Problem Chosen	2021 MCM/ICM Summary Sheet	Team Control Number
C		2113014

A Control Solution for Vespa Mandarinia Based on Time Series Analysis and Broad Learning System

Summary

Based on the actual characteristics of the Vespa mandarinia in Washington State and the analysis of existing data, we provide a reference solution for the Giant Hornet prevention and control in Washington State on the premise of rationalizing the use of resources and improving the effectiveness of the solution.

Firstly, we preprocessed the data. We divide the obtained data into table data and image data for processing. For tabular data, we found and cleared some redundant information that was useless for work; for image data, we performed data cleaning and data conversion, and only retained the image type data.

Next, we selected the report data of the area of Washington State where the Vespa mandarinia has been confirmed, and applied **ARIMA** to the time series prediction of the new variables that we defined representing the geographic location. We found that the pest did not spread on a large scale over time. It has something to do with their characters

Then, we analyzed the image data labeled as positive and negative in the Washington area, but the data ratio is seriously unbalanced. Therefore, we use Data Augmentation methods to randomly generate positive data, and use an efficient and fast **Broad Learning System (BLS)** to classify the resulting data set, which achieved good results. In later work, we also rely on the Residual Network-based **Stacked Broad learning System (SBLS)** to update the model, and according to the seasonal analysis of the swarm activity, determine the model update frequency, which provide us a model updating strategy

Due to the limited resources available to the government, for reports from different regions, We have established a scoring system based on factors including prediction results, geospatial and historical detection to determine the priority of the investigation in different regions degree,which will be the basis for the priority investigation.

Finally, we used the results and models of the above work, combined with the method of exterminating Vespa mandarinia, to determine the basis for us to judge the eradication of the Asian Giant Hornets in Washington state,and provide a low-cost and high-reliability reference solution for Washington State.

Key words:ARIMA ; Broad Learning System;Update Strategies; Judgements

Contents

1Introduction

1.1 Restatement of problems

Vespa mandarinia is the largest species of hornet in the world.It is very toxic,which is equivalent to a poisonous snake.If people are stabbed repeatedly by its sting,it can cause human death.Additionally,it is a predator of European honeybees,which is capable of invading and destroying their nests.Vespa mandarinia can destroy a whole colony of European honeybees in a short time.In the fall of 2019,a nest and workers of Vespa mandarinia were discovered on Vancouver Island.Because the number of bee populations in the United States is declining,people worry that the invasion of Vespa mandarinia will wipe out honeybee populations.

Due to the potentially serious impact of wasps on the local honeybee populations,The State of Washington has established a helpline and a website for people to report sightings of these hornets.

We solve the current problems through the following work:

- In order to study the relationship between pest spread and time, we used ARIMA model to analyze the relevant data.
- Based on the BLS,we established pest classification model and provide a suitable update and maintenance program for it.
- We established a priority scoring system based on predictions, geography, and historical reporting factors to provide the government with a reference for investigating resource allocation.
- Based on the results of the model and the literature, we put forward the judgment basis for the complete eradication of the Asian giant wasp in Washington State.

2 General Assumptions and Variable Description

2.1 Assumptions

In order to simplify our models, we make some assumption in our paper. The details are as below:

- In a small area,the influence caused by the surface of the earth is almost negligible,so we approximate this area as a flat area.On this assumption,we calculate the azimuth and distance between two locations.
- In the data sets and data tables we have obtained,there are no identification errors caused by human factors.
- In a certain period of time, the event has and maintains sufficient attention and influence in the local society, and has a reliable amount of data reported by the public to ensure that the data obtained truly reflects the situation of the event in this period of time, and will not cause obvious data omission due to other factors.

2.2 Variable Description

Symbol	Explanation
bp_f	The first positive detaction location point in Washington state
d	The distance from bp_f
θ	The angle between the line from bp_f to the subsequent point and the direction of the meridian
la	The latitude label of data
lo	The longitude label of data
N_1	The number of nodes in each window of the Feature mapping layer
N_2	The number of Feature mapping layer windows
N_3	The number of enhancement layers nodes
i	The i th block of Stacked-BLS
$Acci$	The accuracy of the i th block of Stacked-BLS
F_1	The monthly average number of reports at queens' stage outside of the nest
F_2	The monthly average number of reports at workers' stage outside of the nest
F_3	The monthly average number of reports at over-wintering stage
p	The predicted value of the Broad Learning System
num	The number of 'negative' feedback received within 50km of the area

3 Data processing

The works we have done in data processing are as follows.

3.1 Data Classification

We classify the data into image data and table data for processing.There is a corresponding relationship between image data and table data.

3.2Data Screening and Inspection of Value involved

In table data,we find some useless redundant information,such as Note and Lab Comments,which have no effect on subsequent model.So we delete they.And we use "screening" and "positioning control" operations in Excel to check for missing values in the remaining data.After that,we don't find the existence of missing values.

3.3Data Cleaning and Data Conversion

For image data,in order to facilitate model building and training,we maintain the ***.jpg*** files , ***.png*** files and ***.jpg*** files which are decompressed from the ***zip*** file.And we delete Others files which are not conductive to training the model,such as ***.mp4*,*.mov*** file.According to the "GlobalID" of the cleaned file,we extract the corresponding information from the data of the corresponding two tables and merge them into one table(the relevant codes are shown in Appendix A).Then we check that there are multiple duplicate "GlobalID".After comparing with the corresponding images,we decide to maintain the data with repeated "GlobalID" to improve the accuracy of the model.For the "Lab Status" of each data,we replace the "positive ID" with **1** and replace "negative ID" with **0** respectively.**0** and **1** can represent tags for data to facilitate loading model training.

4 Model Establishment and Solutions

4.1 Model-1 Prediction of Vespa mandarinia Transmission Based on ARIMA Model

4.1.1 Model Establishment and Solving

Based on the analysis of the given data and problems,we find that most of the positive detection took place in British Columbia in Washington state.so we use report datas from areas where the Asian Giant Hornet has appeared,that is,we analyze the data whose status is "positive ID"on the data label "Lab Status".By tracing the data of "positive ID" on the map, we find that they are mostly concentrated in a range of about 50 kilometers,so we use time series analysis to predict the location of the data of "positive ID" in order to describe the possible transmission of the Vespa mandarinia in this small area.The modeling process is as follows.

First,We have learned that the first colony of Vespa mandarinia appeared in Nanaimo on Vancouver Island, which is unrelated to the first colony of Vespa mandarinia appeared in Washington state. That is to say, the Asian Giant Hornets that appeared in Washington were newly introduced, not transmitted by the Hornets on the Nanaino. Therefore, we will exclude this report data.

Second,in order to indicate the location of the data of "positive ID",we define two variable sequences,$\boldsymbol{d}$ and $\boldsymbol{\theta}$,and set the location where the first Asian Giant Hornet appeared in Washington as the base point $\boldsymbol{bp_f}$. For $\boldsymbol{bp_f}$, we will set its $\boldsymbol{d}$ and $\boldsymbol{\theta}$ to zero respectively.And it does not participate in modeling.

The formula of $\boldsymbol{d}$ and $\boldsymbol{\theta}$ are as follow:

$$d = \sqrt{\left[\left(la_1 - la_2\right)\times 111\right]^2 + \left[\left(lo_1 - lo_2\right)\times 111\times\cos\left(49^\circ\right)\right]^2}$$

$$\theta = arc\tan\left\{\left[\left(la1 - la2\right)\times 111\right]\div\left[\left(lo1 - lo2\right)\times 111\times\cos\left(49^\circ\right)\right]\right\}$$

Third,we revert the predicted distances and angles to latitude and longitude information.Because the detection time interval of the given report is not uniform, we do not consider the specific time of the reports.Instead, we redefine a natural number sequence from

1 to 20 as a time measurement, in which **[1,10]** are real data, and **[11,20]** are data to be predicted and analyzed.

Forth,we find that the two sequences are both stationary through model recognition.And then,by optimizing their **Akaike Information Criterion(AIC)**,We calculated the ARIMA model parameters ***p d q***, which are **ARIMA (4,0,0)** and **ARIMA (4,0,0)** respectively.

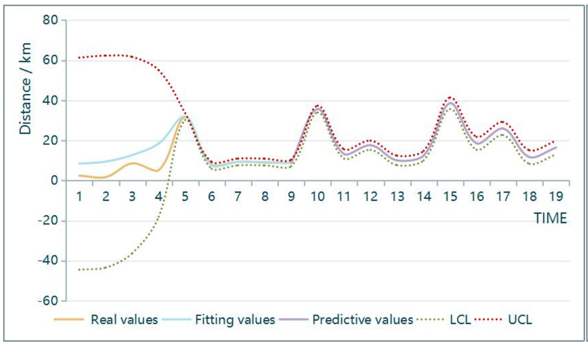


Figure4-1:ARIMA Model of ***d***

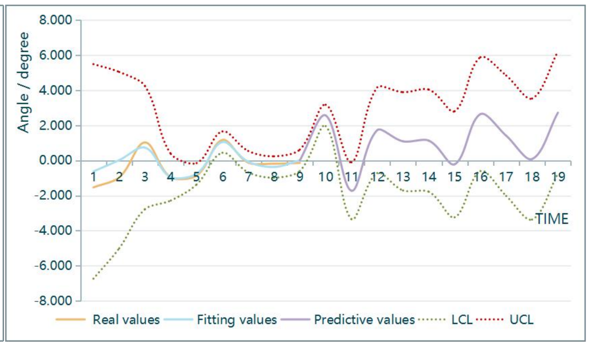


Figure4-2:ARIMA Model of ***θ***

4.1.2 Analysis and Evaluation of results

Through the analysis of fitting and prediction results, we got the fitting and prediction curve as shown in **Figure 4-5** and **Figure 4-6** ,as well as the metrics shown in **Table4-1**

Table4-1: the analysis data of ARIMA model

	ARIMA(4,0,0)	**ARIMA(4,0,0)**
R^2	0.536	0.706
RMSE	9.826	0.811

In statistics,**the coefficient of determination(R^2)** is an metrics to measure the effectiveness of regression model.If **R^2** is closed to 1 ,it means that the regression or prediction model fits the data well. The value of **R^2** of ARIMA model of ***d*** and ARIMA model of ***θ*** are both between 0.5 to 0.8,which indicates that The model has a good fitting effect on the data.In addition,according to the residuals ACF graph and PACF graph as shown in **figure4-3,figure4-4,figure4-5,figure4-6**,we can know that they both have low value of autocorrelation and partial autocorrelation,which means the model is effective.

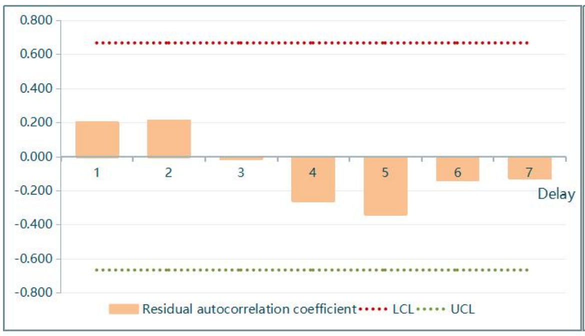
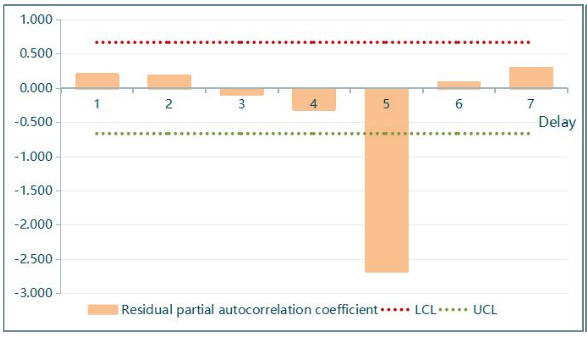


Figure4-3:Residual ACF of ***d***

Figure4-4: Residual ACF of ***θ***



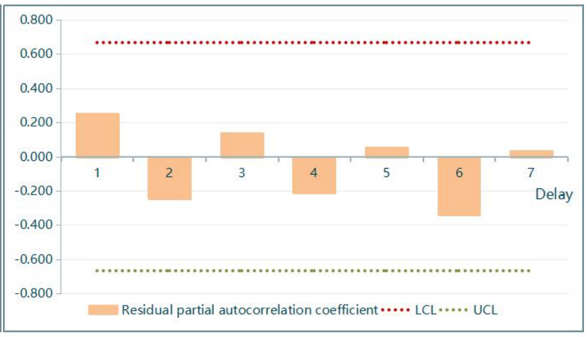


Figure4-5:Residual PACF of ***d*** **Figure4-6**:Residual PACF of ***θ***

After that,we mark the latitude and longitude we calculate on the map.

We restore the predicted value of ***d*** and ***θ*** to the longitude and latitude coordinates.

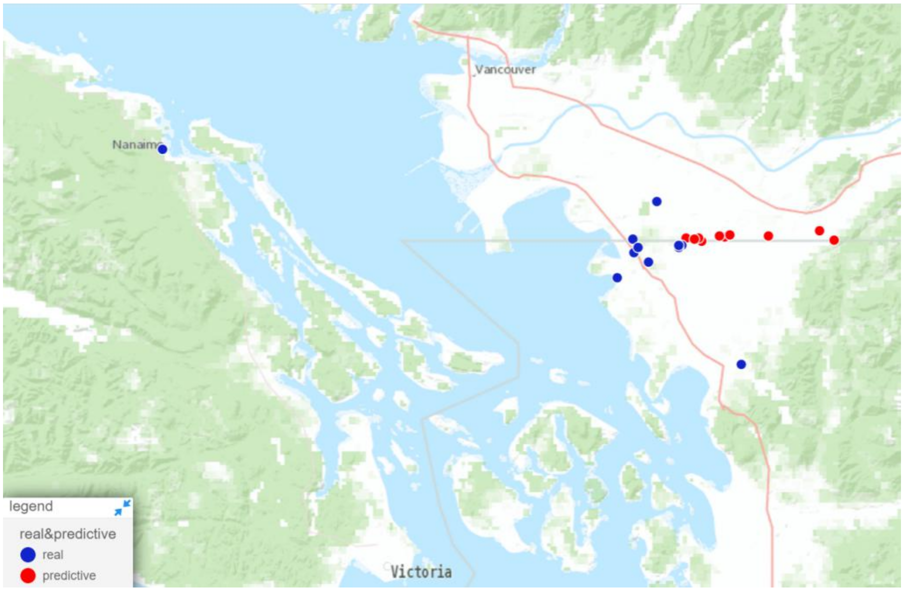


Figure 4-7:Map showing our calculated and actual values

The **Figure4-7** indicate that**:** Although Vespa mandarinia spread to the southeast basically, it is going to spread in a small range and small scale, and there is no obvious law of large-scale spread. Finally, we come to the following conclusion**:** The transmission of Vespa mandarinia at the current stage is not strong. Based on the time and area of the discovery, Vespa mandarinia were found along the west coast between May to October. This is also related to the fact that the species is more active in summer and autumn, and the temperate oceanic climate of the west coast is similar to the temperate/subtropical monsoon climate characteristics of its source region.

4.2 Model-2 Pest Classification Model Based on Broad Learning System

4.2.1 Problem Analysis and Image Data Pre-processing

Based on the analysis of the problem and the given data, we decided to use the neural network to classify the data pictures and get the accuracy of the classification.In this way,we can solve the problem of predicting the error classification probability. Because the data-set size is too small, it is not suitable for complex neural networks, such as **CNN**, **ResNet**.Therefore, we select an efficient and fast neural network, **Broad Learning System**.

Broad Learning System is a new discriminant learning algorithm proposed by **C. L. Philip Chen** and his students in 2018. It accepts input $\boldsymbol{x}$ and generates feature mapping nodes and enhancement nodes, and connects with output $\boldsymbol{y}$. Broad Learning System has better computational performance and scalability, lower computational costs and higher classification accuracy than common Deep Learning Network when dealing with small amounts of data.This allows us to use the Broad Learning Network to train and predict the image data quickly, and calculate the prediction accuracy.

We selected the image data , whose "Lab status" label are"positive ID" and "negative ID", as the data-set needed for modeling.In previous data pre-processing, we labeled "positive ID" and "negative ID", corresponding to **1** and **0**. And we used them as labels for images.At the same time, we resized all images to **200*200**.Then, we divided the data into training sets and test sets at a ratio of **5:2**.

However, the volume of data labeled 1 is too small (only 14 photos).The number of different categories of data in the data-set is severely unbalanced, which seriously affects the training accuracy of the model.So we performed **Data Augmentation** method on the data labeled 1 in the training set and testing set.We used **keras API in python** for image generation, using the **Class ImageDataGenerator** to randomly rotate, horizontally flip, shift, scale, and shear the image.After that,we generate 9 derivative images for each image.

To simplify the operation better, we convert the **three-channel image data** into **a single-channel grayscale image.**

Here is the data-set we got and one example of generated images in data-set.

Table 4-2: Our Data-set

	Training Set	**Test Set**	**Total**
Num of class 1	100	40	140
Num of class 2	2220	885	3105
Total	2320	925	3245
Size		200×200	
Type		Grayscale Images	

Figure4-8: Generated Images Example in Data-set

4.2.2 Model Establishment and Solving

Broad Learning System was constructed based on random vector functional-link neural network(**RVFLNN**).It includes the input layer,feature layer,enhancement layer and output layer.The input layer obtains each group of the group of feature nodes by multiplying with the sparse weight obtained by the alternating direction multiplier method **(ADMM)** optimization.Here is the following formula.

$$Z_i = \varphi(XW_{ei} + \beta_{ei}), i = 1, \ldots\ldots, n$$

$$\begin{aligned}
&where: \\
&\quad \varphi \ is\ linear\ function \\
&\quad Z_i \ is\ the\ i\text{-}th\ group\ of\ feature\ nodes \\
&\quad W_{ei} \ is\ sparse\ weights \\
&\quad \beta \ is\ randomly\ genarated
\end{aligned}$$

And the feature layer Z^n is shown below.

$$Z^n = [Z_1, \cdots, Z_n]$$

Enhancement nodes are generated by feature nodes.The enhancement nodes $\boldsymbol{H_m}$ is shown below.

$$H_m \equiv \zeta\left(Z^n W_{hm} + \beta_{hm}\right)$$

$$\begin{aligned}
&where: \\
&\quad \zeta \ is\ linear\ function \\
&\quad W_{hm} \ is\ randomly\ generated\ weights \\
&\quad \beta \ is\ randomly\ generated
\end{aligned}$$

The enhancement layer H^m and P are :

$$H^m = [H_1, \cdots, H_m]$$

$$P = [Z^n \mid H^m]$$

And we got AW=Y. The loss function can be optimized by ridge regression algorithm.

$$\underset{W}{\arg\min}: \|PW - Y\|_v^{\sigma_1} + \lambda \|W\|_u^{\sigma_2}$$

The optimal solution can be obtained by generalized inverse.

$$\begin{aligned}
&W = P^{+}Y \\
&P^{+} = \lim_{\lambda \to 0}\left(\lambda I + PP^T\right)^{-1} P^T Y
\end{aligned}$$

So the optimal solution **W** is :

$$W = \left(\lambda I + PP^T\right)^{-1} P^T Y$$

Finally, the weights of BLS network are trained by these formulas.

We read the training set data, the training set label, the test-set data and the test-set label respectively, and process them into the corresponding data matrix, and input them into the Broad Learning System algorithm as parameter components.The Broad Learning System takes the training set data as the input parameter, and the training set label as the expected output to train the corresponding network weight, which shows the corresponding training accuracy.Then, with the trained weights as the system weights, the system validates the data and labels of the test set and shows the corresponding test accuracy to reflect the classification effect of the network on the data-set.

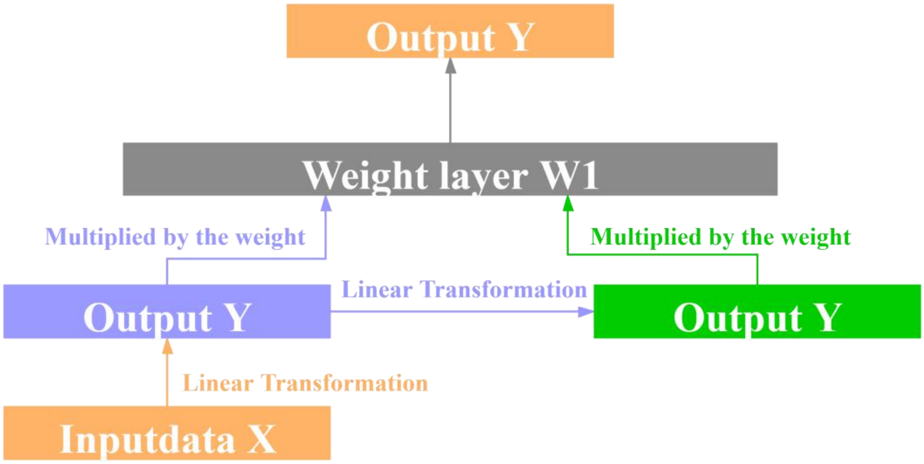


Figure 4-9:the structure of BLS

4.2.3 Analysis and Evaluation of results

In order to test the effect of the Broad Learning System on the data set, we use different parameters to train and test the Broad Learning Network.By adjusting the **N1, N2, N3** parameters, we get different results.Here are the results.

Table 4-3: Classification accuracy for different number of feature nodes (partial data)

No.of Model	N_1	N_2	N_3	Training Accuracy(%)	Train Times(s)	Testing Accuracy(%)	Testing Time(s)
1	5	5	500	97.111	3.1582	95.351	0.1166
2	10	10	500	97.155	10.7299	95.243	0.3457
3	20	20	500	97.585	42.2484	95.568	0.7985
4	50	50	500	100	270.1907	92.649	2.9311
5	70	70	500	100	2213.517	91.892	16.6887

Table 4-4:Classification accuracy for different number of enhancement nodes (partial data)

No.of Model	N_1	N_2	N_3	Training Accuracy(%)	Training Accuracy(%)
1	10	10	50	95.687	95.676
2	10	10	100	95.687	95.676
3	10	10	200	95.903	95.784
4	10	10	800	98.619	94.703
5	10	10	1000	99.569	93.189

The training accuracy and test accuracy of the Broad Learning System measure the correctness of the system for the classification and prediction of the data set.

From **Table4-3** and **Table 4-4**,we preliminarily reached the following conclusions:

- When the number of enhancement layer nodes is constant,if the number of windows in the feature mapping layer and the number of nodes in each corresponding window of the feature mapping layer tend to be higher,the training accuracy improvement degree of the

Broad Learning System is relatively smaller,and the overall testing accuracy interval is between **95.1%** and **95.8%**.But it costs more time of calculating.

- With the change of the number of nodes in the enhancement layer, the network will achieve the highest testing accuracy around ***N3*=200**.However, too many enhancement nodes and too many feature nodes have a negative impact on the data, and the model has a phenomenon of over-fitting, which reduces the accuracy of the correlation and the effect of classification.
- For the current discriminant data-set of Vespa mandarinia, the Broad learning system model we used maintains good classification accuracy and computational performance near ***N1*=10**, ***N2*=10**, and ***N3*=200**, with the lowest possibility of classification errors.

4.3 Model-3 A Scoring System for investigating the priority of detection reports

4.3.1 Problem Analysis

In order to determine the criteria for conducting a priority survey of the reports received, a scoring system was established by selecting multiple indicators from “Unprocessed” and “Unverified”.Although “Unverified” has a large amount of data, only a portion of the information of data contains corresponding images in the data-set.Therefore,we only select the information of data containing the corresponding image for scoring.

First, we use the trained Broad learning system model to predict images which without classification labels from “Unprocessed” and “Unverified” , and get the probability that the report is positive.Although this indicator can preliminarily judge how to determine the standard of priority investigations, it can not well reflect the comprehensiveness of the standards.Therefore,we also need to consider other factors related to geographic location and historical information about sighting reports.

We consider that the spread of Vespa mandarinia is related to distance.The closer to the confirmed location of Vespa mandarinia, the more likely it is to be positive and the more priority should be given to investigation.We use the distance between the sites of the reports from “Unprocessed” and “Unverified” and the earliest pest occurrence sites as a negative indicator to show the impact of the distance between the area and the confirmed pest occurrence area on the the priority of the survey.

In addition, based on the consideration of historical report,we consider that the number of negative reports received in the past in a certain area also affects the priority of the survey.Theoretically, the more negative feedback received in the past in a certain area, the less likely the new report from the area is positive.So we take the total number of negative reports received in the past within 50km of a certain location as a negative indicator.

4.3.2 Model Establishment and Model Solving

- the value that x_{ij} standardizes X_{ij}

Because the dimension and magnitude of each evaluation index are different, we should first standardize them.

$$X_{ij}=\frac{x_{ij}-min\{x_{ij},\cdots,x_{nj}\}}{max\{x_{ij},\cdots,x_{nj}\}-min\{x_{1j},\cdots,x_{nj}\}}$$

where :

xij is j-th original value of i-th evaluation index

$\max\{x_{ij}\}$ *and* $\min\{x_{ij}\}$ *are maximum and minimum of indicators*

- The proportion of i-th evaluation index to the sum of j-th evaluation **P_{ij}**

$$P_{ij}=\frac{X_{ij}}{\sum_{i=1}^{n}X_{ij}},i=1,\cdots,n,j=1,\cdots,m$$

- Information entropy value of j-th evaluation index **E_j**

$$E_j=-K\sum_{i=1}^{n}P_{ij}\ln(P_{ij})$$

$$K=\frac{1}{\ln(n)}$$

- Information entropy redundancy **D_j**

$$D_j=1-E_j$$

- Weights of Indicators **W_j**

$$W_j=\frac{D_j}{\sum_{j=1}^{m}D_j}$$

- Comprehensive score for each sample **S_i**

$$S_i=\sum_{i=1}^{m}W_j\bullet P_{ij}$$

4.3.3 Analysis and Evaluation of results

According to the formulas on **4.3.2**,we can calculate the weight of each indicator and the priority evaluation score of each reporting location of "Unprocessed" and "Unverified" respectively.

We sorted each reporting location according to the reporting time, and used the longitude and latitude of the location to represent the location. All the score data are shown in the attached list.

Table 4-5:The Weight of each Evaluation index

The Evaluation index	Weight
values	0.2668

d	0.1812
num	0.5520

Table 4-6:Some values and their Score Table

No.	*la*	*lo*	*values*	*d**	*num*	Total score
1	46.7144	-122.9834	0.266	253.8515	93	2.2236
2	47.804	-121.549	0.2687	156.5180	217	2.0887
3	48.7238	-122.3544	0.3123	39.2490	229	2.3683
4	46.9006	-122.2743	0.254	234.4335	271	1.8378
5	48.9392	-122.6613	0.2376	6.7634	182	2.1888
6	47.5861	-122.0073	0.2552	164.2604	823	0.8103
7	45.7359	-122.6456	0.3603	361.6631	66	2.5527
8	48.9102	-122.3571	0.2795	26.7963	183	2.3384
9	48.0542	-122.1566	0.2006	111.6216	556	1.1762
10	47.829	-122.0981	0.1919	136.5805	745	0.7349

*d** refers to the distance between the point and the center of each positive point, and the longitude and latitude of this center are generated by the average longitude and latitude of each point.

Based on our results,we reached the following conclusions:

- Among the three evaluation indicators, the number of negative feedback received in the vicinity of the reporting point has a significant impact on the priority score of the region.The Broad Learning System ranked second in the importance of the predicted image classification values. The distance between the report site and the first witness confirmed the location of Vespa mandarinia was less important.
- The higher the priority score, the greater the likelihood that pests are present in the region, and the greater the need to investigate the priority allocation of resources.
- Considering that in previous models, we found that Asiatic giant wasps are still in small-scale spread and do not have a trend of large-scale spread, this scoring system will be effective for rapid analysis of reports and provide a reference for the government in investigating priority decisions.

4.4 Classification Model updating Strategy

4.4.1 Model updating method based on Stacked Broad Learning System

(1)The analysis of the problem

With the growth of time, the number of reports received by the government and the amount of data that needs to be classified will be more and more.The original Broad Learning System has no significant effect on more complex data,so we use **stacked BLS**, a Broad Learning System with **ResNet**.It forms structure like Resnet by stacking multilayer BLS,and has better classification effect for more complex data.

We assume that the new data received to update our model have been labeled as positive or negative by laboratory,and they are not in “Unprocessed” or “Unverified”.

(2)The establishment and analysis of the model

①The introduction of the model

Stacked is a framework based on **BLS**.In **Figure4-10** ,there are two **BLS** blocks stack structures(2 blocks).The first block is equivalent to the original **BLS**, where the input and expected output are the original data **X**, **Y**, and the input of the subsequent block is the output U_i of the previous block,

$$\begin{cases} X_1 = X \\ X_i = g(U_{i-1}), i = 2, \cdots, n \end{cases}$$
$$\text{where :}$$
$$g \text{ is identify function}$$

TThe expected output is the **residual error** of the previous block:

$$Y_i = Y - \sum_{k=1}^{i-1} U_k$$

After multilayer residual approximation, we can get the final output:

$$Y \approx \sum_{k=1}^{n} U_k$$

Compared with traditional deep network, this cascade residual approximation method requires much less calculation, and has higher computational efficiency and speed.At the same time, Stack BLS has better fitting effect and classification accuracy than the original Broad learning system.Different blocks in Stack BLS can adjust hyperparameter for different task needs and data-sets to improve the generalization ability and adaptability of the model.

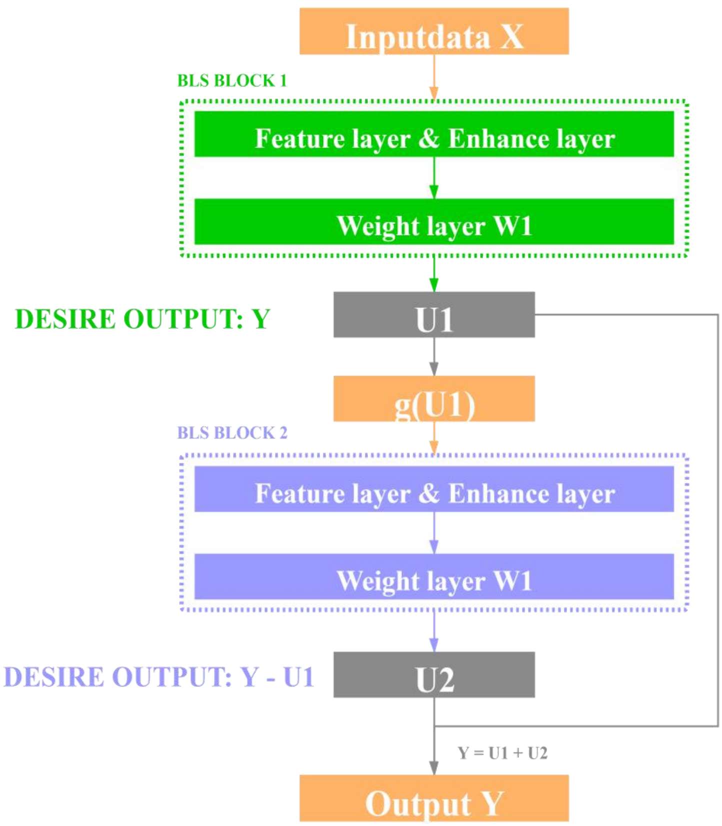


Figure 4-10:the structure of **Stacked-BLS**

②The method of updating the model

Although the number of sample sighting reports submitted in Washington State has increased over time, it is uncertain whether the original classification model (**Model 2**) is valid for new data.

To address the question of whether and how to update the model, we analyze the accuracy of each block and update the model through the following steps:

Step1:All the image data are put into two **Stacked-BLS** with two layers for classification analysis. If the accuracy of the second block is less than that of the first block, no model updates will be made.Otherwise, go to Step2.

Step2:Continue training and testing in i-th (block I i, I = 3,4,5...) block.Repeat this step if the block's testing accuracy is higher than the previous one, otherwise go to Step3.

Step3:Update the model to layer i-th **Stacked-BLS**.

The strategy solves the problem of if to update the model and the degree of model update.It can classify image data of Vespa mandarinia more effectively.

③The analysis of the results

Due to the limitation of the amount of data, we re-divided the original picture data-set into training and testing set randomly and proportionally.In order to verify the validity of Stacked-BLS, we only select three layers of **Stacked-BLS** to test. The results are as follows.

Table4-7:Results of validating **Stacked-BLS**(partial data)

	N1	N2	N3	Training Accuracy(%)	Test Accuracy(%)
Block 1	10	10	3000	100	85.996
Block 2	5	5	470	100	86.678
Block 3	2	7	6	95.688	95.676

The results show that although the second block has no change in accuracy compared with the first, the possible reason is that the data-set is severely unbalanced, which affects the classification results.

4.4.2 Update Frequency Analysis Based on the seasonal activity of pests

①seasonal activity of pests

The activity of Vespa mandarinia is seasonal and has different characteristics in different periods.In Spring and Summer,when the climate is mild,workers will go out to look for food.In winter,when the climate is cold, new queens will spend the colder months overwintering.Based on these information,we can figure out behavioral regulars of pests,which reflects seasonal changes in the number of sighting reports by the public. Therefore,we can calculate how often the model is updated in different time periods.

Based on the data we consulted, we roughly divided the pattern of pests' activity into three stages during the year.

Table4-8: the pattern of pests' activity

Stage	activity
Dec to Mar	New queens spend the colder months overwintering,they seldom go out.
Apr to Jul	New queens go out for building a nest, foraging, laying eggs, and caring for

	young.At that time,the number of sighting reports begins to rise.
Aug to Nov	Many workers begin to go out for forage,covering a range of 2km to 8km.The probability of public witnessing to the colony is further increased.

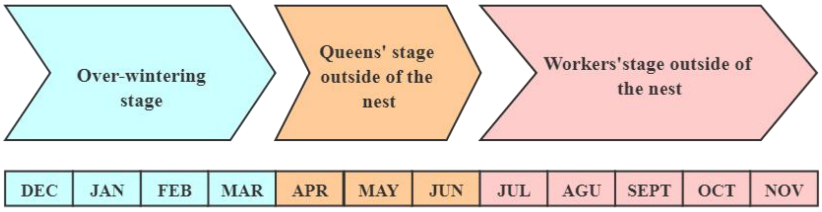


Figure 4-11:the stages of pest's pattern which we divided

②Determination of model updating frequency

Based on detection date,we made frequency statistics on the number of reports.Because the number of reports before 2020 is too small and incomplete, we selected the data for 12 months of 2020 and made a statistical analysis of its frequency based on **Assumption 1** above.Here is the figure we analyzed.

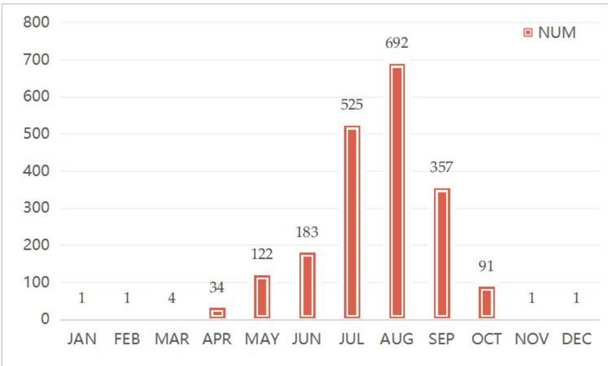


Figure 4-12:frequency statistics on the number of reports in 2020

We found that the number of reports was consistent with the colony activity cycle.So we take the monthly average of the total number of reports for each of the three stages as the update threshold.

Table4-9:number of reports per month for each 3 stages

	F_1	F_2	F_3
Num of reports per month	1.75	113	416.5

We set up a record database for each stage to store reports,which facilitates separate analysis of the reports collected by each stage to further explore the activity characteristics of each stage's pests population.We check the number of reports in the database once a month,When the number of new reports in the database reaches the threshold, a model update is performed.If a report still exists in the database when the stage ends, it is added to the next model update, which prevents the stage's report from being shelved until the next year.

It is worth noting that for the update of the update threshold, we take the monthly average of the total number of reports in the same stage from 2020 as the new update threshold.

4.4.3 Our Strategy

As shown in the **figure4-11**, based on the number of reports growing over time,we divide the year into three stages,and examine reports for each stage monthly.When the number of new reports reaches the corresponding update threshold, the model is updated, all the picture data obtained is input into the training model in **Stacked-BLS**, and the accuracy between blocks is compared to determine whether updates are needed,.Finally we get models that don't need to be updated or that have already been updated.

4.5 Evidence of Judging Pest Eradication

4.5.1 Analysis of the Basic Situation

Asian giant hornets are social hornets, with an annual colony cycle.Because of their subterranean nesting habit, locating the nests of Asian giant hornets can be very difficult.Due to the limited public resources of the government, it is impossible to conduct a comprehensive and detailed inspection of all parts of Washington state.From ***New Pest Response Guidelines***,we know that there are some approaches for eradication,such as targeting queens or targeting nests.We will combine data analysis, forecasting and site visits manner, in as much as possible to reduce government operating costs in this respect the conditions determined in accordance with the Asian giant hornet is eradicated.

4.5.2 The judgment of eradicating pests

Based on the pest’s pattern,we divide the obtained report in 2020 into three parts,which is corresponding to the over-wintering stage,queens' stage outside of the nest,and workers' stage outside of the nest.In over-wintering stage, it is hard to find nests or queens.Therefore,we mainly analyze data collected in queens' stage outside of the nest,and workers' stage outside of the nest.

In addition, we aim to target queens in queens' stage outside of the nest.And we target nests in workers' stage outside of the nest.

For the data that may be obtained in the future,We consider the following two situations under the limit resources of government:

- Report image data obtained by the laboratory are already marked with labels:

We count the collected reports by each stage.If there is positive in these labels, it means that pests in this area have not been eradicated.We will investigate the areas labeled positive, and use appropriate measures to hunt and kill the queens or destroy the nests.conversely,the pests are eradicated.

- Report image data obtained by the laboratory are not marked with labels:

We plan to establish separate databases for the data of these two stages and input them as training data sets into the classification model based on the Broad learning system for training, so as to obtain two different pest classification and recognition models with period characteristics.Firstly, the newly acquired image data is input into the trained pest classification and recognition model corresponding to the period, and the expected output is set as the tag value, so as to obtain the prediction tag of the image.Because the framework shows good performance and classification accuracy in the previous training tests, the prediction tag we obtained has high reliability.Therefore, based on this prediction label, we will concentrate resources on investigating the areas whose prediction label is positive.If there is positive in these labels, it means that pests in this area have not been eradicated.We will investigate the areas labeled positive, and use appropriate measures to hunt and kill the queens or destroy the nests.conversely,the pests are eradicated.

On the situation in various regions of the state of Washington a year to analyze statistics, if all regions have confirmed the pests has been eradicated, to some extent, be considered a pest in Washington has been eradicated.

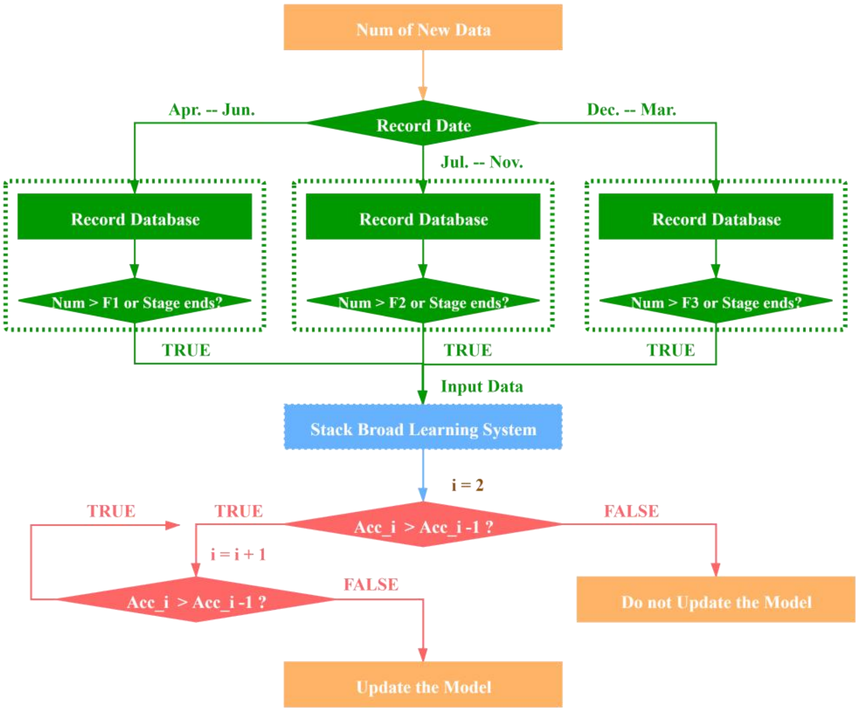


Figure4-13:Model strategy flow chart based on S-BLS and pest seasonal activity

5 Sensitivity analysis

For our ARIMA model, we changed the parameters: autoregressive process of order **p**,difference order **d** and sliding average process of order **q** [p,d,q]reset to[3,0,0] and [0,0,2].This makes the fitting effect of our model to the data worse than the original parameters, and the predicted value does not match the expected trend.We think this phenomenon may be that the amount of postive data we obtained is too small, and the data is not very relevant. Therefore, the emergence of the Asian giant hornet in this area cannot be well represented over time.

For the width learning system model of pest classification, we modified the number of enhancement nodes. We found that as the number of nodes increases, the overall accuracy of the model has a peak.The highest accuracy is maintained at about 95.784%, but the time cost will increase accordingly.This shows that for the current data set, it is not that the more nodes are built, the better the effect. In order to maximize the training effect of the model,We have to choose a more balanced interval between accuracy and calculation performance according to the actual situation.

These trends indicate that the model we use is sensitive to the balance of the data set and the amount of data. If the number of categories in the data obtained is sufficient and the proportions are appropriate, our model will show better performance.

6 Evaluation of the model

6.1 Strengths

- Our pest classification model is flexible, fast and efficient because of the special structure of the Broad Learning System.Based on the Broad Learning System, we established pest classification model.Compared with the traditional deep network model, it has no gradient descent process, requires fewer calculation parameters, faster speed,and less prone to over-fitting.Additionally,it can flexibly adjust the number of nodes to obtain the best model effect.It also has lower hardware requirements, which helps the government maintain databases and update maintenance models at a lower cost.
- The priority scoring system we have established takes full account of factors such as time, space, and the prediction results of our model.It comprehensively considers historical data, spatial geographic location, and data predicted values to obtain a more reliable priority survey standard.This can provide an effective reference for the government and help focus limited public resources on solving problems in key areas.
- The update strategy we provide takes into account both data and reality features.It not only chooses an appropriate S-BLS model for the analysis of actual data, effectively provides the basis for judging whether and how to update the model, but also determines the update frequency according to the actual activity of pests, which provides an effective reference for model updating.

6.2 Weaknesses

- For the pest transmission model, because the amount of positive report data is too small, a model with strong generalization ability cannot be obtained.It will lead to a decline in the accuracy of the prediction results.
- Our model depends on the balance of data and the amount of data.When the data is seriously unbalanced or the amount of data is too small, the effect of our classification model will be poor, so we use the method of data enhancement to randomly generate a small number of data, which effectively alleviates the problems caused by the data imbalance.
- Because of the less number of the data,we only considered the report status within one year and did not carry out cross year comparative analysis.

7 Memo

From:Team 2113014,MCM 2021
To:the Washington State Department of Agriculture
Date:February 8, 2021
Subject:Analysis and Suggestions of Asian Giant Hornet

In order to determine whether the Asian giant hornet has expanded its existence over time in Washington State and nearby areas, we find the current report that confirmed the occurrence of the Asian giant hornet.we conduct time series analysis on current data and forecast future conditions.Finally, we conclude a reliable pest transmission situation.

At the same time, our team adopt a classification model BLS for the image data submitted by the public that the government has obtained and its corresponding status tag. By training and testing it with the processed data set, we obtain a model that can effectively classify Asian giant hornet and other colonies.We use a stacked BLS model to update the model, and determine the update frequency according to the seasonality of pests activity, and get an update strategy of pest classification model, which makes our model more reliable.

In order to use resources more rationally and effectively, we establish a scoring system, which considering reports of a certain area category prediction results, geographic space and historical data of the area.The scoring system is used to evaluate the probability of positive reports in this area, and based on this, it can be used as the priority reference for the government to investigate the areas where these reports are located.

Finally, we use the results of the above work, combine with the actual available methods of location and elimination of pests, and put forward the basis for us to determine that the Asian giant hornet was eliminated in the Washington State area.This basis will be established according to different situations and using different methods, and considering the rational use of the government's limited resources. It also has certain validity and accuracy.

Here are our results:

- The Asian giant hornet is less aggressive in regions, and there is no obvious trend of large-scale spread over time, and it is only confined to a small area along the coast of British Columbia. Its population size is also relatively small, and it don’t have a significant impact on the interior of Washington State.
- Based on the category prediction results of the report, the distance from the Asian giant hornet and the number of historical negative reports within a certain range,we established a survey priority evaluation model.The results show that the number of historical negative reports within a certain range should be the first consideration index of priority investigation. If the number of negative reports received from a small area is more, the possibility of pests is less.
- According to the seasonal activity characteristics of the pests,they are most active during the period from July to November, and the number of sighting reports received at this time will increase accordingly. Due to the increase in the amount of data, during this period of time, we need more complex models and the more frequent model updates.
- We determine the basis for judging whether pests are eradicated.For a certain area, we only conduct statistical analysis on the sighting reports of the two stages where the pest

are more active. When the reports of various areas in Washington State within one year have been identified or predicted as well as targeted on-site inspections, no pests have been found, we can declare that the pests have been eradicated in Washington State.

The proposal and strategies are as follows:

- In view of the fact that the Vespa mandarinia has not been found to have the characteristics of large-scale spread and harm the local ecology. The government can carry out relevant publicity to the public, display and explain the analysis and prediction results of relevant data, and soothe public sentiment.
- According to the time characteristics of the activity of the swarm, the government can divide the year into three phases. Different phases set different thresholds for the database storing public reports, and regularly input the accumulated data into the system for updating, which better reflects the development status of the bee colony in the region. At the same time, it can also reduce the government's operating costs.
- Under the condition of limited public resources, the focus of the elimination of Vespa mandarinia should be focused on finding nests of them and eliminating queen bees. In the active period of the queen bee, you can directly hunt the queen bee or locate and destroy the nest not far away by identifying the results of the image data; while in the active period of the worker bee, you can set traps, Locate the nest and destroy it together with the reproductive queen bee in the nest.

8 Reference

[1] Time Series;BLS;Stacked-BLS; ARIMA
[2] C. L. P. Chen and Z. Liu, "Broad Learning System: An Effective and Efficient Incremental Learning System Without the Need for Deep Architecture," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 1, pp. 10-24, Jan. 2018,
[3] United States Department of Agriculture,New Pest Response Guidelines.
[4] Fangqiong Luo,Chunmei Wu.A Summary of the Theory and Application of Time Series Analysis[J].Journal of Liuzhou Teachers College,2009,24(3):113-117
[5] Z. Liu, C. L. P. Chen, S. Feng, Q. Feng and T. Zhang, "Stacked Broad Learning System: From Incremental Flatted Structure to Deep Model," in IEEE Transactions on Systems, Man, and Cybernetics: ystems,vol.51,no.1,pp.209-222,Jan.2021,
[6] XU Qi-li Design and Implementation of Comprehensive Evaluation System Based on Entropy Weight Method[J],2020,41(01):79-84
[7] Francois Chollet,Deep Lreaning of Python,2018
[8] https://www.nationalgeographic.com/animals/2020/05/asian-giant-hornets-arrive-united-states/

9 Appendix

Appendix A Model-3 Priority evaluation score sheet for untreated/unclassified areas

No.	*la*	*lo*	*values*	*d*	*num*	Total score
1	46.7144	-122.9834	0.266	253.8515	93	2.2236
2	47.804	-121.549	0.2687	156.5180	217	2.0887
3	48.7238	-122.3544	0.3123	39.2490	229	2.3683
4	46.9006	-122.2743	0.254	234.4335	271	1.8378
5	48.9392	-122.6613	0.2376	6.7634	182	2.1888
6	47.5861	-122.0073	0.2552	164.2604	823	0.8103
7	45.7359	-122.6456	0.3603	361.6636	66	2.5527
8	48.9102	-122.3571	0.2795	26.7963	183	2.3384
9	48.0542	-122.1566	0.2006	111.6216	556	1.1762
10	47.829	-122.0981	0.1919	136.5805	745	0.7349
11	47.0561	-122.7623	0.5923	215.1419	339	3.1163
12	47.5015	-122.1875	0.178	169.8484	882	0.3680
13	48.9936	-122.4017	0.2264	21.8867	168	2.1548
14	48.7126	-122.4974	0.2832	34.6031	236	2.2391
15	47.3136	-122.4296	0.2539	187.5702	767	0.8926
16	47.2551	-122.4416	0.2467	193.9334	697	0.9964
17	47.3886	-122.1327	0.237	182.9469	770	0.8219
18	47.6578	-122.3258	0.2534	150.8138	878	0.7070

19	48.4462	-122.3096	0.1621	67.1807	252	1.6736
20	47.637	-117.231	0.2325	425.9435	70	1.9502
21	47.7391	-121.9911	0.2331	148.5931	732	0.9183
22	46.5805	-118.2146	0.2505	422.5641	10	2.1482
23	47.9633	-122.0014	0.2514	125.2641	596	1.2908
24	47.448	-121.7701	0.2464	184.5273	515	1.3700
25	48.4887	-122.4819	0.2851	58.3295	257	2.1801
26	47.3807	-122.0214	0.2923	185.8004	720	1.1470
27	47.7677	-122.2633	0.2587	139.8110	796	0.9044
28	46.706	-120.481	0.2861	301.0914	44	2.3550
29	48.9934	-122.7527	0.2412	3.6723	170	2.2312
30	48.5013	-122.2346	0.2432	64.4208	244	2.0270
31	48.3487	-122.4002	0.2518	74.9177	297	1.9448
32	47.656	-122.4058	0.2417	150.0658	875	0.6654
33	47.7195	-122.337	0.2498	143.9343	837	0.7814
34	47.5264	-122.9836	0.2434	164.1767	330	1.7502
35	47.2449	-122.3127	0.2532	196.2043	700	1.0150
36	47.0456	-122.8306	0.2262	216.4631	298	1.6881
37	48.0909	-122.1712	0.3029	107.4366	520	1.6741
38	47.7184	-122.201	0.3119	146.2132	821	1.0671
39	47.8223	-122.3615	0.2748	132.3919	788	0.9949
40	48.9979	-122.7299	0.2976	2.0629	169	2.4670
41	47.7546	-122.278	0.3033	140.9855	805	1.0691
42	47.4154	-122.3161	0.2954	177.4523	830	0.9480
43	48.7523	-122.4837	0.2507	31.1794	230	2.1207
44	48.8186	-122.4537	0.2335	26.5713	219	2.0768
45	47.3534	-122.0566	0.2494	188.0678	718	0.9716
46	48.5122	-122.6327	0.2961	53.7114	265	2.2140
47	47.3979	-122.6243	0.2341	177.2473	780	0.7958
48	47.3922	-121.9765	0.2882	185.4800	702	1.1663
49	47.4912	-121.7166	0.2728	181.5923	464	1.5842
50	47.5022	-122.164	0.264	170.1557	882	0.7219
51	48.9942	-122.7535	0.2861	3.7328	170	2.4159
52	48.055	-122.1654	0.2042	111.3122	560	1.1833
53	45.5826	-122.3628	0.296	379.4613	67	2.2667
54	48.9803	-122.5971	0.2438	7.8035	171	2.2352
55	47.5793	-122.3879	0.2217	158.6759	894	0.5356
56	47.2578	-122.4589	0.3165	193.5226	697	1.2846
57	48.8312	-122.5515	0.2143	21.1297	220	2.0016
58	45.6515	-122.5147	0.3915	371.2605	67	2.6691
59	48.7871	-122.4884	0.2108	27.7369	223	1.9742
60	47.6245	-120.6787	0.2785	211.7119	81	2.3436

61	47.6517	-122.6928	0.2472	148.9838	731	0.9778
62	47.6298	-120.6895	0.3371	210.7359	82	2.5842