Open Access Library Journal
2022, Volume 9, e9474
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Distribution Estimation of Invasive Species Based on Crowdsourcing Reports

Yuxin Shi[1], Siyuan Liu[2], Tingzhen Liu[1]

[1]Information Science and Technology College, Dalian Maritime University, Dalian, China
[2]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
Email: 863480069@qq.com

How to cite this paper: Shi, Y.X., Liu, S.Y. and Liu, T.Z. (2022) Distribution Estimation of Invasive Species Based on Crowdsourcing Reports. *Open Access Library Journal*, **9**: e9474.
https://doi.org/10.4236/oalib.1109474

Received: October 20, 2022
Accepted: November 27, 2022
Published: November 30, 2022

Abstract

Species invasion will cause certain harm to the local ecosystem. *Vespa mandarinia*, discovered on Vancouver Island, is harmful to agriculture and predators of European honeybees. The government tried to use a crowdsourcing system to collect information and formulate policies to eliminate *Vespa mandarinia*. However, the information provided by the local population about *Vespa mandarinia* is not entirely accurate. For this problem, we build a method to mine trusted information in massive crowdsourcing *Vespa mandarinia* reports. We consider providing the date and location of the report, and establishing a credibility calculation model for further analysis. For the report date, we calculate the normal distribution parameters based on the frequency of the report in each season to measure the reliability of a single report. For report location, we use K-means cluster analysis to find the location of the center point, which is regarded as a hive, count the report points in each hive radiation range, and use these points to generate two-dimensional normal distribution parameters to normalize the data and eliminate statistical errors. We take the probability density of the report at its location as the reliability of the reports. Through credibility, we can screen out reports that are more likely to be positive for prioritizing investigation. In order to better analyze the newly discovered reports in the future and ensure the timeliness of the model, we set up distributed incremental adjustment model to modify normal distribution parameters, and update the existing model.

Subject Areas

Biophysics, Biotechnology

Keywords

Data Mining, Public Health, Biotechnology, Cluster Analysis, Crowdsourcing Data

1. Introduction

Vespa mandarinia, discovered on Vancouver Island in the fall of 2019, is harmful to agriculture and predators of European honeybees. Therefore, it is necessary to study this hornet and the spread of them over time, officially developed a crowdsourcing system [1] for *Vespa mandarinia*. But citizens did not know *Vespa mandarinia* very well, many witnesses provided wrong information. Part of the information has been tested by the laboratory and verified. However, there are still many reports that cannot be verified by the laboratory due to a lack of information, and some reports have not yet been processed.

How to eliminate errors in crowdsourcing data to extract trusted information is a topic of social data analysis research [2]. Willett *et al.* [3] used clustering analysis and user interface optimization to improve the yield of crowdsourcing data. Koswatte *et al.* [4] used the naive Bayesian network to assess the credibility of crowdsourcing rescue information in the 2011 Australian flood event. Loganathan *et al.* [5] used logical regression to classify whether crowdsourcing data is reliable. Shamir *et al.* [6] used the performance of the supervised learning model to judge the noise level in the crowdsourcing data. Silverman *et al.* [7] evaluated the conditions under which the sample mean of crowdsourcing data can measure data reliability based on the maximum entropy principle.

In this paper, we will build a model to mine the information in crowdsourcing *Vespa mandarinia* reports [8]. The information available from the report includes their submission date, longitude and latitude. We build two models to analyze their submission date and longitude and latitude. For the submission date, since the life habits of *Vespa mandarinia* are affected by the seasons, we build a model to analyze the probability density of the number of reports in each month as the season's credibility.

For longitude and latitude coordinates, we use K-means [9] to determine the cluster center, then, analyze to determine the distribution of hives. Since the radiation range of a hive is 30 km, we first need to find the reports within 30 km of each hive, then, use these reports to calculate the two-dimensional normal distribution parameters [10] radiated by each hive. In this way, for each unverified and unprocessed report, the nearest hive can be found, and the probability density of its corresponding position in the two-dimensional normal distribution radiated by this hive can be calculated. This probability density value is regarded as the location's credibility. The season's credibility and location's credibility are combined to calculate the final credibility, so that public health agencies can take can first investigate reports with high credibility. When a new report is received, we quickly update the model by incrementally estimating the parameters of the normal distribution.

2. Data Pre-Processing

Since we are handing a problem with big data, there is a diversity of data with different types. Besides, the data interact with each other to some degree. We must deeply analyze the data to dig out the meaning of each column and the va-

lidity of each data set.

In order to analyze the distribution of hornets over time, we first simply process the data, and select data from the past two years and exclude negative reports. We use Python to draw the distribution map of hornets over time (including positive reports, unverified reports and unprocessed reports), as shown in **Figure 1** and **Figure 2**.

We found that the number of positive reports is very small and most of them are concentrated in one area. Therefore, the information that positive reports can provide is very limited. It is necessary to find a way to mine information from unified and unprocessed reports. In addition, no obvious trends can be seen from the year. We will build a model to find out the trends of its seasons and geographic locations.

3. Credibility Calculation Model

In order to make better use of the information provided in the reports, more accurately judge the correctness of each report, We calculate their credibility based on the reported Detection Date (season) and the reported longitude and latitude (location).

3.1. Calculate Reliability Based on Season

As climatic conditions have a great influence on the survival of *Vespa mandarinia*, the detection date reported is an important factor for judging an unverified or unprocessed report. Taking 2020 as an example, we have compiled the number of reports for each month, as shown in **Figure 3**.

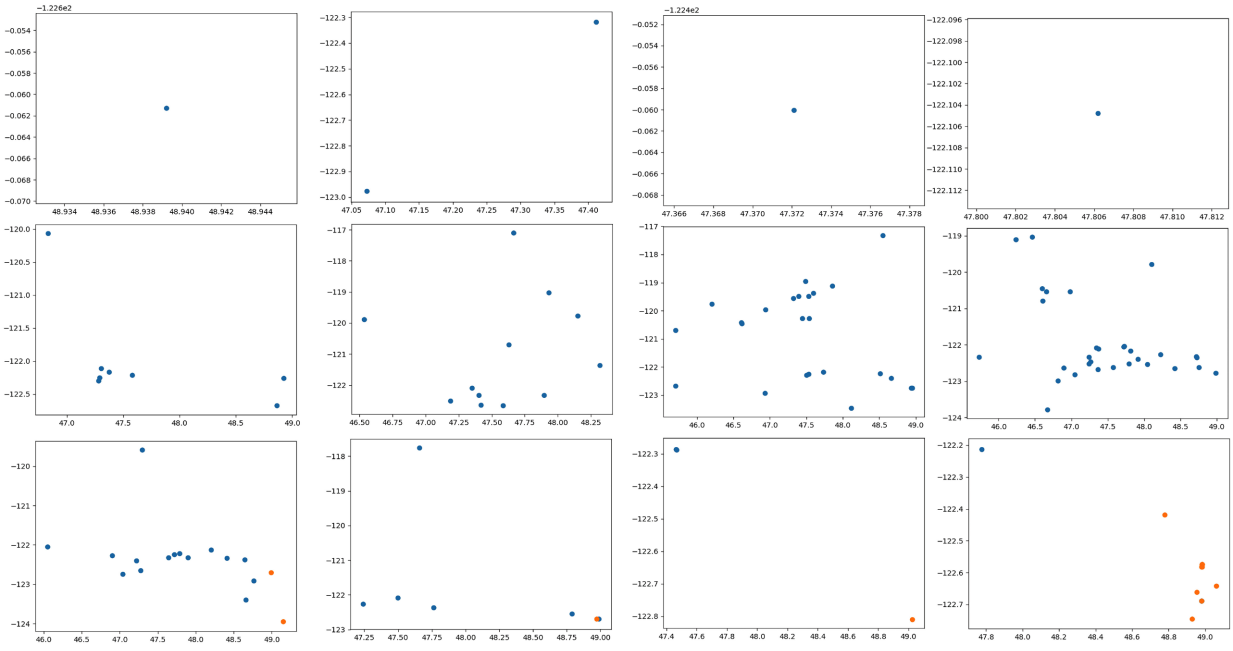

Figure 1. The distribution of the hornets over time in 2019. **Note:** The orange dots represent positive reports, and the blue dots represent unverified and unprocessed report. The abscissa is latitude, and the ordinate is longitude. In each picture, from left to right and from top to bottom is January to December.

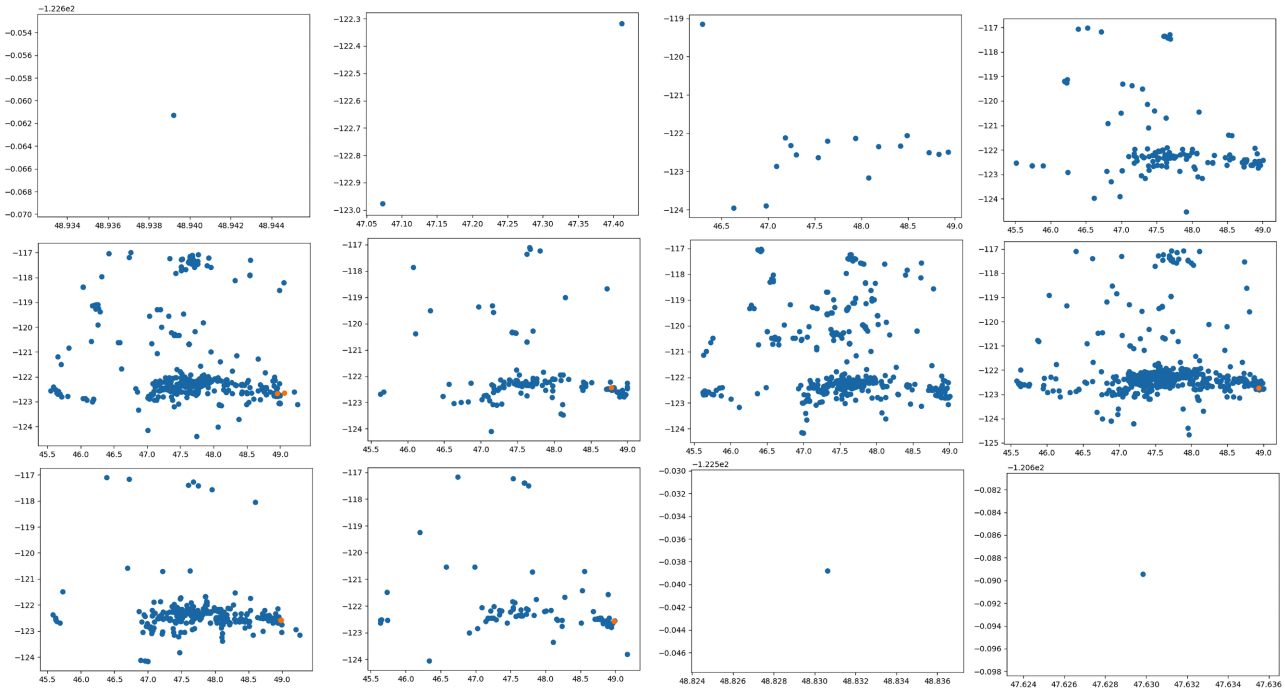

Figure 2. The distribution of the hornets over time in 2020. **Note:** The orange dots represent positive reports, and the blue dots represent unverified and unprocessed report. The abscissa is latitude, and the ordinate is longitude. In each picture, from left to right and from top to bottom is January to December.

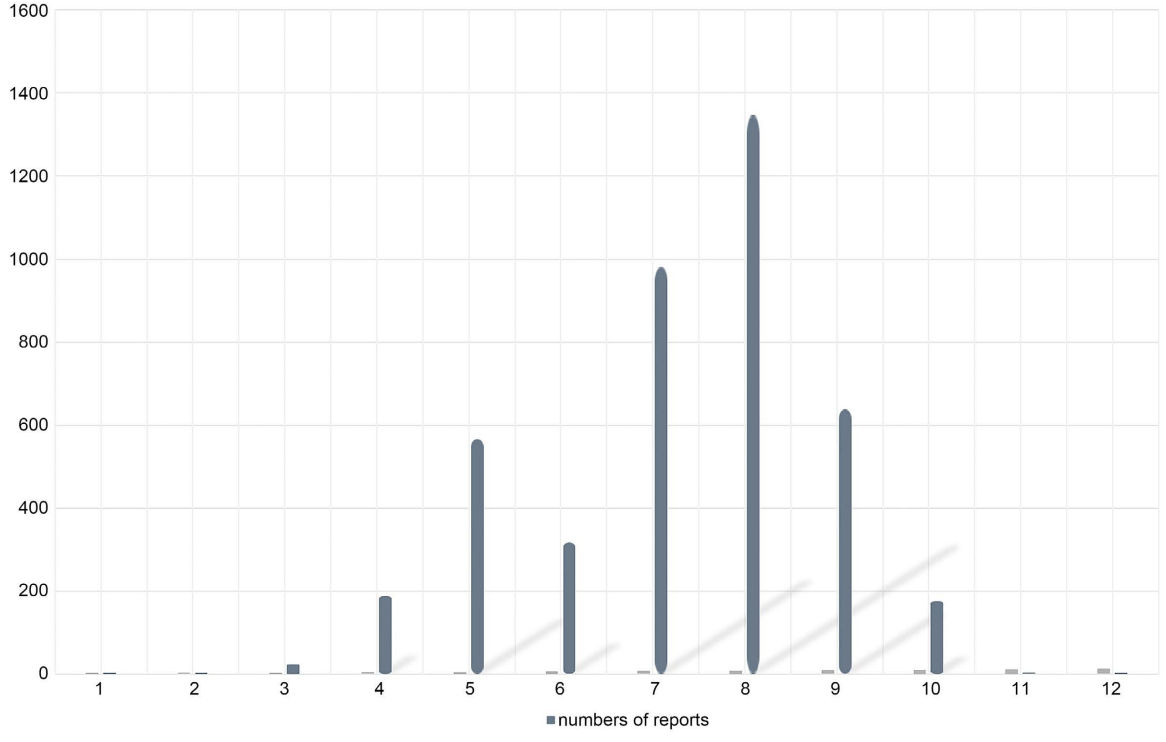

Figure 3. Number of reports per month in 2020.

It can be observed that August has the largest number of reports. Since the number of *Vespa mandarinia* is affected by many factors. The amount of data in

practice is large and independent of each other. Therefore, theoretically, the number of reports per month should follow a normal distribution. However, the number of reports in June in the picture is less than in the two adjacent months. This may be caused by errors in data collection. In order to solve this error, we re-normalize the data based on the statistics of the normal distribution.

The mean value of month $\mu = 8$, its variance $\sigma^2 = \sum_{k=1}^{12} (x_k - 8)^2 p_k$, p_k is the normalized result of the number of reports. After calculation, we get Table 1.

After calc After calculation we get $\sigma^2 = 3.034234234$. We get a normal distribution with $\mu = 8$, $\sigma^2 = 3.034234234$.

The month M is a random variable, and the reliability of each month's reports obeys this distribution, written as $M \sim N(8, 3.03)$.

Substituting the months from January to August into the normal distribution, we get the credibility of the probability density function for these months. Because in reality, the number of *Vespa mandarinia* in winter decreases faster than it grows in spring. The purpose of using the normal distribution is to make the data on the left and right sides of the mean of the month have their own monotonicity, but at the same time make the data on both sides have symmetry. So we have to deal with it to make it asymmetrical. The probability density from September to December is replaced by the probability density from April to January.

3.2. Calculate Reliability Based on Location

Since a new queen usually has a range estimated at 30 km for establishing her hive [11], the reported longitude and latitude are also important factors for judging credibility. Next, we build a model based on the reported location.

Table 1. The normalized result of the number of reports.

Month	Number of reports	Number of normalized reports
1	1	0.000675676
2	3	0.000900901
3	23	0.002702703
4	188	0.033108108
5	566	0.168243243
6	316	0.074774775
7	980	0.200675676
8	1346	0.324324324
9	637	0.140315315
10	176	0.054279279
11	2	0
12	1	0

We divide hives into two types, one is the location that has been identified as a positive report, namely hives. The other is the gathering point of the report, we call them uncertain hives. The calculation steps are as follows.

3.2.1. Use K-Means Cluster Analysis to Find Uncertain Hives

Step 1. Analyze the data with Elbow Method [12] to determine the cluster center *k*, that is, the number of uncertain hives (**Figure 4**).

The point of maximum slope change rate is 3.75. So there are 4 cluster centers, that is, the number of uncertain hives is 4.

Step 2. Perform K-means clustering and divide it into *k* categories to get **Figure 5**.

The purple part in the lower right corner of the picture is the actual area with positive reports. And K-means clustering results show that the purple part is the smallest category. It shows that this area is the densest, which is consistent with the existing positive reports. This situation can show that the model is reasonable.

Step 3. Find the positive point and all points within 30 kilometers from each cluster center point. This is the maximum radiation range of this hive or uncertain hive. We get Schematic diagram of reports under each hive radiation (**Figure 6**).

Step 4. Since the random variables are longitude and latitude, their correlation coefficient is 0. Calculate the mean variance of these points, and each hive or uncertain hive can get a two-dimensional normal distribution. The credibility of each point can be the probability density of the point on the distribution.

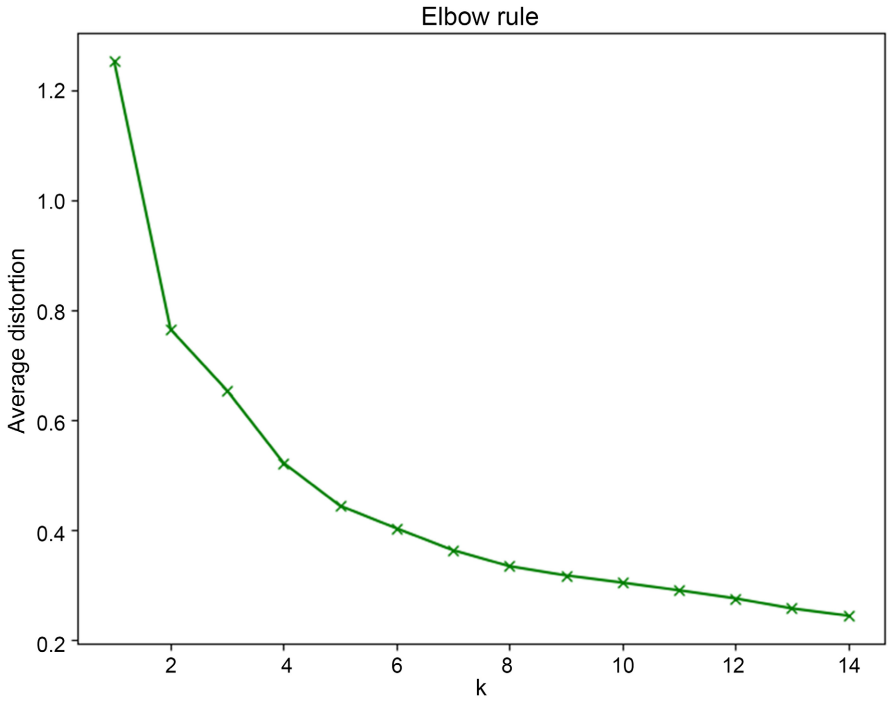

Figure 4. The relationship between *k* and cost function.

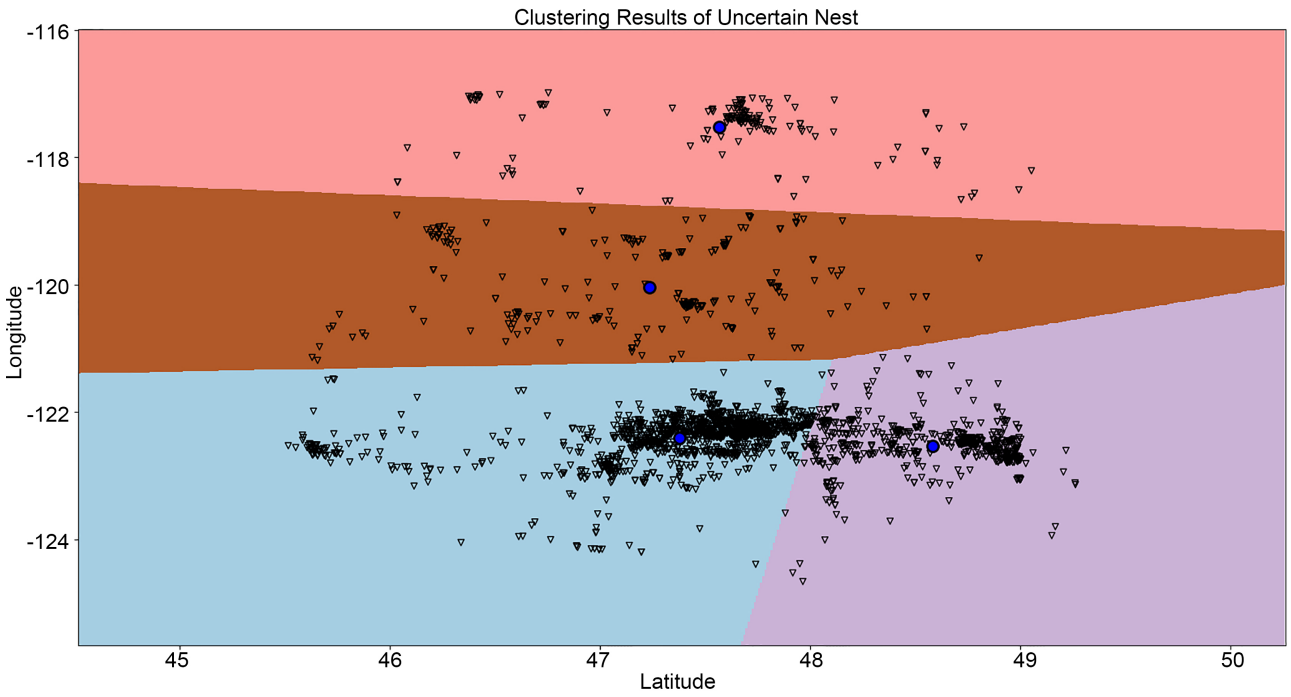

Figure 5. The result graph of performing K-means clustering. **Note:** The pink, orange, blue and purple areas represent different classes. The blue dot represents uncertain hive. The black triangle represents unverified or unprocessed report.

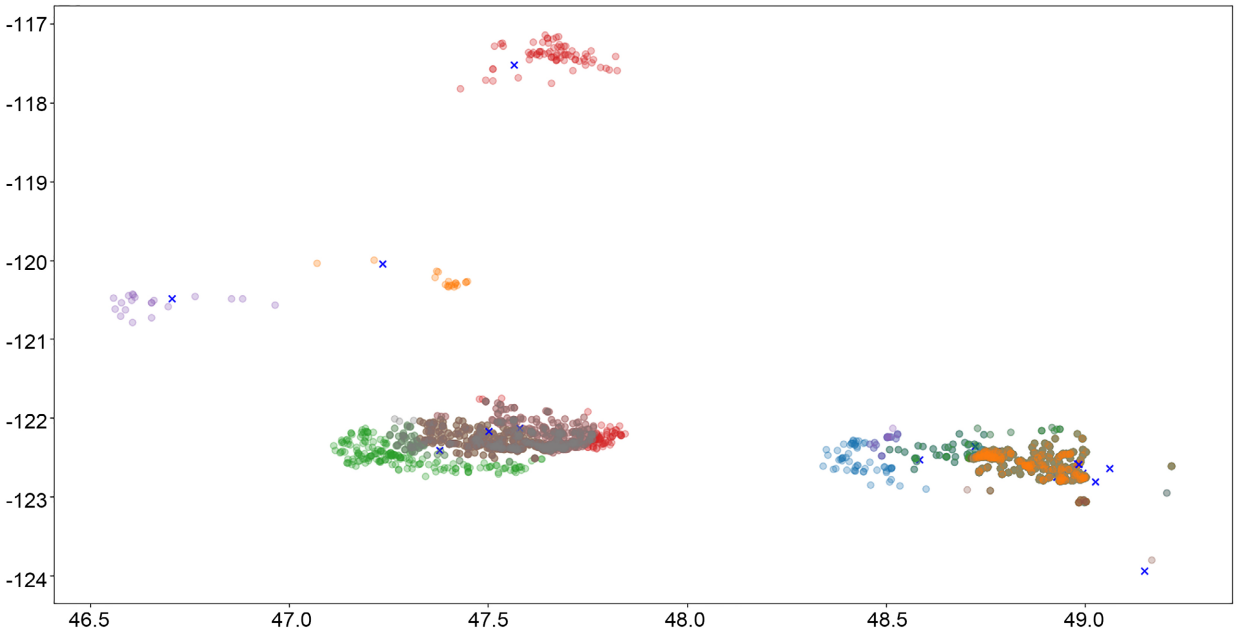

Figure 6. Schematic diagram of reports under each hive radiation. **Note:** Blue cross means positive report or uncertain hive. Different colored dots represent reports radiated by different hives. The abscissa is latitude and the ordinate is longitude.

After calculation, we get the two-dimensional normal distribution parameters of radiation (Table 2).

3.2.2. Calculate Credibility of Each Unverified Report

Step 1. Find the nearest hive or uncertain hive to this point.

Table 2. The two-dimensional normal distribution parameters of radiation.

s_1^2	s_2^2	u_1	u_2
0.021559586	0.02275687	47.37780629	−122.4103369
0.006087037	0.01913566	47.56529326	−117.5211199
0.023101033	0.015975192	48.58362288	−122.5277207
0.008281684	0.01116439	47.23487839	−120.042999
0.014257235	0.015585253	48.723779	−122.354431
7.17E−05	0.004930427	49.149394	−123.943134
0.006557754	0.020917926	48.993892	−122.702242
0.007117487	0.021407771	48.971949	−122.700941
0.004096438	0.019428031	49.025831	−122.810653
0.006893771	0.021722665	48.980994	−122.688503
0.003876809	0.017631808	49.060215	−122.641648
0.008632947	0.020996337	48.955587	−122.661037
0.011103904	0.017261207	48.777534	−122.418612
0.015007777	0.01841511	47.579579	−122.124218
0.011930149	0.009424903	46.706048	−120.481003
0.007938466	0.020856617	48.927519	−122.745016
0.008228704	0.016522836	48.984269	−122.574809
0.017076043	0.018369782	47.50218	−122.16402
0.008228655	0.01652283	48.98422	−122.574726
0.008228608	0.01652283	48.984172	−122.57472
0.008273893	0.016495515	48.979497	−122.581335
0.008141993	0.016516265	48.983375	−122.582465

Step 2. Calculate the probability density of this point on the hive corresponding distribution as the reliability. If no hive is found within 30 kilometers of this point, the reliability of this point is 0.

3.3. Final Credibility

Further, considering that seasonal factors are relatively fixed and location factors are more differentiated, we will combine the season credibility and the location credibility in 4:6 ratio to get the final credibility. Among them, if the distance credibility is 0, the final credibility is directly 0.

After calculation, we get the credibility of all reports. Due to the huge amount of data, we show some data in Table 3.

3.4. Distributed Incremental Adjustment Model

In the future, people may continue to discover *Vespa mandarinia* and provide new reports. For new reports that people discover later, we use the following algorithm to online update our model.

Table 3. Final reliability of partial reports.

ID	Final credibility
{17EBEDE2-7DFF-4342-A66E-376185CD95DE}	0
{204DA998-2F64-40CD-9A52-30D25D272CF5}	0.207743828
{47BBB2BB-2996-4BF5-8329-73F28A7E5778}	0.065569054
{DB01EEBC-66F4-4012-A16B-DDE2365FAE24}	0.916002762
{589E0AD2-4EEF-4588-8969-158B084727AE}	0.207819937
{A0D45071-DFD3-4F0E-9851-FFBD637FBF58}	0
{EB6CACF8-D71E-4C68-AE03-D2AEF2E42421}	0.207766593
{7D0FB65F-A3EA-4436-AA6D-B379565CFE87}	0.776898177
{7E522AE9-2155-4ECE-A3CD-34C4CB91650C}	0.207851304
{CD18D749-4333-4A1A-96FC-1287A8F66B32}	0.207789183

μ_1 is the expectation of the model composed of n data, s_1^2 is the variance of the model composed of n data. μ is the expectation of the updated model composed of $n+1$ data, s^2 is the variance of the model composed of $n+1$ data after the update. We get the following equation:

$$\mu = \frac{1}{n}\left(a_1 + a_2 + \cdots + a_n\right)$$

$$s^2 = \frac{1}{n}\left[\left(a_1 - \mu\right)^2 + \left(a_2 - \mu\right)^2 + \cdots + \left(a_n - \mu\right)^2\right]$$

$$\mu_1 = \frac{1}{n+1}\left(a_1 + a_2 + \cdots + a_n + a_{n+1}\right)$$

$$s_1^2 = \frac{1}{n+1}\left[\left(a_1 - \mu_1\right)^2 + \left(a_2 - \mu_1\right)^2 + \cdots + \left(a_n - \mu_1\right)^2 + \left(a_{n+1} - \mu_1\right)^2\right]$$

Since the value of n is very large, when calculating the variance of the $n+1$th data, it can be regarded as $\mu = \mu_1$.

That is, the variance expression at this time can be written as:

$$s_1^2 = \frac{1}{n+1}\left[\left(a_1 - \mu\right)^2 + \left(a_2 - \mu\right)^2 + \cdots + \left(a_n - \mu\right)^2 + \left(a_{n+1} - \mu\right)^2\right]$$

We can get:

$$\mu_1 = \frac{1}{n+1}\mu + \frac{a_{n+1}}{n+1}$$

$$s_1^2 = \frac{n}{n+1}s^2 + \frac{\left(a_{n+1} - \mu\right)^2}{n+1} = \frac{n}{n+1}s^2 + \frac{\left(a_{n+1} - \mu_1\right)^2}{n+1}$$

4. Conclusions

In this paper, we build three models to analyze the known data and get the final credibility. Using K-means clustering-normal distribution model, and using a normal distribution to repair errors in known data. In theory, the random inde-

pendent time affected by multiple factors is normally distributed. So, we normalize the data to a normal distribution and two-dimensional normal distribution to repair the errors. This can improve the accuracy of the data and the rationality of the results. Using distributed incremental adjustment model modifies the normal distribution parameters, updates the credibility calculation model, and ensures the timeliness of the model. The update frequency is that every new report can be updated. Final credibility is the likehood of correct classification. The calculation formula for the probability of misclassification is $e = 1 - y$.

We sort the reports according to their final credibility. The top-ranked reports are the reports that are investigated first, and they are most likely to be positive sightings.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Keating, M., Rhodes, B. and Richards, A. (2013) Crowdsourcing: A Flexible Method for Innovation, Data Collection, and Analysis in Social Science Research. In: Hill, C.A., Dean, E. and Murphy, J., Eds., *Social Media*, *Sociality*, *and Survey Research*, Wiley, New York, 179-201. https://doi.org/10.1002/9781118751534.ch8

[2] Yuen, M.-C., King, I. and Leung, K.-S. (2011) A Survey of Crowdsourcing Systems. 2011 *IEEE Third International Conference on Privacy*, *Security*, *Risk and Trust and* 2011 *IEEE Third International Conference on Social Computing*, Boston, MA, 9-11 October 2011, 766-773. https://doi.org/10.1109/PASSAT/SocialCom.2011.203

[3] Willett, W., Ginosar, S., Steinitz, A., Hartmann, B. and Agrawala, M. (2013) Identifying Redundancy and Exposing Provenance in Crowdsourced Data Analysis. *IEEE Transactions on Visualization and Computer Graphics*, **19**, 2198-2206. https://doi.org/10.1109/TVCG.2013.164

[4] Koswatte, S., McDougall, K. and Liu, X.Y. (2017) VGI and Crowdsourced Data Credibility Analysis Using Spam Email Detection Techniques. *International Journal of Digital Earth*, **11**, 520-532. https://doi.org/10.1080/17538947.2017.1341558

[5] Loganathan, V., Subramani, G. and Bhaskar, N. (2020) Crowdsourcing Data Analysis for Crowd Systems. In: Ranganathan, G., Chen, J. and Rocha, Á., Eds., I*nventive Communication and Computational Technologies*, Springer, Singapore. https://doi.org/10.1007/978-981-15-0146-3_117

[6] Shamir, L., Diamond, D. and Wallin, J. (2015) Leveraging Pattern Recognition Consistency Estimation for Crowdsourcing Data Analysis. *IEEE Transactions on Human-Machine Systems*, **46**, 474-480. https://doi.org/10.1109/THMS.2015.2463082

[7] Silverman, M.P. (2019) Extraction of Information from Crowdsourcing: Experimental Test Employing Bayesian, Maximum Likelihood, and Maximum Entropy Methods. *Open Journal of Statistics*, **9**, 571-600. https://doi.org/10.4236/ojs.2019.95038

[8] Shi, Y.X. (2022) *Vespa Mandarinia* Crowdsourcing Reports. Figshare. Dataset. https://doi.org/10.6084/m9.figshare.21333966.v1

[9] Wang, Z., Liu, Q. and Chen, E. (2009) A K-Means Algorithm for Optimizing the Initial Center Point. *Pattern Recognition and Artificial Intelligence*, **22**, 299-304.

[10] Xia, X.-F., Liu, X. and Li, X.-M. (2010) User-Item Missing Ratings Complement Based

on Two-Dimensional Normal Distribution. 2010 2*nd International Workshop on Database Technology and Applications*, Wuhan, 27-28 November 2010, 1-6. https://doi.org/10.1109/DBTA.2010.5658988

[11] Huang, S.K. (2001) The Preliminary Report on *Vespa mandarinia* and Other Arthropods in Its Cave. *Journal of Fujian Agricultural University* (*Natural Science*), **30**, 99-102.

[12] Wu, G.J., Zhang, J.L. and Yuan, D. (2019) Automatically Obtaining K Value Based on K-Means Elbow Method. *Computer Engineering & Software*, **40**, 167-170.