Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2103942

Response Strategies for the Biological Invasion of Hornets

Since the September 2019 incident of hornet attacks on humans in Canada, there have been reports of hornet sightings in Washington State, US. The presence of wasps like hornets can cause panic among the public because they are highly destructive to both humans and the ecosystem. The Washington State Department of Agriculture receives a large number of sighting reports from the public, and due to limited resources, only some of these reports have been verified. Therefore, we need to build a suitable model to analyze the existing data to help the government identify the sighting reports that are more likely to be correct and allocate the resources for follow-up investigation.

Two models were developed for this purpose. The prediction model was used to estimate the occurrence of bee swarms in future periods. Since the number of samples before 2019 was too small and too long ago to be meaningful for predicting 2021 of data, we only selected 2019 and 2020 years of data as the sample data. Considering that latitude and longitude are not intuitive to describe the distance between sample points, we choose to convert the latitude and longitude data into data in kilometers. All the data were sorted by quarter, and a suitable circular area was selected so that it contained the most sample points under that area size, and the number of sample points in that area was counted for each quarter. Winters' additive was used to forecast the data for the four quarters of 2021 and the predicted values in each of the four areas were counted, and the wasp spread paths could be determined based on the number of predicted values of wasps in different areas.

The classification model is used to determine the probability that the sighting report is a positive report. Since the data set contains image, text and time and location data, this model treats different categories of data separately, using SVM and MLP methods for image, plain Bayesian classification for text, and SGD for time and location data, respectively, to obtain four probability indicators after processing, and then using Logistic Regression processing to get the probability of reporting as positive report.

The Unverified sighting reports are input into the model to get their probabilities, and then ranked among the contemporaneous reports, and with limited resources, those with higher probabilities are investigated first.

Since all the methods used in the model support partial fit, if new data are available, the model weights can be updated without retraining all the data and using only the new data. Also, according to the survey, it is known that the activity of giant tiger bees is extremely seasonal, so it is only necessary to update the model once every quarter.

The Winters' additive was used to predict 2022 of data from 2019, 2020 and the four quarters of 2021, and when the predicted values of all four quarters of 2022 were zero, the data of 2021 could be used as evidence that the wasps had been eliminated in Washington State.

Keywords: Winters' additive, Naïve Bayers, SVM, TF-IDF, Stochastic Gradient Descent

Table of Contents

1 Introduction

1.1 Problem Background

Asian giant hornets are aggressive species that prey on arthropods such as bees, causing significant damage to the beekeeping industry, and also shows potential threat to human. The occurrence of the nest is an alarming of biological invasion. However, Asian giant hornets typically build their nests underground, which makes their nests highly difficult to be located.

In September 2019, a nest of Asian giant hornets was discovered and destroyed on Vancouver Island, British Columbia. Meanwhile, Asian giant hornets also invaded the United States. In December, the Washington State Department of Agriculture confirmed a dead specimen had been found in Washington. It's still unclear if the hornets are established and reproducing in North America, or how widespread they are in the Pacific Northwest. Based on the available evidence, it's likely that they are not widely established, and there may have been multiple independent introductions of the wasps. Although official news about them is scarce, the action to find and eradicate them from North America will likely to launch.

1.2 Restatement of the Problem

The habits, distribution status and official information about hornets given in the Annex are known. Combined with the detailed information available online, the questions we need to solve are as follows:

- How can we interpret the data provided by the public reports?
- What strategies can we use to prioritize these public reports for additional investigation given the limited resources of government agencies?

2 Assumptions and Notations

2.1 Assumptions

To simplify the problem, we make the following basic assumptions, each of which is properly justified.

- **Assumption 1:** *Unless there is human intervention, the number of wasps will not suddenly drop in the coming year.*
- **Assumption 2:** *Wasps will not undergo irregular mass migration.*
- **Assumption 3:** *Every reporting person is honest.*
- **Assumption 4:** *Each sighting saw a different individual insect.*

2.2 Notations

The key mathematical notations used in this paper are listed in Table 1.

Table 1 Notations used in this paper

Symbol	Description	Symbol
m	Cycle length (Quarterly data is taken as 4)	m
α	Horizontal smoothing parameter	α
β	Smoothing parameter of the trend	β
γ	Seasonal smoothing parameter	γ
$\hat{x}_{t+h}$	The predicted value of period h	$\hat{x}_{t+h}$

3 Preprocessing of Dataset

By analyzing the 2021MCM_ProblemC_Files.rar, 2021MCM ProblemC_ Images_by_GlobalID.xlsx, we present the structure of the dataset (consider two or more images in one file as a single file):

Table 2 structure of the dataset

Format	Image	Video	Text
Number of Files	3212	92	1

3.1 Image & Video

As we know the GlobalID of each report of sighting, we can divide the image & video file as two kind: a GlobalID with single file, and a GlobalID with multiple files.

3.1.1 Video

Considering the model we used cannot process video files, we extract the key frames of each video as images.

3.1.2 Image

In order to fit our model, we do thing same transform on images and the key frames from videos:

- Convert the image into grayscale.
- Resize the image to 64px by 64px.
- Normalize the pixel to the value between 0 and 1.
- Save as array.
- Decompose the components of the image by using PCA (Principal Component Analysis) and also save as array.

Decomposition PCA that extracts the components of the image that may contain hornets.

Given a set of points in Euclidean space, the first principal component corresponds to a line that averages the points through the multidimensional space, while ensuring that the sum of squares of the distances from each point to this line is minimized. After removing the first principal component, the second principal component is obtained in the same way. And so on. The singular values in Σ are all the square roots of the eigenvalues of the matrix XXT. Each eigenvalue is proportional to the variance associated with them, and the sum of all eigenvalues is equal to the sum of the squares of the mean point distances of all points to their multidimensional space.

Usually, to ensure that the first principal component describes the direction of the maximum variance, we perform principal component analysis using mean subtraction. If average subtraction is not performed, the first principal component is likely to correspond more or less to the mean of the data. In addition, to find the minimum mean square error of the approximate data, we must choose a zero mean.

Assuming a zero empirical mean, the principal component w1 of the data set X can be defined as:

$$\boldsymbol{w}_1 = \underset{\|\boldsymbol{w}\|=1}{\operatorname{argmax}} Var\{\boldsymbol{w}^T\boldsymbol{X}\} = \underset{\|\boldsymbol{w}\|=1}{\operatorname{argmax}} E\{(\boldsymbol{w}^T\boldsymbol{X})^2\}$$

3.2 Text

Notes and Lab Comments from each observation record were extracted and annotated, and the textual information in the attachment was extracted together.

3.3 Date & Coordinate

Standardize features by removing the mean and scaling to unit variance

The standard score of a sample x is calculated as:

$$z = \frac{(x - u)}{s}$$

where u is the mean of the training samples, and s is the standard deviation of the training samples.

Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Mean and standard deviation are then stored to be used on later data using transform.

For instance many elements used in the objective function of a learning algorithm assume that all features are centered around 0 and have variance in the same order. If a feature has a variance that is orders of magnitude larger that others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected.

4 Model I: Predictive Model

4.1 Data preprocessing

The data before 2019 are considered to be of low value because of the too long interval and small sample size, so they are discarded. According to the background information, hornets die as a whole in late fall or winter, but leave newly produced queens for the winter, and new queens start a new life cycle in spring when the temperature rises. Therefore, the number of samples in winter is not meaningful for predicting the future, but given the small amount of data in winter, it is not removed here.

The data given in the question is the latitude and longitude, due to the shape of the earth, the length of the longitude at each degree interval is different under different latitudes, so we

choose to take the smallest values of longitude and latitude in all sample points as the reference value, and the latitude and longitude of the rest of the points are differenced from this reference value. After this treatment, all the sample points will be distributed in the first quadrant, and it can be seen through the survey data: the difference of 111km per degree of latitude under the same longitude, and the distance of longitude per degree of latitude under different latitude conditions is shown in table.

Table 3

Latitude	Length of the meridian at one-degree intervals (per mile)
45°	84 506.138
46°	83 358.926
47°	82 191.231
48°	81 003.346
49°	79 795.563
50°	78 568.182

The scatter plot consisting of all sample points after processing (horizontal and vertical coordinates are both in km) is shown in Fig.

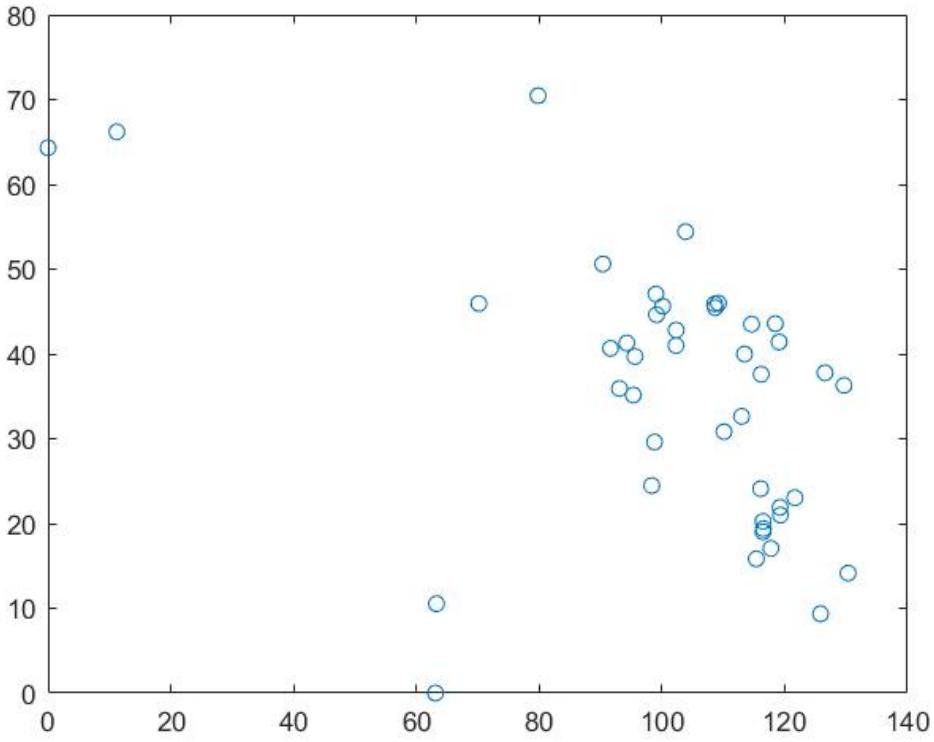


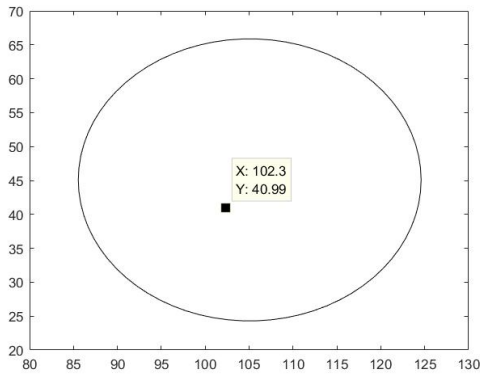
Figure 1

4.2 Division of Model Area

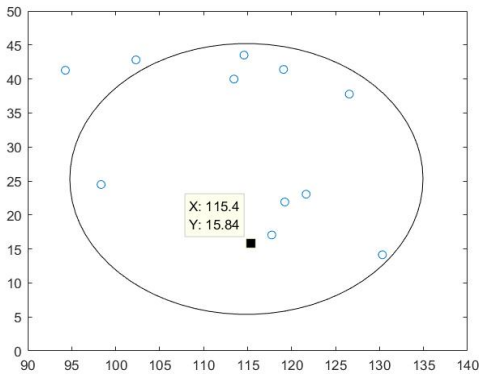
With the help of the background information given in the question, we know that hornets have strong seasonal migration characteristics, and hornets are more active in summer and

autumn. Combined with the statistics of 2019 and 2020, we choose to count the number of samples by season, thus we divide all the data into 8 categories.

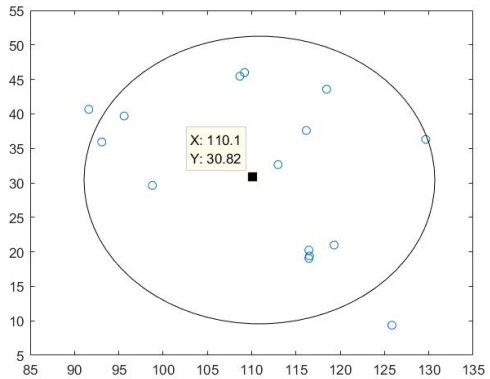
The scatter plots of the first quarter of the two years 2019 and 2020 were drawn, and the distribution of sample points in a common area was analyzed for both, and a suitable area, noted as area A, was selected, making the largest number of sample points contained in the same area. After several experiments, it was found that when the area is a circle with a radius of 20 km (the center of the circle is chosen reasonably according to the actual situation of each quarter), the number of sample points contained is higher, and the area is smaller at this time, which can achieve the goal of containing as many sample points in as small an area as possible. Achieve the original intention of containing maximum amount of sample in the minimum area. (Note: Since there are no sample points in the first quarter of 2019 and the second quarter of 2019, there is only one scatter plot for both the first quarter and the second quarter)



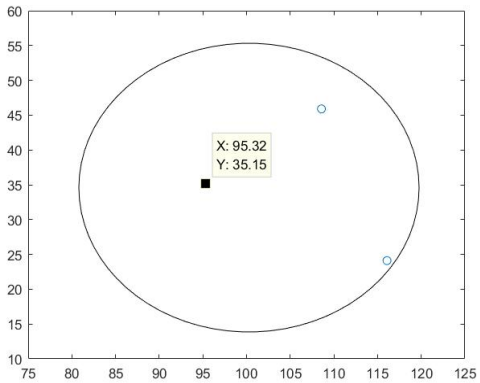
Q 1 2020



Q 2 2020



Q 3 2020



Q 4 2020

Figure 2

The number of samples within region A in each of the four quarters of 2019 and 2020 was counted, and so on, the number of samples within region B, region C, and region D. Summarize all the data to get the table shown below.

Area A	Area B	Area C	Area D

date	number	date	number	date	number	date	number
Q1 2019	0	Q1 2019	0	Q1 2019	0	Q1 2019	0
Q2 2019	0	Q2 2019	0	Q2 2019	0	Q2 2019	0
Q3 2019	1	Q3 2019	1	Q3 2019	1	Q3 2019	1
Q4 2019	2	Q4 2019	2	Q4 2019	3	Q4 2019	3
Q1 2020	1	Q1 2020	1	Q1 2020	1	Q1 2020	1
Q2 2020	6	Q2 2020	10	Q2 2020	8	Q2 2020	6
Q3 2020	10	Q3 2020	12	Q3 2020	14	Q3 2020	11
Q4 2020	3	Q4 2020	1	Q4 2020	3	Q4 2020	3

Figure 3

Table 4 Descriptive Statistics

Area	Number	Minimum	Maximum	Total	Standard Deviation	Mean Value
Area A	8	0	10	23	3.482	2.88
Area B	8	0	12	27	4.779	3.38
Area C	8	0	14	30	4.892	3.75
Area D	8	0	11	25	3.758	3.13

4.3 Model Building

4.3.1 Basis and Method of Model Construction

With the help of the background information in the topic and the visualized data in the figure above, we can see that the distribution of hornets in various places is highly seasonal, with the highest numbers in summer and autumn. Therefore, we consider using the expert modeler in the SPSS software time series model to predict the population size of hornets.

4.3.2 Model Introduction

Table 5 Model Statistics

Model	Model Type	Area	Model Fit Statistics		
			R-squared	Stationary R-squared	Normalized BI
Hornet population prediction model	Winters' additive	Area A	0.645	0.871	2.576
		Area B	0.540	0.857	3.469
		Area C	0.567	0.870	3.455
		Area D	0.581	0.870	2.893

From the above table, it can be seen that: the optimal model given by the expert modeler is Winters' additive, R-squared is around 0.6 and the model prediction is good.

Introduction of Winters' additive ：

$$\begin{cases} l_t = \alpha(x_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1}) \text{ , } (Horizontal\ smoothing\ equation) \\ b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1} \text{ , } (Trend\ smoothing\ equation) \\ s_t = \gamma(x_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m} \text{ , } (Seasonal\ smoothing\ equation) \\ \hat{x}_{t+h} = l_t + hb_t + s_{t+h-m(k+1)} \text{ , } k = \left[\frac{h-1}{m}\right] \text{ , } (Prediction\ equation) \end{cases}$$

4.3.3 Model Parameters

Table 6 Model Parameters

Area		Estimate	Standard Error	t	Significance
Area A	α	.206	.230	.894	.412
	β	1.713E-5	.114	.000	1.000
	γ	.001	.630	.002	.999
Area B	α	.207	.206	1.005	.361
	β	4.424E-7	.104	4.259E-6	1.000
	γ	3.152E-5	.561	5.623E-5	1.000
Area C	α	.205	.235	.876	.421
	β	2.572E-7	.120	2.150E-6	1.000
	γ	4.699E-5	.628	7.479E-5	1.000
Area D	α	.204	.243	.838	.440
	β	.000	.128	.001	.999
	γ	.001	.641	.002	.999

The residual plots ACF and PACF are shown below:

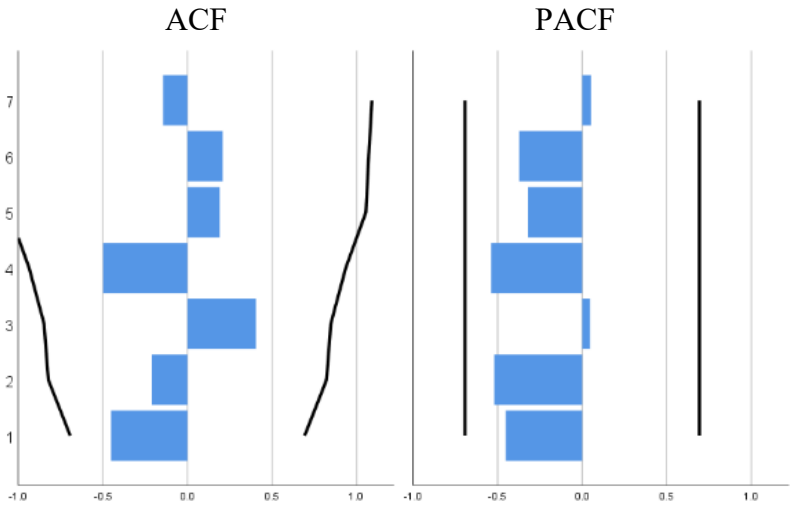


Figure 4

From the ACF and PACF graphs of the residuals, we can see that the autocorrelation coefficients and partial autocorrelation coefficients of all lag orders are not significantly different from 0. Therefore, we consider that the residuals are white noise series, and therefore Winters' additive can identify all sample point data well.

5 Model II: Classification Model

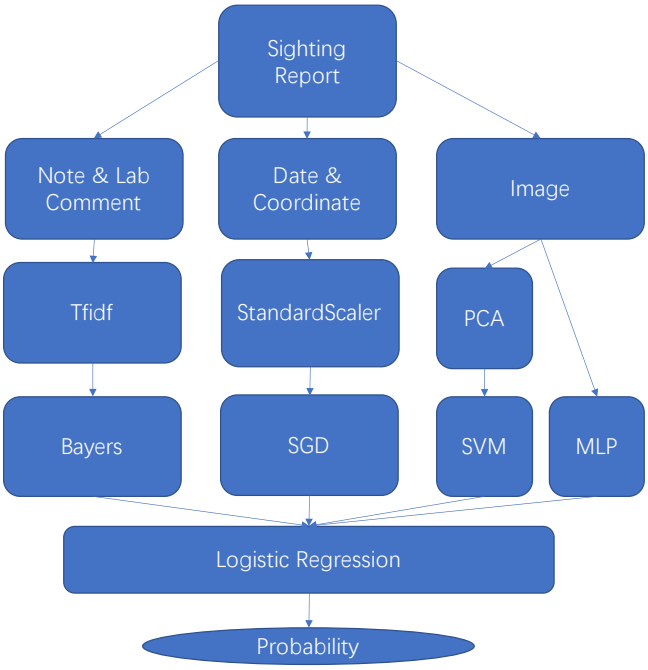


Figure 5

5.1 Model of Image

For the image data in the dataset, we think to perform two considerations. Analysis of the components of the images, and analysis of the whole images. The final two metrics are derived on the image dataset.

5.1.1 Image component classification model

To consider the principal component in the image, i.e., hornets. Use the image set decomposed by PCA. Define a Support Vector Machines, SVC (C-Support Vector Classification) to classify the image after feature extraction. Define a Model Selector, GridSearchCV (Grid Search with Cross Validation) for optimizing the hyperparameters of the support vector machines.

C-Support Vector Classification

SVC are very efficient in high-dimensional spaces and are often used in classification or regression problems and are still effective in cases where the data dimension is greater than the number of samples, so we use this method to classify images.

Definition

SVC construct a hyperplane or a series of hyperplanes in a high or infinite dimensional space that can be used for classification, regression or other tasks. Intuitively, a good segmentation is achieved by using hyperplanes to maximize the distance between the closest training data points in any class (the so-called functional margin), because a larger margin usually results in a lower generalization error of the classifier.

In both classes, given a training vector $x_i \in \mathbb{R}^p, i = 1, \ldots, n$ and a vector$y \in \{1, -1\}^n$, SVC can solve the following main problems:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \xi_i$$

subject to $y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, n$

The pairing is

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha$$

subject to $y^T = 0$

$$0 \leq a_i \leq C, i = 1, \ldots, n$$

Where e is all vectors，$C > 0$ is the upper bound，Q is an n by n semi-positive definite matrix，and $Q_{ij} \equiv y_i y_j K(x_i, x_j)$，where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel. So the training vector is indirectly reflected to a higher dimensional (infinite) space through the function ϕ.

The **decision function** is:

$$sgn\left(\sum_{i=1}^{n} y_i a_i K(x_i, x) + \rho\right)$$

Practical approach

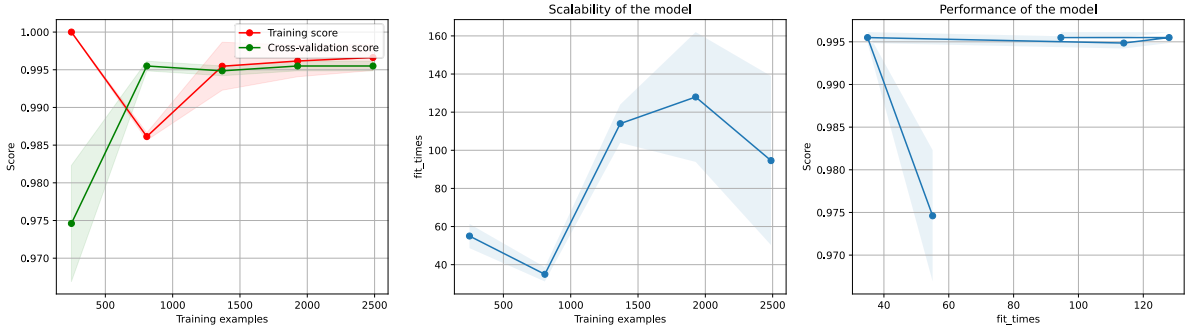


Figure 6

Use rbf as the kernel function. Since the ratio of the number of two classes in the original data is too disparate, the weight (weight) is set for the samples by class: n_samples / (n_classes * np.bincount(y)), and the predicted probability is calculated using 5-fold cross-validation.

5.1.2 Entirety Image classification model

Except for component that may contain hornet, we think the environment of picture can also reflect the possibility of the hornet occur. So we define a Neural Network, MLPClassifier (Multi-Layer Perceptron Classifier), learning features from the whole image.

Multi-Layer Perceptron Classifier

Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function $f(\cdot): R^m \rightarrow R^o$ by training on a dataset, where m is the number of dimensions for input and o is the number of dimensions for output. Given a set of features $X = x_1, x_2, \ldots, x_m$ and a target y, it can learn a non-linear function approximator for either classification or regression. It is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers. Figure 1 shows a one hidden layer MLP with scalar output.

The input layer consists of a set of neurons $\{x_i | x_1, x_2, \ldots, x_m\}$ representing the input features. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation $w_1x_1 + w_2x_2 + \ldots + w_mx_m$, followed by a non-linear activation function $g(\cdot): R \rightarrow R -$ like the hyperbolic tan function. The output layer receives the values from the last hidden layer and transforms them into output values.

MLPClassifier implements a multi-layer perceptron (MLP) algorithm that trains using Backpropagation.

Learning occurs in the perceptron by changing connection weights after each piece of data is processed, based on the amount of error in the output compared to the expected result. This is an example of supervised learning, and is carried out through backpropagation, a generalization of the least mean squares algorithm in the linear perceptron.

After fitting (training), the model can predict the labels of new samples. the MLP algorithm uses backpropagation. More precisely, it uses some form of gradient descent for training, where the gradient is computed by backpropagation. For the classification problem, it minimizes the cross-entropy loss function to give a probability estimate $P(y \mid x)$ in vector form for each sample x.

5.2 Model of Text

For the data in the text dataset, we consider that these descriptions present the shape of hornets to some extent and therefore can be considered as an indicator. The submitter's notes and lab comments contained in the text dataset are considered equivalent in terms of the information describing hornets, and thus both are used as a single dataset.

Text classification model:

Define a feature extraction TfidfVectorizer (Term Frequency times Inverse Document Frequency Vectorizer) for tokenizing the text and counting the word frequency, and finally normalizing it into a statistically usable vector. Define a statistical classifier, distribute Multinomial Naïve Bayes.

Since the sequence of symbolic text cannot be directly passed to the algorithm, it needs to be transformed into a fixed-length numerical matrix feature vector. Since simple word frequency statistics are affected by the length of the text, e.g., although it may be describing the

same topic, longer texts have a higher average number of word occurrences compared to shorter texts. Therefore, the number of occurrences of each word in each text is divided by the total number of all words in the document: a new feature frequency TF (Term Frequencies) is obtained.

The IDF (Inverse Document Frequency) is obtained by decreasing the weight of words that appear in many texts in the training set, thus highlighting the information content of words that appear in only a small portion of the documents in the training set, taking into account words that appear less frequently but may play a larger role.

Definition and Practice Methodology

Count Vectorizer implements tokenization and occurrence counting:

We use it to tokenize and count the occurrence of words in a parsimonious text corpus:

The default configuration tokenizes the string by extracting words of at least 2 letters. The function to do this step can be called explicitly:

Each term found by analyzer during the fitting process is assigned a unique integer index, corresponding to a column in the resulting matrix.

The inverse mapping from feature name to column index is stored in the vocabulary:

Thus, in future calls to the transform method, words not seen in the training corpus will be completely ignored:

Note that in the previous corpus, the first and the last document have exactly the same words, because are encoded as the same vector. In particular, we lose the information that the last document is a questionable form. To prevent reversal of word order, we can extract 2-grams of words in addition to the 1-grams of the monadic model:

The vocabulary extracted by the vectorizer thus becomes larger and at the same time allows to disambiguate when locating patterns:

In a large text corpus, some words will appear many times and therefore have little meaningful information about the actual content of the document. If we provide unhandled count data directly to the classifier, these frequent word groups will mask words that we care about but rarely occur.

To recalculate the feature weights and transform them into floating point values suitable for use by the classifier, it is therefore common to use the TF-IDF transformation.

TF denotes the term frequency, while TF-IDF denotes the term frequency multiplied by the converted document frequency.

Multinomial Naïve Bayes

Since the features of text data are independent of each other, in order to determine the probability of a sample, we use the Multinomial Naïve Bayes algorithm, which is also commonly used in text analysis problems.

Definition

The plain Bayesian approach is a set of supervised learning algorithms based on Bayes' theorem, which "simply" assumes that each pair of features is independent of each other. Given a category y and a vector of related features from x_1 to x_n , Bayes' theorem states the following relationship:

$$P(y|x_1,\ldots,x_n)=\frac{P(y)P(x_1,\ldots,x_n|y)}{P(x_1,\ldots,x_n)}$$

Using the naive assumption that each pair of features is independent of each other:
For all :i to hold, the relation can be simplified to

$$P(y|x_1,\ldots,x_n)=\frac{P(y)\Pi_{i=1}^{n}P(x_i\mid y)}{P(x_1,\ldots,xn)}$$

Since it is a constant in the given input$P(x_1,\ldots,x_n)$, we use the following classification rule:

$$P(y\mid x_1,\ldots,x_n)\ \propto P(y)\prod_{i=1}^{n}P(x_i\mid y)\Downarrow$$

$$\hat{y}=\arg\max_{y}P(y)\prod_{i=1}^{n}P(x_i\mid y)$$

We can use the Maximum A Posteriori Probability (MAP) to estimate; the former is the relative frequency of the categories y in the training set.

Most of the differences between the various plain Bayesian classifiers come from the different assumptions made when dealing with the $P(x_i\mid y)$ distribution.

Despite its simplistic assumptions, plain Bayesian works well in many practical situations, especially document classification and spam filtering. These efforts require a small training set to estimate the necessary parameters. (For theoretical reasons why plain Bayesian performs well and what types of data it is applicable to, see the references below.)

Compared to other more complex methods, the plain Bayesian learner and classifier are very fast. The decoupling of the distribution of classification conditions means that each feature can be estimated independently and separately as a one-dimensional distribution. This in turn helps to alleviate the problems associated with dimensional catastrophes.

Practical approach

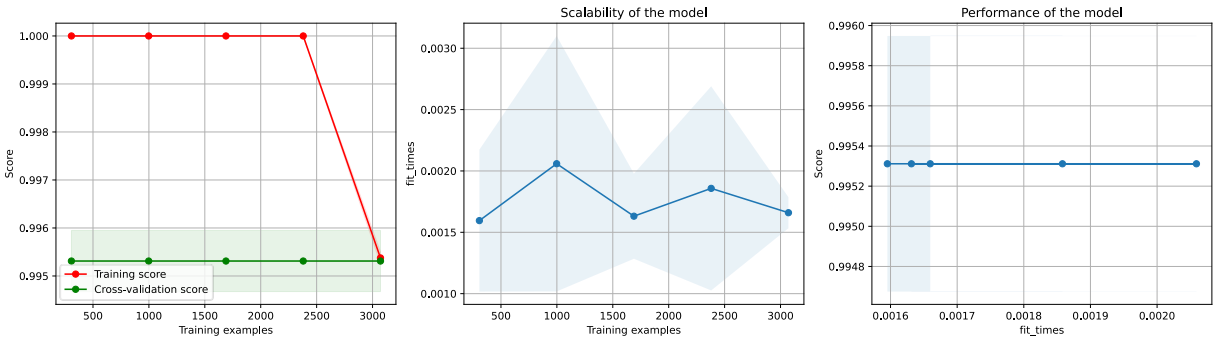


Figure 7

The multinomial distribution plain Bayesian algorithm implements the plain Bayesian algorithm for data subject to multinomial distribution and is one of the two classical plain Bayesian algorithms used for text classification (a field where data is often represented as word vectors, although in practice TF-IDF vectors perform well in prediction). The distribution parameters are determined by each class y, where n is the number of features (for text

classification, it is the size of the vocabulary) θ_{yi} is the probability $P(x_i \mid y)$ of belonging to feature i in class y in the sample.

The parameter θ_y is estimated using the smoothed maximum likelihood estimation method, i.e., relative frequency counts:

5.3 Spatiotemporal model

The date and location of discovery are provided in the observation report, and we believe that these two can be considered together to give a probability indicator through the model.

Preprocessing: The time in the dataset is in date format, which cannot be read by the program, so a method is proposed to convert the date into a timestamp in days, with the timestamp origin being the UNIX timestamp origin (1970-01-01), which visualizes the distance between each date.

Stochastic Gradient Descent

Due to the sparse and discrete spatiotemporal data, the more efficient SGD method is used for easy debugging and optimization.

Definition

We describe here the mathematical details of the SGD procedure.

Given a set of training examples (x1, y1), …, (xn, yn) where xi∈Rm and yi∈R (yi∈ −1,1 for classification), our goal is to learn a linear scoring function $f(x) = wTx + b$ with model parameters w∈Rm and intercept b∈R. In order to make predictions for binary classification, we simply look at the sign of f(x). To find the model parameters, we minimize the regularized training error given by

$$E(w,b) = 1n\sum i = 1nL(yi, f(xi)) + \alpha R(w)$$

where L is a loss function that measures model (mis)fit and R is a regularization term (aka penalty) that penalizes model complexity; α>0 is a non-negative hyperparameter that controls the regularization strength.

Different choices for L entail different classifiers or regressors:

- Hinge (soft-margin): equivalent to Support Vector Classification. $L(yi, f(xi)) = max(0,1 - yif(xi))$.
- Perceptron: $L(yi, f(xi)) = max(0, -yif(xi))$.
- Modified Huber: $L(yi, f(xi)) = max(0,1 - yif(xi))2\ if\ yif(xi) > 1$, and $L(yi, f(xi)) = -4yif(xi)$ otherwise.
- Log: equivalent to Logistic Regression. $L(yi, f(xi)) = log\ (1 + exp\ (-yif(xi)))$.
- Least-Squares: Linear regression (Ridge or Lasso depending on R). $L(yi, f(xi)) = 12(yi - f(xi))2$.
- Huber: less sensitive to outliers than least-squares. It is equivalent to least squares when $|yi - f(xi)| \leq \varepsilon, and\ L(yi, f(xi)) = \varepsilon|yi - f(xi)| - 12\varepsilon 2$ otherwise.
- Epsilon-Insensitive: (soft-margin) equivalent to Support Vector Regression. $L(yi, f(xi)) = max(0, |yi - f(xi)| - \varepsilon)$.

Popular choices for the regularization term R (the penalty parameter) include:

- L2 norm: $R(w) := 12\sum j = 1mwj2 = ||w||22$,

- L1 norm: $R(w):=\sum j=1m|wj|$, which leads to sparse solutions.
- Elastic Net: $R(w):=\rho 2\sum j=1nwj2+(1-\rho)\sum j=1m|wj|$, a convex combination of L2 and L1, where ρ is given by 1 - l1_ratio.

Practical approach

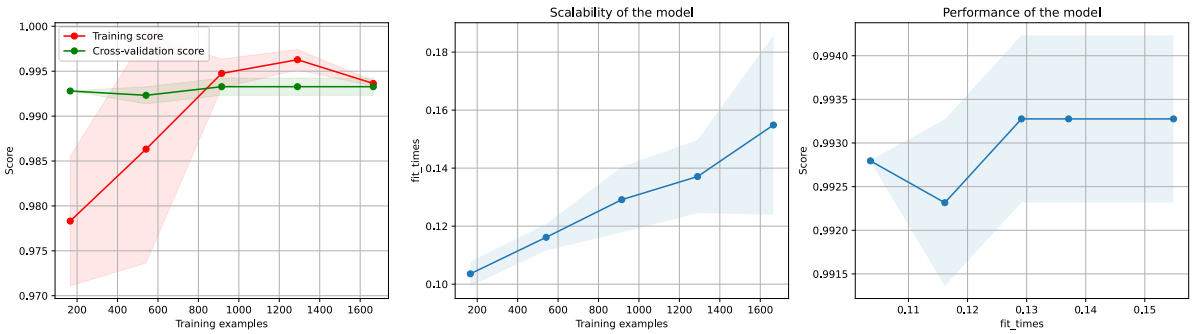


Figure 8

Stochastic gradient descent is an optimization method for unconstrained optimization problems. In contrast to (batch) gradient descent, SGD approximates the true gradient of E(w, b) by considering a single training example at a time.

The class SGDClassifier implements a first-order SGD learning routine. The algorithm iterates over the training examples and for each example updates the model parameters according to the update rule given by

$$w \leftarrow w-\eta[\alpha\partial R(w)\partial w+\partial L(wTxi+b,yi)\partial w]$$

where η is the learning rate which controls the step-size in the parameter space. The intercept b is updated similarly but without regularization (and with additional decay for sparse matrices, as detailed in Implementation details).

The learning rate η can be either constant or gradually decaying. For classification, the default learning rate schedule is given by

$$\eta(t)=1\alpha(t0+t)$$

where t is the time step, t0 is determined based on a heuristic proposed by Léon Bottou such that the expected initial updates are comparable with the expected size of the weights (this assuming that the norm of the training samples is approx. 1).

5.4 Integrated model

An observation report can get 4 probability metrics after passing the above three models, which are image feature probability, image probability, text probability and spatiotemporal probability. In practice, we found that since the difference between the data volume of the two categories in the training set is very large, the negative data is about 150 times of the positive data volume, so the sensitivity of the trained model to the positive samples is weak, and the state of the samples cannot be judged directly for a single indicator. Therefore, we believe that we need to statistically analyze the above four probabilities in order to obtain intuitive classification data. For this we introduce the logistic regression algorithm.

Logistics Regression with Cross-Validation:

Since we need to combine the four indicators to obtain a probability indicator, and we believe that the hornets' appearance follows a linear law to some extent, so the data set can be regressed in a function, that is why we use Logistics Regression. To explore the best hyperparameters, we apply Cross Validation for this method.

Definition

Log-odds distribution formula

$$P(Y = 1|X = x) = \frac{e^{x'\beta}}{1 + e^{x'\beta}}$$

where the parameter β is commonly estimated by maximum likelihood.

Practical approach

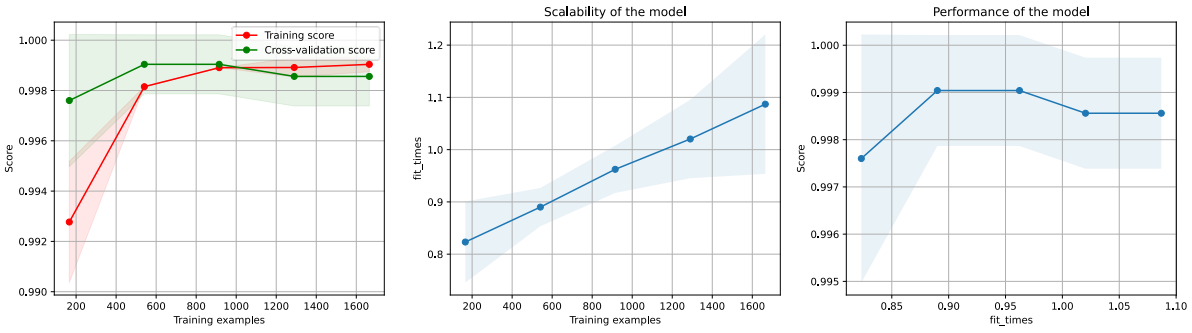


Figure 9

As an optimization problem, binary class ℓ2 penalized logistic regression minimizes the following **cost function**:

$$C \sum_{\{i=1\}}^{n} log\left(exp\left(-y_i(X_i^T w + c)\right) + 1\right).$$

5.5 Model Optimization (Hyperparameter Optimization)

Grid Search with Cross Validation

Since support vector machines require multiple hyperparameter tuning to achieve better performance, for some parameters the best values can only be obtained by trial-and-error methods. grid search is used to optimize SVC by testing the parameters in a given range to achieve the best performance.

Definition

The provided grid search generates candidates from the full range of grid parameter values determined by the param grid parameter. When "fitting" on the dataset, all possible combinations of parameter values are evaluated and the best combination is calculated.

Practical approach

In the current practice, the C (Regularization Parameter) and gamma (Kernel coefficient) of SVC need to be tuned, so the best value is tried out by GSCV based on the empirical value

setting range.

6 Model III: Validation Model

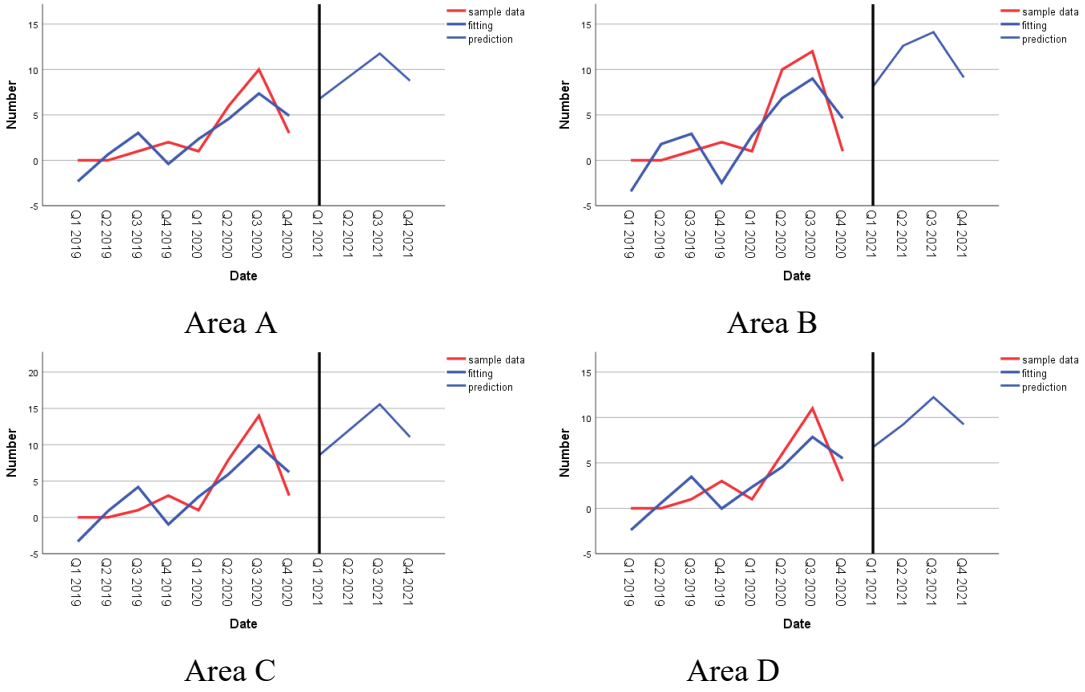
The model was constructed in a similar way to model one, using the expert modeler of SPSS software to construct Winters' additive to achieve data prediction.

7 Conclusion

7.1 Aspect of Prediction

7.1.1 Prediction of Data Using Model I

The fitted and predicted curves are shown in the following figures:



Area A Area B

Area C Area D

Figure 10

Table 7 Statistical table of predicted values

region	date	predicted value	interval prediction	
region A	Q1 2021	7	0	13
	Q2 2021	9	3	16
	Q3 2021	12	5	18
	Q4 2021	9	2	15
region B	Q1 2021	8	-2	18
	Q2 2021	13	3	23
	Q3 2021	14	4	24
	Q4 2021	9	-1	20

region C	Q1 2021	9	-1	18
	Q2 2021	12	2	22
	Q3 2021	16	5	26
	Q4 2021	11	1	21
region D	Q1 2021	7	-1	14
	Q2 2021	9	2	17
	Q3 2021	12	5	20
	Q4 2021	9	1	17

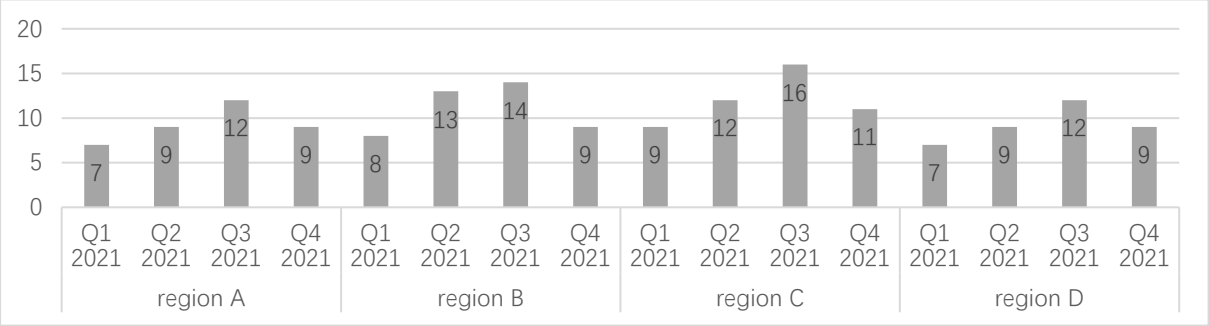


Figure 11

7.1.2 Data Analysis

From the predicted values, it can be obtained that Region C has the largest population in the first quarter of 2021, followed by Region B. Region B has the highest number of populations in the second quarter, followed by Region C. Region C has the highest number of populations in the third quarter, followed by Region B. Region C had the highest number of populations in the fourth quarter, and the rest of the regions had approximately the same number.

From the above analysis, it is clear that region B and region C have been the hardest hit areas, always occupying the first two places in the sample size. This shows that the spread of the population is predictable and is mainly distributed in region C in the first, third and fourth quarters and in region B in the second quarter.

Since circular regions do not facilitate the direct derivation of latitude and longitude relationships, we used the method of taking the outer square quadrilateral of a circular region to determine the population aggregation area. Therefore, the latitude and longitude ranges corresponding to region B were 48.6149°N~48.9753°N 122.2513°W~122.7526°W, and the latitude and longitude ranges corresponding to region C were 48.6600°N~49.0204°N 122.3139°W~122.8152°W.

(Note: Since the range of the hive was 30 km, the actual range of latitude and longitude is slightly larger than this range)

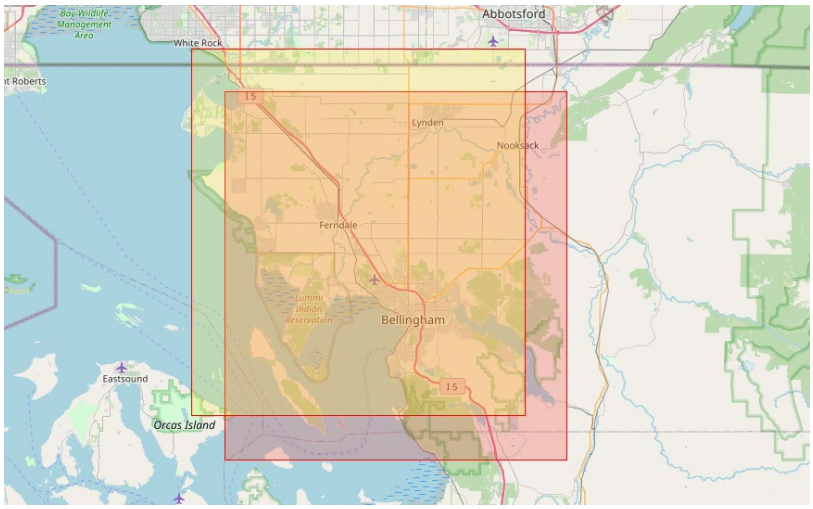


Figure 12

7.2 Aspect of Classification

For this problem, we consider that predicting the probability of misclassification is equivalent to determining the probability that the bee reported by sighting is Vespa Mandarina. Therefore, for the dataset files and image files, we create a set of probability-based classification models.

The model output is the probability that the sighting report is positive, and since the Negative sample size is much larger than the Positive sample size in the identified Lab Status data, the probability that the sighting report is positive can be considered a small probability event, so only the investigation priority needs to be determined based on the model's probability ranking.

All algorithms in this model support partial fit, and all newly given reports can be updated directly on the original model weights. Also from the previous section, the activity of Vespa Mandarina is very seasonal, so the model should be updated once every quarter.

7.3 Aspect of Validation

By varying the sample sizes of the four quarters of 2021, it is found that when the sample sizes of the four quarters of 2021 are 0, 1, 2, and 0, the predicted values of the four quarters of 2022 obtained by time series forecasting are all zero, and the confidence intervals of the four quarters are (-5,3), (-6,7), (-6,10), and (-9,9) at the 95% confidence level. Also, since all sample point data for 2019 and 2020 used for the prediction were distributed in Washington State, if the statistical values for the four quarters in 2021 are less than 0,1,2,0, it can be judged that in 2022, the species of hornet will no longer be present in Washington State, i.e., this pest has been eliminated in Washington State.

8 Strengths and Improvement

8.1 Strengths

- For the prediction of the number of wasp populations, the seasonal factor was taken into account.
- Base on probability, reflected less by the samples.
- Use various model for different type of data, make fully use of dataset.
- Can be updated by single new data by not need to train from the beginning.

8.2 Weakness

- The given sample data is too small, resulting in an R-squared of only about 0.6 and a poor fit, which will inevitably lead to less accurate predicted values.
- Wasps may migrate to other regions and do not necessarily congregate only in the four regions.
- Still need artificial judgment.
- Did not consider the climate and terrain of the place which could influence the behavior of hornet.

8.3 Possible Improvements

- Use more accurate data train the model.

9 Memorandum

MEMORANDUM FOR the Washington State Department of Agriculture
FROM: TEAM2103942
DATE: Feb 9th, 2021
SUBJECT: Strategies for Determining the Confidence Level of Hornet Sighting Reports and Allocating Government Resources to Investigations

---

We examined the public sighting reports of hornets submitted between January 15 and October 23, 2020, and found that although there were more public sightings, very few reports were identified as positive sightings. To make more effective use of government resources to curb the infestation, we examined the sighting report data and proposed three models: a classification model to determine the probability that the report was for the hornets, a prediction model to provide information on the likely occurrence of the swarm, and a verification model to infer the conditions under which the infestation disappeared.

First, the classification model was used to process the sighting reports by classifying the images to determine the morphological characteristics of the bees, analyzing the text to obtain the descriptive characteristics of the most likely to be a hornet, and clustering the time and location to obtain the population characteristics of the most likely to be a hornet colony. The model gives the probability that the bee in the sighting report is a hornet by the above processing. Therefore, for new sighting reports, two approaches can be taken to determine this. First, Individual sighting reports can enter the model to obtain a confidence level for the model judgment, which is judged by setting an empirical threshold, and reports above this threshold are investigated. Meanwhile, according to our investigation, the activity of hornets is extremely seasonal, and in fact it is only necessary to input new sighting reports into the model every quarter to get a series of probabilities and rank them from highest to lowest probability, and then prioritize resources to investigate the high probability reports.

By analyzing the data, we built a prediction model based on known data as well as data judged by a classification model. Assuming that each wasp sighting was a different individual, we were able to predict the number of observations for a future period. We found a 6.5-fold increase in the observed number of hornets in 2020 compared to 2019 and predicted a 1.8-fold increase in 2021 compared to 2020. We believe that the pest control measures taken by the relevant authorities have been effective, and that the hornet population will grow at a lower rate after a certain level of development due to population competition and other effects, so a slower growth rate in 2021 years is expected.

At the same time, by analyzing the sightings of this species of wasp in two areas with overlapping parts in Washington State, we obtained a basis for judging that the infestation might disappear in the state, i.e., when the number of positive sightings in the state was determined to be [0, 1, 2, 0] in all four seasons of a given year, we can assume that the infestation will not occur in the state in the following year.

In addition to the analysis of the data, we should consider how to obtain additional information. As far as the analysis of the available information is concerned, most of them have been confirmed not to be wasps, and the other uncertain parts have been confirmed not to be wasps after our modeling processing analysis. First of all, we should reassure the public not to make them panic so that they can provide more reliable data. At the same time, the authorities should publish some characteristic points of the samples confirmed to be hornets, so that the public can make the first round of simple judgments by themselves to reduce the workload of the government.

FOR THE Officer
TEAM 2103942

Attachment:
Paper

References

"Automatic Capacity Tuning of Very Large VC-dimension Classifiers", I. Guyon, B. Boser, V. Vapnik - Advances in neural information processing 1993.

"Support-vector networks", C. Cortes, V. Vapnik - Machine Learning, 20, 273-297 (1995).

"Learning representations by back-propagating errors." Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams.

"Stochastic Gradient Descent" L. Bottou - Website, 2010.

"Backpropagation" Andrew Ng, Jiquan Ngiam, Chuan Yu Foo, Yifan Mai, Caroline Suen - Website, 2011.

"Efficient BackProp" Y. LeCun, L. Bottou, G. Orr, K. Müller - In Neural Networks: Tricks of the Trade 1998.

"Adam: A method for stochastic optimization." Kingma, Diederik, and Jimmy Ba. arXiv preprint arXiv:1412.6980 (2014).

H. Zhang (2004). The optimality of Naive Bayes. Proc. FLAIRS.

"Solving large scale linear prediction problems using stochastic gradient descent algorithms" T. Zhang - In Proceedings of ICML '04.

"Regularization and variable selection via the elastic net" H. Zou, T. Hastie - Journal of the Royal Statistical Society Series B, 67 (2), 301-320.

"Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent" Xu, Wei

Pattern Recognition and Machine Learning, Chapter 4.3.4, Christopher M. Bishop

Minimizing Finite Sums with the Stochastic Average Gradient, Mark Schmidt, Nicolas Le Roux, and Francis Bach

SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives, Aaron Defazio, Francis Bach, Simon Lacoste-Julien

Appendices

Source Code for Coordinate Data Visualization

```
load('data.mat')
w_min = min(W);
w_max = max(W);
l_min = min(L);
l_max = max(L);
w1 = W - repmat(w_min, size(W, 1), 1);
l1 = L - repmat(l_min, size(W, 1), 1);
w2 = W - repmat(46.5, size(W, 1), 1);
l2 = l1 .* (w2 <= 0) * 83.358926;
l2 = l2 + ((0 < w2) & (w2 <= 1)) * 82.191231 .* l1;
l2 = l2 + ((1 < w2) & (w2 <= 2)) * 81.003346 .* l1;
l2 = l2 + ((2 < w2) & (w2 <= 3)) * 79.795563 .* l1;
w3 = w1 * 111;
X = [l2, w3];
plot(l2, w3, 'o')
```

Source Code for SVM

```
from sklearn.decomposition import PCA
pca = PCA(n_components=150, whiten=True)
X_train_pca = pca.fit_transform(train_set.data)

from sklearn.model_selection import GridSearchCV
from sklearn.svm import SVC
param_grid = {'C': [1e3, 5e3, 1e4, 5e4, 1e5], 'gamma': [0.0001, 0.0005,
0.001, 0.005, 0.01, 0.1], }
clf = GridSearchCV(SVC(kernel='rbf', class_weight='balanced', probabil-
ity=True) , param_grid)
clf = clf.fit(X_train_pca, train_set.target)
```

Source Code for MLP

```
from sklearn.neural_network import MLPClassifier
clf = MLPClassifier(random_state=1, max_iter=300)
clf.fit(train_set.data, train_set.target)
```

Source Code for SGD

```
from sklearn.preprocessing import StandardScaler
std = StandardScaler()
x_train = std.fit_transform(train_dataset.data)
y_train = train_dataset.target.astype(np.int8)
from sklearn.linear_model import SGDClassifier
from sklearn.model_selection import GridSearchCV
param_grid = {'alpha': 10.0**-np.arange(1,7), }
clf = GridSearchCV(SGDClassifier(max_iter=1000, tol=1e-3,loss='modified_huber',shuffle=True) , param_grid)
clf.fit(x_train, y_train)
```

Source Code for Bayers

```
from sklearn.feature_extraction.text import TfidfVectorizer
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(train_set.data)

from sklearn.naive_bayes import MultinomialNB
from sklearn.model_selection import learning_curve
clf = MultinomialNB()
clf.fit(X, train_set.target)
```

Source Code for Logistic Regression

```
X = sum_train[['tl_pos', 'svm_pos', 'mlp_pos', 'nc_pos']]
Y = sum_train['Lab Status']
X = X.dropna(axis=0,how='any')
Y = Y.dropna(axis=0,how='any')

from sklearn.linear_model import LogisticRegressionCV
clf = LogisticRegressionCV()
clf.fit(X,Y)
```