Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2107239

A report analysis and resource allocation method based on time series and data stream

Summary

In our paper, we build an **Autoregressive Moving Average Model(ARMA)** to analyze existing data. At the same time, using univariate nonlinear regression combined with the time series analysis to predict the future trend. And resource priority allocation model called **Time Topographic Longitude Latitude Population Model(TTLLP)** to assist the government in deploying limited resources.

As for primary problem I, we aim to construct a **Convolutional Neural Network(CNN)** as a classifier so that we are capable to recognize the new coming image and give the supporting decision to save the limited resource, so we use the basic image preprocessing way to resize, blur the images in order to get the clearing set to train the model. Since the number of positive data sets is very small compared to the number of negative data sets, we use **Generative Adversarial Networks(GAN)** to expand the positive samples while ensuring that no new external data is introduced. After validating, we find that CNN could only classify the negative images but weak in predicting the positive images. In improving this weakness, the positive part is selected separately, and the model is corrected using the principle of grey prediction and other Holt-Winters' additive method.

In addition to mining the information of image data, we also make predictions about future trends. Based on the pest's natural life cycle, combining with relevant policies and measures, we conduct a time series analysis on it. The **ARMA and univariate nonlinear regression models (AMRA-UNLR)** are used to analyze and summarize the existing information, and finally get the development trend.

As for primary problem II, we create a model which can give the government advice about allocating limited resource, so based on the given attribute, and first we come up with a model called **Time-Longitude-Latitude(TLL)**, we concentrate on the time and distance of each new coming report, but this model is too simple as it overlooks the other effect factors. And we provide a revision model called TTLLP, which is involved the population and terrain elements. Also our model can update as the time going and the update frequency should be in the range of two weeks to a month. While maintaining this update frequency, the government's active measures will expel this pest from Washington State, so we have given a basis for judgment through the model: when the probability of occurrence is lower than a certain value within a period of time, we believe that the pest has been eradicated.

We finally conduct sensitivity analysis, dissect pros and cons of our model and present a memo of our work to the state agricultural department.

Keywords: ARMA; CNN; GAN; priority allocation; nonlinear regression; TTLLP;

Contents

1 Introduction

1.1 Problem Background

The ecological impact caused by biological invasion is very severe. Invasive species may threaten the survival of local species, and have a great impact on economic development and even human health. Nowadays, the invasive species Vespa mandarinia has emerged in Washington state, which influences the local ecological environment. However, the sighting reports of the masses are sometimes not accurate enough. In order to make effective use of existing resources, and carry out further follow-up investigation. A model is needed to process and analyze the report data provided.

We proposed to decompose this problem into two main parts:

- Process the data provided.
- Use model to analyse these reports.

1.2 Literature Review

A number of researchers have previously contributed to neural network models for prediction. The study of Krizhevsky A et al. use deep convolutional neural networks to classify images into different groups and show the advances in neural information processing systems[1]. LeCunY's et al. literature gave us a way to apply gradient-based learning to document recognition[2]. Barabassi A.L's research provided us with a reference for basic characteristics: the idea and method of analyzing the dynamic topology of complex network systems.

1.3 Overview of Our Work

The first main question is about analyzing and interpreting the data reported by the public. What we first adopt is to research and process the images in the given sample set. We use the confrontation generation network model to expand the number of existing image samples. Thereby increasing the number of training sets, and then use the convolutional neural network to classify and learn the image to obtain a classification model, but due to the overfitting of the classification effect, part of the research on the image ends here. Then we conducted relevant statistics and research on Detect Date in the given data set.

The second main problem is that for the priority allocation of limited resources, we first gave a rough TLL model. Because the TLL model is less robust and not strong enough at the same time, we proposed it on the basis of the TLL model. The TTLLP model maximizes the integration based on the limited resources contained in the problem, thus giving a model that can provide the government with priority allocation. For the migration of models over time.

2 Assumptions and Notations

2.1 Assumptions

Due to the lack of necessary data and the limitation of our knowledge, we make the following assumptions to help us model. These assumptions are the basis of our subsequent analysis.

- Assuming that all the reported data in the table is representative, you can use this part of the data to infer the whole.
- Assuming that the data reported by public are real information obtained by themselves.
- Assuming that the Asian giant hornets species that appeared in Washington State are similar to those in other regions, we can use the habits of such hornets in other regions to infer the survival law of the target species.
- We assume that the time of detection contains more information than the time of submission.
- Assuming that for those positive reports, the time they detected the Asian giant hornets should not be longer than two weeks before the local hornets began to appear.
- Assuming that the time of the first positive report detected is the approximate time (within two weeks) when this species of hornets first appeared in Washington State.
- Assuming that if there is no new positive report within a month, it can be approximated that this species of hornet does not exist in this area.

2.2 Notations

Symbol	Description
t	Time effect factor
To	Topographic
Lon	Longitude
Lat	Latitude
Po	Population
t_d	Detect date
t_l	The latest date of positive ID detect in this area
p	Possibility of unverified becoming possitive ID
α	The enhanced weight of positive ID
Δd	The distance of two area(space effect factor)
Δt	Time bias(should be positive)
β	Base weight for time and space
R	Radius of the earth

3 Data Preprocessing

3.1 Database Processing

Given that we have some years of sighting reports in Washington State, the original database is complicated. Therefore, we should conduct data processing based on the completeness and

usefulness of the information. We use MySql database to narrow down the data.

1 Excel2MySql
Import the data in the Excel table into the MySql database, firstly we check the inaccurate data in MySql.Among them, there are data rows in the format "30/8/1899" and "" in the year. It is also found that the globalId of these samples does not have a corresponding FileName in the mapping table, and the number of such data rows is small, so we delete them directly.

2 Merging
Use the globalId in the two Excel tables as the associated key, and combine the two tables into a new table with no globalId and only other valid information as the subsequent start data table.

3.2 Image Stream Processing

For the ".MOV" type files that exist in the given data set, use the python opencv tool to extract the key frames in the video and output them as new picture files into the data set, delete the original ".MOV" type files. In this way, we obtain a dataset with only image files.

4 Public reports data interpretation

4.1 ARMA and univariate nonlinear regression model

4.1.1 Problem Introduction

From many reports from the citizens, It is clear that the presence of Vespa mandarinia has caused seriously anxiety. In order to appease the people, it becomes urgent to predict and take measures.

Firstly, using the data we draw a line chart. According to its data characteristics and order of magnitude, we divided the data into three groups: detect date between 2010 and 2019, detect data in 2020 and submission date (all in 2020). From the chart 1, between 2010 and 2018, the number of detection was within single digits, and there was not much fluctuation. But from 2019, the number has started to increase dramatically. In 2020, the number of monthly reports has risen directly to the order of hundreds of digits. In August of 2020, the number even exceeded 1,000. As for the date of submission, it is basically concentrated in 2020.

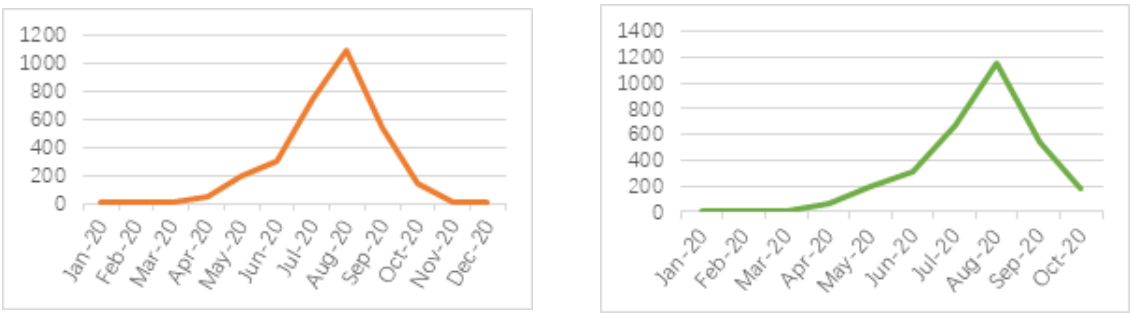

Figure 1: Two types of labstatus distribution

Because this giant hornet has many natural enemies[4], such as spiders, bats, birds and so on. Therefore, in the process of spread, there is a balance point with natural enemies (BPWNE)

in its population. On the basis of previous studies, the equilibrium point of natural enemies should gradually stabilize with time[5]. And satisfy the following formula

$$P_1(\frac{h}{\gamma}, \frac{B\gamma}{\lambda h}, \frac{A\lambda\gamma+\alpha B\gamma-h^2\lambda}{\lambda\beta h})$$

α represents the interaction coefficient between giant hornet and damaged crops $(\alpha \geq 0)$; β indicates the probability of giant hornet being preyed by natural enemies $(0 \leq \beta \leq 1)$; h indicates the lethality rate of chemical pesticides $(0 \leq h \leq 1)$; B indicates the increase in crops per unit time $(B \geq 0)$; λ indicates the mortality of crops due to giant hornet damage $(0 \leq \lambda \leq 1)$; λ represents the growth rate of giant hornet's natural enemies $(\lambda \geq 0)$.

In addition to the influence of natural enemies, existing survey data show that giant hornet likes to build nests in places like natural cavities such as hollow trees and sometimes inside the walls of buildings[6]. Therefore, the prediction range can be limited to this type of environment. With a controllable area, the prediction is achievable and credible.

At the same time, since the life cycle of this hornet is similar to other wasps, follow the cycle shows below. Based on this quantitative law of periodic changes, the prediction accuracy is greatly improved.

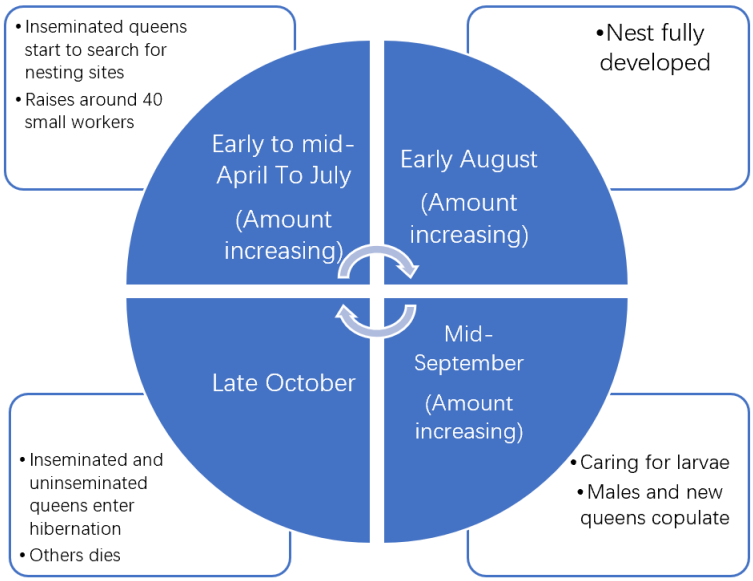

Figure 2: The habits of Vespa mandarinia

According to the data provided, the propagation path is also clear. Therefore, based on the above reference and other criteria, we believe that the spread of the pest can be predicted. Considering the characteristics of the given data and the accuracy of the model, **Autoregressive Integrated Moving Average (ARIMA) model** would be a good way. Next, ARIMA model will be used as the main line, with the aid of other methods such as grey-prediction to forecast the propagation.

4.1.2 Model Introduction

Autoregressive Integrated Moving Average (ARIMA) mode has been widely used in agricultural, econometrics and so on. It is a generalization of an **Autoregressive Moving Average** (ARMA) model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series (forecasting)[7].

ARMA is composed of AR (Auto Regressive Model) and MA (Moving Average Model), so it can be expressed as ARMA (p, q). p is the autoregressive order, and q is the moving average order.

$$x_t = u_t + \phi_1 u_{t-1} + \phi_2 u_{t-2} + \cdots + \phi_q u_{t-q} + \theta_1 x_{t-1} + \theta_2 x_{t-2} + \cdots + \theta_p x_{t-p}$$

It can be seen from the formula that ARMA combines the advantages of the two models. Among them, AR can solve the relationship between current data and later data, and MA can solve the problem of random changes, which is noise. The ARIMA model is also based on a stationary time series. Or it is based on a stable time series after differentiation. ARMA can be seen as a special form of ARIMA, expressed as ARIMA (p, d, q). Same as ARMA, the meaning of p and q has not changed. d refers to the number of differences made when time becomes stationary. It also represents what 'Integrate' means here. After test, ARIMA seems to be a better choice.

We predict the actual spread of pests, so here, only reports that LabStatus is positive are meaningful. According to the hypothesis assumption. It is assumed here that in the days when there is no positive report, no new giant hornets are produced. That is, the number of positive reports generated on that day is 0. After screening and classification statistics, there are 14 reports that LabStatus is positive. This set of data is distributed from 2019.9 to 2020.10. The cases are all distributed between 48.7-49.2 in latitude and 122.4-124.0 in longitude. By querying the topographic map, we learned that the natural environment in this area is mostly low shrubs. It is known that this area is suitable for them to survive, in line with the survival and nesting environment of the giant hornets.

4.1.3 Model Analysis

Next, we take days as the unit to perform time series analysis on the model.

The use of the ARMA model requires that the time series must be stationary, so the first step is to check the stationarity of the original data. There are many test methods, including ADF, KPSS, P-P and so on. ADF test and KPSS test are used here.

An augmented Dickey-Fuller test (ADF) tests the null hypothesis that a unit root is present in a time series sample[8].

Kwiatkowski Phillips Schmidt Shin (KPSS) tests are used for testing a null hypothesis that an observable time series is stationary around a deterministic trend (i.e. trend-stationary) against the alternative of a unit root[10].

When testing the original data directly, adftest=1, kpsstest=1, and so test fails. In order to obtain a stationary time series, first-order difference processing is performed on the data.

Next, determine the order of the ARMA model. The order is determined by the ACF method and PACF method. Autocorrelation function (ACF) is an (complete) auto-correlation function which gives us values of auto-correlation of any series with its lagged values. Partial

autocorrelation function (PACF) gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags.

According to the input value, the following two equations can be obtained. And according to the general characteristics described in the above table, the following rules can be summarized: The last lag value of PACF outside the blue line (that is, outside the threshold) is the p. The last ACF lag value outside the blue line (that is, outside the threshold) is the q.

$$\text{Akaike Information Criterion} : AIC = -2\ln(L) + 2k$$

$$\text{Bayesian Information Criterion} : BIC = -2\ln(L) + \ln(n) * k$$

L is the maximum likelihood of the model, n is the number of data, K is the number of variables of the model.

From equations above, p and q seem to be too large. Therefore, another method is adopted, namely, akaike information criterion (AIC) and bayesian information criterion (BIC) criteria are used to select the order. Finally, p and q are determined to be 4 and 7, respectively.

Then, in order to ensure that the determined order is appropriate, a residual test is also required. The standardized residuals is the residual signal after subtracting the signal fitted by the model from the original signal. The following is the result of the residual test drawn.

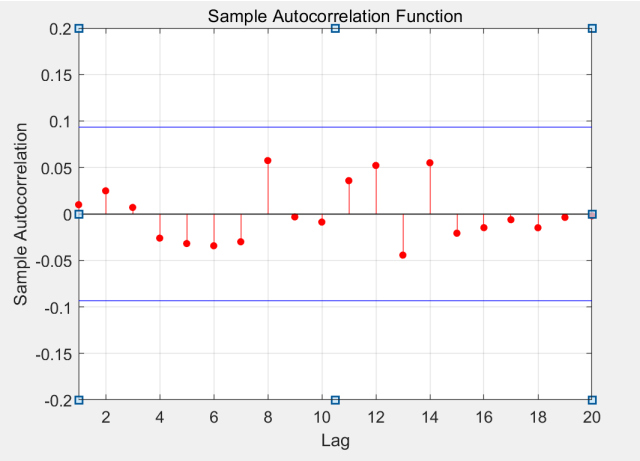

Figure 3: The habits of Vespa mandarinia

The ideal residual should be close to a normal distribution. This condition is basically met here. Think that ARMA modeling meets the requirements.

Finally, a prediction step of 300 days is made from the existing data. As shown in the figure, the black line is the forecast of future values. The red line is the upper and lower limits of the 95% confidence interval, that is, there is a 95% probability that the true value of the future will fall within this range. It can be seen that for a long period of time in the future, the probability of the emergence of giant hornets will be controlled within a relatively small range. And also follow the law of trend.

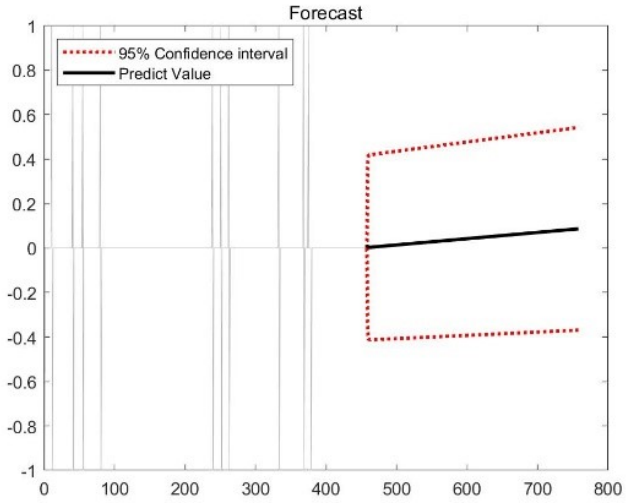

Figure 4: The result of prediction

Combining the life cycle of the bumblebee and the cycle of geographic climate change, it is not difficult to conclude that the occurrence of the bumblebee conforms to the probability distribution with a certain period of time as the cycle. Below, the statistics are calculated on a monthly basis. Perform univariate nonlinear regression processing on this set of data.

Nonlinear regression is a form of regression analysis. The data are fitted by a method of successive approximations. Taking September 2019 as the first month, the number of cases that appear each month is shown in the figure below.

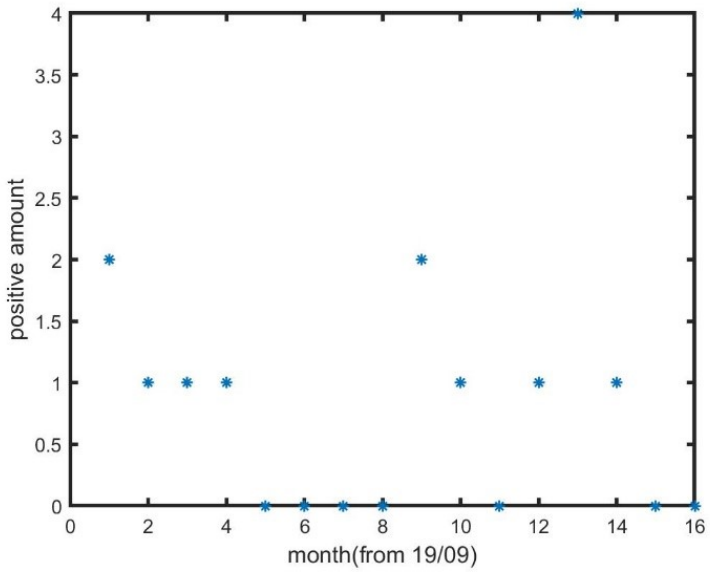

Figure 5: Positive data distribution

Approximately satisfy the distribution of the sum of multiple indices. Through trial and

constant iteration and modification, the formula:

$$f = a_1 * exp(-((x-b_1)/c_1)^2) + a_2 * exp(-((x-b_2)/c_2)^2) + a_3 * exp(-((x-b_3)/c_3)^2) + a_4 * exp(-((x-b_4)/c_4)^2)$$

Table 1: The result of regression analysis

Coefficient	Value	Intervals of 95% confidence bounds
a_{1}	5.093	(-22.31, 32.5)
b_{1}	13.08	(12.46, 13.71)
c_{1}	0.9733	(-1.329, 3.275)
a_{2}	1.574	(-224.7, 227.8)
b_{2}	-28.98	(-3.956e+10, 3.956e+10)
c_{2}	1.007e+05	(-5.703e+13, 5.703e+13)
a_{3}	-1.8	(-21.68, 18.08)
b_{3}	6.055	(0.1965, 11.91)
c_{3}	2.273	(-12.08, 16.62
a_{4}	-2.716	(-32.97, 27.54)
b_{4}	13.47	(-5.626, 32.57)
c_{4}	2.816	(-54.95, 60.58)

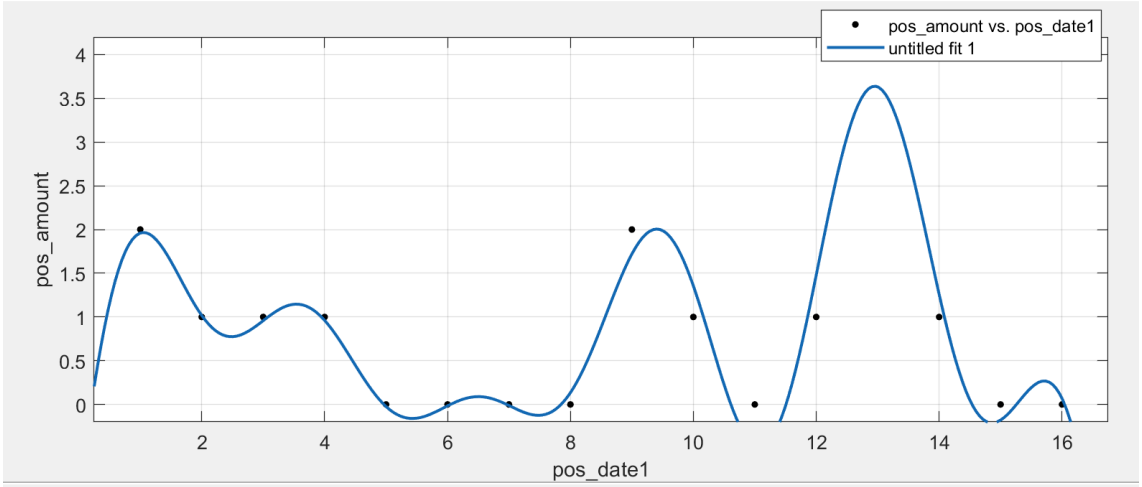

Figure 6: The result of nonlinear regression analysis

According to the forecast one year as a cycle, that is, the number of months 1-12 is a cycle. Bringing 12-16 months of data into the test, it is found that the coincidence rate is higher, indicating that the model is more feasible. At the same time, due to the following laws in the life cycle of giant hornets: the population numbers continue to increase from spring to autumn, and the fall and winter are extremely decreased. The increase or decrease trend of the model is also very consistent with this law. Verify the correctness of the model again.

4.1.4 Precision analysis

Sum of the Squared Errors (SSE): This parameter calculates the sum of squared errors between the fitting parameters and the corresponding points of the original data. The calculation formula

is:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

The smaller the SSE, the better the model selection and fitting.

R-square: R-square is determined by two parameters, SSR and SST. SSR is the sum of squares of the difference between the mean of the predicted data and the original data. The formula is as follows:

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

SST is the sum of the squares of the difference between the original data and the mean.

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Therefore, SST=SSE+SSR, R-square is defined as:

$$R - square = \frac{SSR}{SST} = \frac{SST - SSE}{SST} - \frac{SSE}{SST}$$

It can be known from the above expression that the value range of the determination coefficient is [0,1]. The closer to 1 indicates that the variables of the equation have a stronger ability to explain y and the model fits the data better.

Adjusted R-square: Compared with R-square, this parameter removes the influence of the increase in the number of variables on the result of goodness of fit judgment. The calculation formula is:

$$Ad\,justR - Square = 1 - (1 - r^2)\frac{n-1}{n-k}$$

RMSE: This parameter is the square root of the mean of the sum of the squares of the errors of the predicted data and the corresponding points of the original data. Calculated as follows:

$$RMSE = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y})^2}$$

Because it is the same case as SSE, the fitting effect reflected by SSE is also the same.

In the above formulas, y is the value to be fitted, the mean value is $\bar{y}$, the fitted value is $\hat{y}$, n is the number of samples of y, and k is the number of variables, here k=1.

4.1.5 Model Particularity

ARMA method can process most time series. After the processing difference, the sequence becomes a stationary random process, which meets the modeling requirements of ARMA. By taking more parameters, the accuracy of spectrum estimation is improved and the spectrum resolution is optimized. Considering that the spread of this pest is affected by many factors such as seasons, regions, policies, etc., it is reasonable to use ARMA to deal with and predict. At the same time, due to the non-linear relationship, a non-linear regression prediction model is used. Starting from the original data, combined with relevant biological, social, and agricultural knowledge, the regression function is finally determined.

The above two lines are merged to form our ARMA-UNLR model. This model is clear, flexible, and has a high fault tolerance rate. It also lays a good foundation for finding reasonable measures and evaluating their effects.

4.2 Convolutional Neural Network Model

4.2.1 Model introduction

We observed that the data set is mainly image and video data, and the name of the image can be found in the table. Therefore, we can label the images we have at present, and the sample of the data set is not huge. Based on this characteristic of data, we consider introducing Convolutional Neural Network (CNN) as an image classification method, **trying to get a model that predicts the likelihood of a mistaken classification.**

We refer to the previous paper on AlexNet[1]. After comparing LeNet[2], we find that AlexNet may have higher accuracy to reach classification requirement, so we choose to use AlexNet with more layers and more connections.

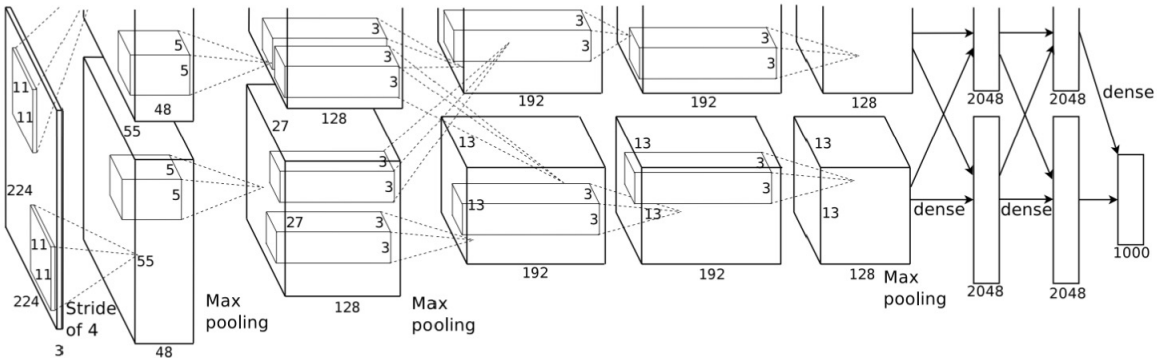

Figure 7: AlexNet introduction[1]

4.2.2 GAN Model Imported

Generative Adversarial Network (GAN) is a class of machine learning frameworks which could help operator to extend the data set and get more samples for trainning model[3].

GAN has a generator and discriminator, generator does a job in generating different types of images and discriminator is response for classifying the generative images. They beating each other and learn from this competition, and finally we will get the extensive samples.

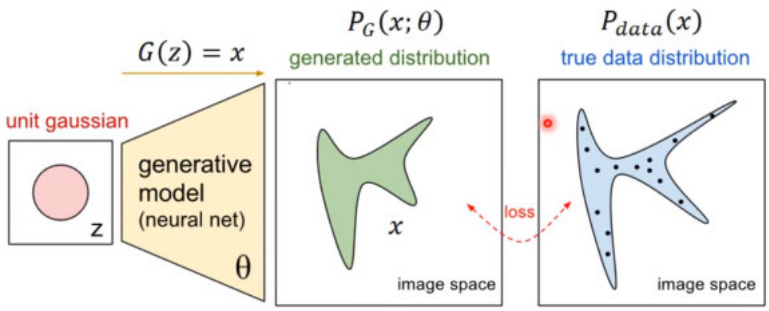

Figure 8: Generator working process

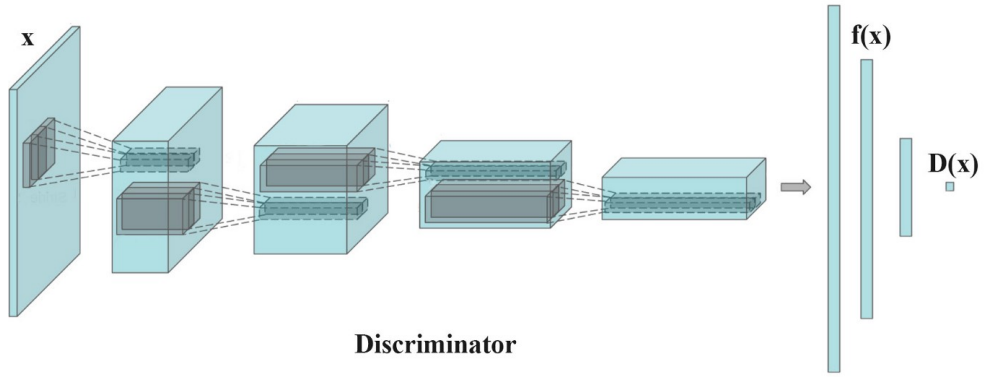

Figure 9: Discriminator working process[13]

4.2.3 Model Construction

Resize all pictures to a standardized input image size of 200 $\times$ 200. Then we Randomly process the image samples by mirroring, blurring, etc.

We find that the number of images with positive ID is small. So we use the GAN(Generative Adversarial Networks) to rich image sample set in order to get a better train result of CNN.

After building a neural network model with 5 layers of convolutional layers and 3 layers of connection layers, we divided the samples into 8:2 (8 is the training set, 2 is the test set), and the test set has 1000 negativeID Picture and 40 positiveID pictures. At the same time we set the learning rate of the neural network to 0.002.

Table 2: Model construction table

layer_name	kernel_size	kernel_num	padding	stride
Conv1	11	96	[1,2]	4
Maxpool1	3	None	0	2
Conv2	5	256	[2,2]	1
Maxpool2	3	None	0	2
Conv3	3	384	[1,1]	1
Conv4	3	384	[1,1]	1
Conv5	3	256	[1,1]	1
Maxpool2	3	None	0	2
FC1	2048	None	None	None
FC2	2048	None	None	None
FC3	1000	None	None	None

4.2.4 Model analysis

Finally, after 5 epoch iterations, the accuracy of the test set we obtained was 0.962. Such prediction results seem to be good. For most pictures, the classifier has a relatively good test result.

However, we found through calculations that, assuming a naive classifier, all pictures are judged as negativeID, that is, 40 positiveID pictures in the test set are judged as negativeID at the same time, and the test accuracy rate obtained

$$\frac{1000}{1040} = 0.962$$

It is basically consistent with the training results of the neural network. It can be seen that using the convolutional neural network AlexNet to classify the picture is not enough and inaccurate, because it will never judge the picture as positiveID.

4.2.5 Model Discussion

In fact, there should be a hybrid model based on image processing and comprehensive consideration of geographical location and time information contained in the image information.

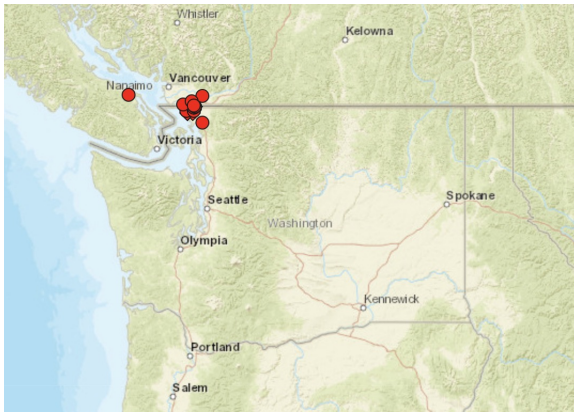

Figure 10: Positive ID Distribution

At the same time, because of the particularity of this problem, the particularity is that most of the reports are invalid, and the positive image distribution is concentrated in a certain area, so the simple image processing method will not achieve a particularly good effect.

Because the topic limits the sample size we use, we cannot use external data to optimize our model. Therefore, the existing models can not achieve very high prediction accuracy. In fact, most models only judge the new image as negative.

Based on the results obtained from the above time series analysis and univariate nonlinear regression, consider the update strategy of the database. According to the machine learning model improvement strategy, similarly, for the reported information, after the lab judges its status, all positives should be put into the database. This piece of data should contain information such as the location, time, and type of giant hornets. From the regression function, the update frequency should be higher than January.

Combining the model to determine whether the Asian Hornet has been eradicated from the following two aspects:

- Whether the frequency of positive has been reduced to 0.01.
- Whether there has been no new positive report in a long period of time.

When both are met, it can be concluded that the pest has been eradicated.

5 Resources allocation strategies

5.1 TLL Model

5.1.1 Model Introduction

TLL model is based on variables including time, longitude and latitude. We are interested in the distribution of each labstatus type, so we plot the several scatter diagram to figure out whether it has some laws of internal distribution.

Figure 11: Four types of labstatus distribution

First, we concern about the evolution of the positive ID's distribution. So we compare the positive ID distribution before 2020 and the distribution after 2020.

Figure 12: Distribution Before 2020

Figure 13: Distribution After 2020

we can find from the heat map that positive ID's location is being changed from time to time, so we give time as a factor t, if the latest positive ID detect time is far from current, then t will become less and if the latest positive ID detect time is close to now, then t will become greater which means this place has higher priority to allocate resource.

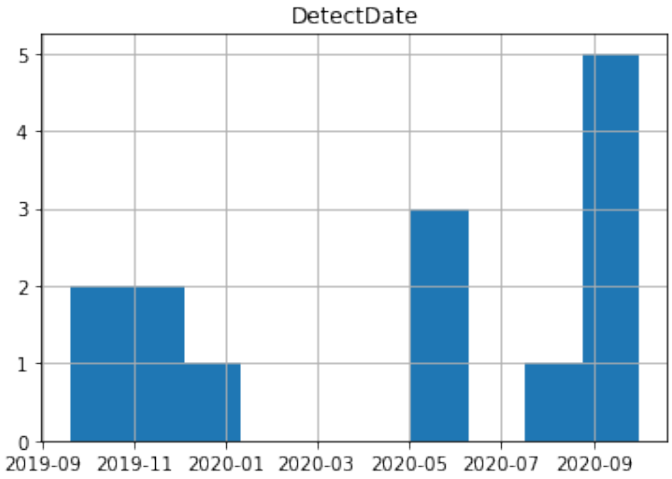

Figure 14: Detect Date of Positive ID

5.1.2 Model Formula

As mentioned above, there are three impact factors, and then we give the interpretation formula(which is the resource distribution formula as well):

when the new reports come, we need to use TLL model to evaluate the priority of each report base on their location and the time when the information detect.

For a new coming report, assume we have the DetectDate t_d, latitude Lat_d and longitude Lon_d of this report, we can find a positive ID happened place which is nearest to this location, whose Latitude is Lat_n and Longitude is Lon_n, this location's positive happening time is t_n. we calculate the distance Δd and DetectDate bias Δt.

$$\Delta t = t_d - t_n \tag{1}$$

$$MLat_d = 90 - Lat_d \qquad MLat_n = 90 - Lat_n \tag{2}$$

$$C = sin(MLat_d) * sin(MLat_n) * cos(Lon_d - Lon_n) + cos(MLat_d) * cos(MLat_n) \tag{3}$$

$$\Delta d = R * Arccos(C) * Pi/180 \tag{4}$$

$$p_i = \theta * f(\Delta t) * g(\Delta d) \tag{5}$$

$$f(\Delta t) = \frac{1}{\theta_1 * \Delta t} \tag{6}$$

$$g(\Delta d) = \frac{1}{\theta_2 * \Delta d} \tag{7}$$

5.1.3 Model Drawback

The TLL model is only based on the distance difference between time and geographic location. The model is too simple, and the final priority order is not very convincing. At the same time, it is easy to cause the resource allocation to be too concentrated in the area around the positive.

So we introduce a more complete model **TTLLP**.

5.2 Model revision:TTLLP

5.2.1 Model Introduction

TTLLP model is based on variables including time, topographic, longitude, latitude and population, which is an expanded version of the original TLL model.

Topographic depends on location, which is related to longitude and latitude.

We find from the paper that Vespa mandarinia typically build their nests inside of natural cavities such as hollow trees and sometimes inside the walls of buildings, so we divide terrain types into four terrains: plateau, mountain, plain and valley. And we give these four types different weights "To".

And due to the risk of human being attacked by Vespa mandarinia, we concentrate on the population density of Washington state, and we give population as a factor 'Po', 'Po' becomes greater if the place is densely populated.

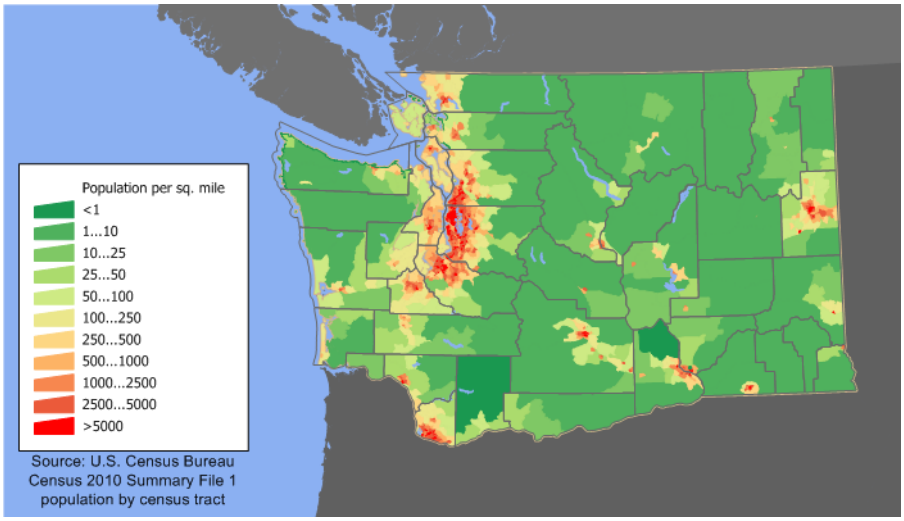

Figure 15: Washington Population Map[14]

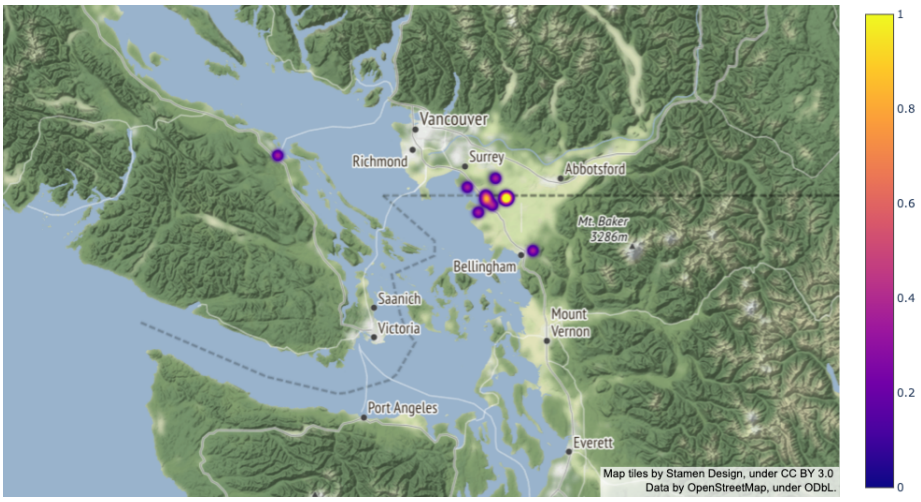

Figure 16: Topographic Map Of Washington State

5.2.2 Model Formula

Variable Announcement

- Po unit: thousand, for example, if the population of the area is 100000, then $Po = 100$.
- To unit: if the terrain type of the area is valley or plain, then $To = 1$, else $To = 0$.(Due to the concentration of positive areas are in river valleys and plains).
- $\theta, \theta_1, \theta_2, \theta_3, \theta_4$ is the normalized operator to ensure the order of magnitude is the same between the functions.

As mentioned above, there are three impact factors, and then we give the interpretation formula(which is the resource distribution formula as well):

when the new reports come, we need to use TLL model to evaluate the priority of each report base on their location and the time when the information detect.

For a new coming report, assume we already have the Δt and Δd which have been calculated in TLL model. And for this report, we can easily get it's terrain To from digital map by inputting it's latitude and longitude, also we can easily get this area's population Po.

so we come to the final resource allocation model:

$$p_i = \theta * f(\Delta t) * g(\Delta d) * h(Po) * k(To) \tag{8}$$

$$h(Po) = \log(\theta_3 * Po) \tag{9}$$

$$k(To) = \begin{cases} \theta_4, & \text{if To = 1,} \\ 1, & \text{else.} \end{cases} \tag{10}$$

The resource allocation strategy

We give the higher probability(p_i) the prior schedule to give out to the lab to figure out the status.

5.2.3 Model Renewal

Since TTLLP contains the relevant function for Δt, for each newly generated report, we will get its detection time, so as to obtain the new time function result, so the model will be given as time is updated new priority scheme.

At the same time for the update frequency of the model and the processing cycle of newly generated reports, On the basis of the results obtained from the above time series analysis and univariate nonlinear regression analysis, consider the update strategy of the database. According to the machine learning model improvement strategy, similarly, for the reported information, after the lab judges its status, all positives should be put into the database. This piece of data should contain information such as the location and time of discovery. From the regression function, the update frequency should be higher than one time a month. It better to put it in the intervals of two weeks to one month.

5.2.4 Model Particularity

Custom function We give the selection of the customized value of $\theta_3$in the population function $h(Po)$in TTLLP. If more people are found to be attacked by Vespa Mandarinia in

the follow-up survey, the value of $\theta_3$can be appropriately increased to increase the weight of population factors in the model

Similarly, if the geographical location of positive is found to be concentrated in valley and plain area in the follow-up survey, the weight of $\theta_4$can also be increased to increase the weight of topographic factors in the model.

6 Sensitivity Analysis

After the model successfully answered the above questions, we tested the robustness of the model. In the above, we used existing data, including positive, negative, unverified and unprocessed status, to predict the future situation. At the same time, check the current error. To verify the robustness of the model, we set up a group of control groups. We randomly selected 20 days in accordance with the period of its outbreak trend law. During these 20 days, cases which Labstatus was positive were detected every day. Substitute this set of data into our model for prediction and testing.

Through ACF and PACF to test the autocorrelation and partial autocorrelation of the residuals, it is not difficult to see that the obtained values are basically within the threshold. Shows that the results are very accurate and ideal. In addition, we also use Durbin-Watson statistic to test the correlation, and its value is 1.89. It can be considered that there is no first-order correlation in the residuals. That means it meets the modeling requirements. In other words, the conclusion is correct.

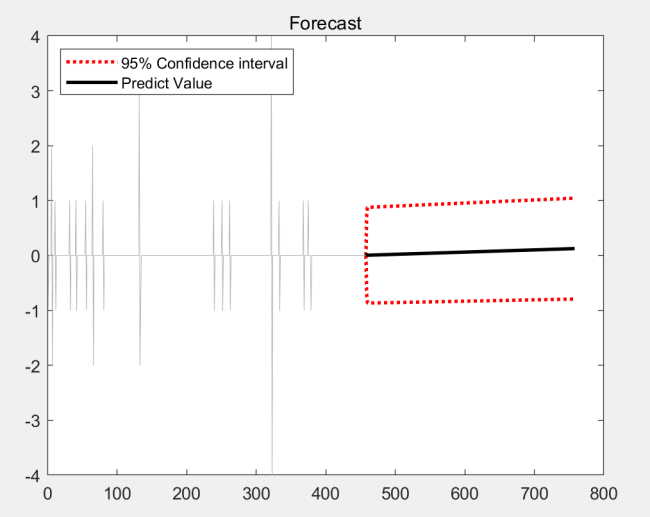

Figure 17: Topographic Map Of Washington State

As for its prediction results, we can see from the front picture that the development trend still maintains the same shape as the above conclusion. Therefore, it is concluded that the model has strong robustness.

7 Conclusions

Our team conducted modeling research on the two main problems and five main aspects required by the topic. Here we conclude our interesting findings.

- Through time series analysis, the development trend of Vespa mandarinia can be predicted. Accuracy can reach more than 86.4%.

- According to our convolutional neural network model calculation, most reported sightings mistake rate at least should be 96.2%
- TTLLP model uses the terrain, population, distance and time difference between the location and the nearest positive point as the classification basis, and provides a solution for priority allocation after comprehensive analysis.
- TTLLP model is a mixed model that includes time variables, and it has the ability to deal with the time problem contained in the newly added data. On the basis of the results obtained from the above time series analysis and univariate nonlinear regression analysis, it's update should occur in range of two weeks to one month.
- Combined with this model, if the latest data show:
  1.The frequency of reporting positive has been reduced to 0.01.
  2.The time span should meet at least two life period without new positive reports.
  Combining the above two points, it can be concluded that the pest has been eradicated.

8 Strengths and weaknesses

8.1 Strengths

- **TTLLP customized function** According to the follow-up development, the model parameters can be manually adjusted for a specific situation, so as to achieve a better optimal allocation effect.
- **Time series model**
  The ARMA model can predict a relatively accurate trend in the time dimension
- **Intuition and completeness**
  Univariate nonlinear regression model, whose regression curve changes dynamically with the increase of data. More intuitive and clear.

8.2 Weakness

- **Data restrictions**
  A limitation of this report is the data constraint, which reduced our available information to TTLLP and ARMA-UNLR data provided. While we were able to derive additional variables from the initial data and include map data from outside sources, our exploration was hindered by the restriction to given socioeconomic factors.
- **Time series model**
  The ARMA model can predict a relatively accurate trend in the time dimension.
- **Difficulties of image recognition**
  The amount of data required for model training is relatively large, but the actual existing data set is not enough. Therefore, if an efficient image classification method based on small samples can be found in the future, the interference of invalid information to government departments can be reduced.

9 Memo

To:Washington State Department of Agriculture (WSDA)
From:Team # 2107239
Date:2021/2/5
Re:Research Results About Asian Giant Hornet

Dear WSDA, we are honored to describe the the research result of Asian giant hornets for you.

Our team has analyzed the given data and conceived several models that successfully interpret the public reports data and develop the resources allocation strategies for controlling the ravages of giant hornets. To better understand the crisis and any opportunities we have to reduce its impact. We explored current research and investigated the data processing and giant hornets colony prediction within our model.

Results

After the careful study of the public reports. Firstly, we utilized the MySql database to narrow down the data. And then obtain a database with only image files by using python opencv tool to extract key frames in the video data. Thus we can put processed data into cnn model to predict the likelihood of a mistaken classification. And in order to most test the reported sightings mistake rate. We aim to build a convolutional neural network as a classifier so that we can recognize newly emerging images and make supportive decisions to save limited resources in the end put all of these into consideration.

Next, we use Generative Adversarial Network(GAN),a basic image preprocessing way to resize, blur the images which can get the clearing set to train the model, to extend existing samples while ensuring that no new external data is introduced. Combine the model can we determine whether the pest has been eradicated. After verifying the model, we found that Convolutional Neural Network(CNN) can only classify images weakly. In order to improve this weakness, we processed the data separately, and corrected the model using gray forecasting principles and other methods.

Secondly, in order to figure out the development trend of Vespa mandarinia, and the accuracy of prediction. We also made predictions about future trends which is based on the pest's natural life cycle, combined with relevant policies and measures. we conducted the Autoregressive Moving Average(ARMA) and Univariate NonLinear Regression models(AMRA-UNLR) analysis to analyze and summarize the existing information, and finally get the development trend. After we established the model, we After we established the above model, we used Autocorrelation Function(ACF) and Partial Autocorrelation Function (PACF) for accuracy analysis. And the result shows the accuracy of our prediction can reach more than 86.4%.

Thirdly, in order to make the work better and the "insect pests" to end as soon as possible, We built an model to help the government allocate resource. So we start with proposing a model called Time-Longitude-Latitude Model(TLL), we focus on the time and distance of each upcoming new report, however the model is so simple that it is ineffective. So we provide a revised model called Time-Topographic-Longitude-Latitude-Population Model(TTLLP) in which involving population and terrain elements. which finally provides a solution for priority allocation after comprehensive analysis. And so far we have solved the problem of allocating

resources.

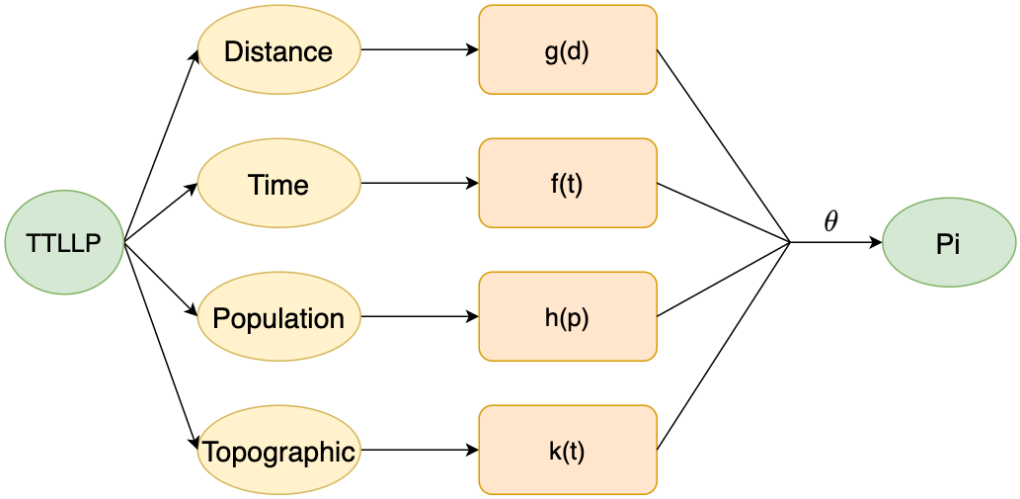

Figure 18: TTLLP FlowChart

What's more. Our model can update its database as the time going and the update frequency should be once in a month. the government's active measures will expel this pest from Washington State.

Proposal

As we analyzed in our model, The following proposals are our conclusions:

1. Control the update frequency to once a month to ensure the use of the TTLLP model.
2. Judging from the report, the misjudgment rate of the masses is relatively high. It is recommended to increase the popularity of relevant information.
3. Use the following two indicators as criteria for judging that pests have been eradicate:
   1).The frequency of reporting positive has been reduced to 0.01.
   2).The time span should meet at least two life period without new positive reports.

4 Incorporate factors such as time into the use of the model, and add relevant data to optimize the model results.

References

[1] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25: 1097-1105.

[2] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[3] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. arXiv preprint arXiv:1406.2661, 2014.

[4] http://www.kumifeng.com/zhuanti/hutoufeng/651.html

[5] ”Research on the spreading model and prediction of alien species.”

[6] 2021MCM_ProblemC_Vespamandarina.

[7] https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average

[8] https://en.wikipedia.org/wiki/Augmented_Dickey%E2%80%93Fuller_test

[9] "Archived copy". Archived from the original on March 2, 2009. Retrieved April 2,2008.

[10] "Testing the null hypothesis of stationarity against the alternative of a unit root."

[11] Significance of ACF and PACF Plots In Time Series Analysis.

[12] https://en.wikipedia.org/wiki/Partial_autocorrelation_function

[13] https://towardsdatascience.com/gan-ways-to-improve-gan-performance-acf37f9f59b

[14] https://commons.wikimedia.org/wiki/File:Washington_population_map.png

Appendices

Here are simulation programmes we used in our model as follow.

Input Python source:

```
import torch.nn as nn
import torch


class AlexNet(nn.Module):
    def __init__(self, num_classes=1000, init_weights=False):
        super(AlexNet, self).__init__()
        self.features = nn.Sequential(
            nn.Conv2d(3, 48, kernel_size=11, stride=4, padding=2),
            nn.ReLU(inplace=True),
            nn.MaxPool2d(kernel_size=3, stride=2),  # output[48, 27, 27]
            nn.Conv2d(48, 128, kernel_size=5, padding=2),  # output[128, 27, 27]
            nn.ReLU(inplace=True),
            nn.MaxPool2d(kernel_size=3, stride=2),  # output[128, 13, 13]
            nn.Conv2d(128, 192, kernel_size=3, padding=1),  # output[192, 13, 13]
            nn.ReLU(inplace=True),
            nn.Conv2d(192, 192, kernel_size=3, padding=1),  # output[192, 13, 13]
            nn.ReLU(inplace=True),
            nn.Conv2d(192, 128, kernel_size=3, padding=1),  # output[128, 13, 13]
            nn.ReLU(inplace=True),
            nn.MaxPool2d(kernel_size=3, stride=2),  # output[128, 6, 6]
        )
        self.classifier = nn.Sequential(
            nn.Dropout(p=0.5),
            nn.Linear(128 * 6 * 6, 2048),
            nn.ReLU(inplace=True),
            nn.Dropout(p=0.5),
            nn.Linear(2048, 2048),
            nn.ReLU(inplace=True),
            nn.Linear(2048, num_classes),
        )
        if init_weights:
            self._initialize_weights()

    def forward(self, x):
        x = self.features(x)
        x = torch.flatten(x, start_dim=1)
        x = self.classifier(x)
        return x

    def _initialize_weights(self):
        for m in self.modules():
            if isinstance(m, nn.Conv2d):
                nn.init.kaiming_normal_(m.weight, mode="fan_out", nonlinearity='relu')
                if m.bias is not None:
                    nn.init.constant_(m.bias, 0)
            elif isinstance(m, nn.Linear):
                nn.init.normal_(m.weight, 0, 0.01)
                nn.init.constant_(m.bias, 0)
```