The network structure of Yolo is very simple. In the front is a fully convoluted network that has been pre-trained. In the back, four convolution layers are added. Finally, two fully connected networks
are added.

And our loss function is Cross-Entropy loss function. Entropy is an index used to measure the amount of information in a system, indicating the expectation of all the information, also called information entropy.

Definition of entropy:

$$H(x) = -\sum_{x \in X} p(x) \log \left( \frac{p(x)}{q(x)} \right)$$  \hspace{1cm} (18)

KL divergence is also called relative entropy, which is used to describe the difference between two probability distributions P and Q. The greater the difference between the two distributions, the greater the divergence. The formula is defined as follows:

$$D_{KL}(p || q) = \sum_{i=1}^{N} p(x_i) \log \left( \frac{p(x_i)}{q(x_i)} \right)$$

$$= H(p(x)) = \left[ -\sum_{i=1}^{N} p(x_i) \log (q(x_i)) \right]$$  \hspace{1cm} (19)

Among them, \(p(x_i)\) is the probability of the real event, which is equivalent to the label, and \(q(x_i)\) is the fitted distribution value, which is equal to the predicted value. Therefore, the first term of formula (1) becomes \(-H(p(x))\), which is the entropy of \(x\), which is constant during the optimization process, so it is not considered. The latter part is the cross-entropy loss function, \(N\) represents the number of samples, and the mathematical definition is:

$$H(p,q) = -\sum_{i=1}^{N} p(x_i) \log (q(x_i))$$  \hspace{1cm} (20)

b. Fine-tune

In this part, we send the honeybee with the cut graph to resnet101 for training, and only modify the last layer of the neural network and training parameters. The code of this part will be shown in the appendix.

c. Resnet 101

Then the image is processed by model method to get a 100-dimensional vector. We find that most of the dimensions have few values, and only six dimensions are clustered.

Finally, we extract them, draw the distribution, determine the centroid, and draw the range.

![Fig.4 Distribution of six look-like species](image)

In the above figure, the red range is the range of hornets, and the other six colors are the range of hornets that are most likely to be mistaken for Hornets. The names and habits of these bees are as follows. These six species are most likely to be mistaken for hornets.
Tab.1 Features of six look-like species

| Category            | Feature                                           |
|---------------------|---------------------------------------------------|
| European hornets    | Nets are usually six feet above the ground        |
| Eastern cicada killers | Terminal abdominal segments completely black    |
| Baldfaced hornets   | Smaller size, black and white coloration         |
| Yellow jackets      | Smaller size (up to 0.5 inches)                  |
| Aegeriidae          | Fly near the flowers                              |
| Horseflies          | Feed on the blood of mammals                     |

4. Seasonal ARIMA Model

In this section, the task we need to complete is to explain how to update additional reports over time and how often updates occur. According to the actual situation and the conclusion of the previous question, we know that both Hornets and other bees will move with the passage of time, so reports should be updated in real-time according to a certain frequency. In order to solve this problem, we are going to take advantage of Seasonal Arima Model.

![Fig.5 Distribution in the time series](image)

Tab.2 Number of Positive ID

|       | JAN | FEB | MAR | APR | MAY | JUN |
|-------|-----|-----|-----|-----|-----|-----|
| 2019  | 0   | 0   | 0   | 0   | 0   | 0   |
| 2020  | 0   | 0   | 0   | 2   | 1   |     |

|       | JUL | AUG | SEP | OCT | NOV | DEC |
|-------|-----|-----|-----|-----|-----|-----|
| 2019  | 0   | 0   | 2   | 1   | 1   | 1   |
| 2020  | 0   | 1   | 4   | 1   | 0   | 0   |

We draw the sequence diagram of the original data, and the basic trend of data can be seen from the sequence diagram.

![Fig.6 Sequence diagram](image)

![Fig.7 Season index](image)
From the sequence diagram, we can see that the sequence has an obvious trend, so it is not stable. It does not satisfy the premise of stationary time series, so it cannot build an ARMA model.

However, there is a relatively apparent seasonal trend, so it is more appropriate to use SARIMA. Approximate seasonal fluctuation index showed above.

Then we establishment of growth season model as in picture 8.

By observing and comparing the above pictures, we can know that the number of models reaches the peak in September every year, and the frequency of updating models needs to be higher from August to October. According to the previous map, when there is a red circle, we need to update the model.

5. Conclusion

In this article, we have done a lot of work in data processing, image set format conversion, and data cleaning of table data. Through the use of grey relational analysis, seasonal ARIMA, Yolo, resnet101 models for analysis and verification, the following conclusions are obtained.

1) Through the research of this article, we know that the six species of bees will mistake the wasps and the update frequency of the report, use the Gaussian process regression model to determine the migration range and conclude that the wasps will move northeast over time, it can provide a certain theoretical basis for relevant departments to take preventive measures and formulate countermeasures in a timely manner.

2) If the wasps migrate to the northeast, they will migrate to Vancouver Island and spread from south to north in Washington State. The disappearance of the wasp started in the warmer areas of southern Washington state and extended to the north and northeast. Considering the close geographical relationship between Vancouver Island and Washington State, and the location relationship is consistent with the migration direction of wasps, we can think that only when there are no wasps on Vancouver Island, the wasps disappear entirely in Washington State.

References

[1] Ding W, Taylor G. Automatic moth detection from trap images for pest management [J]. Computers and Electronics in Agriculture, 2016,123:17-28.

[2] Wen Yi Z, Tian Zhong J, Shan Chun Y, et al. Studies on prediction models of Dendrolimus superans occurrence area based on machine learning [J]. Journal of Beijing Forestry University 2017.

[3] Thies L, Bethge M. Generative Image Modeling Using Spatial LSTMs [J].2015.

[4] Gao H, Mao J, Zhou J. Are you talking to machine? Dataset and methods for multilingual image question answering [J].Computer Science, 2015:2296-2304.

[5] Zhang Jun, Zhang Ting, Yang Zheng ling, etc. Car model recognition based on deep convolutional neural network Method [J]. Sensors and Microsystems, 2016, 35(11): 19-22.