文法错误

4.2.1 Detection time

For months, the main consideration is the probability of correct sighting reports for each month.Here, starting from the reverse side, calculate the probability of false sighting reports, P_w, and the probability of being correct, take P_w.

We took error reports for each of the last two years, with respect to the frequency distribution of the months, because there are too few correct samples, if you look at the positive side, the sample error is huge, but too many wrong samples to take advantage of. The frequency is used as an estimate of the total error probability. Notice the data in the PDF attached, points out that population is at its peak in August, about logically, hornet is found that the probability is higher, in the period of 2019-2020, however the last witness report and no 7, 8 month of data, in addition, there is also a case of correct sightings in winter. According to the data, except for the female hornets that survive the winter, the rest are basically gone in the winter.

Considering that there will be 100% false positives in a certain month, it is not accurate to estimate only through the original data. Therefore, data smoothing is needed. In reference to the hornets life cycle, the probability of sightings is lower in winter and 2 and higher in other seasons, so there is the following data smoothing strategy:

Note that the number of wrong samples for the jth month within two years is W_j, the number of correct samples is R_j, the total sample number is T_j, and the probability of reporting an error is P_i. Let

$$max_w = maxW_j$$
$$max_r = maxR_j$$
$$min_w = minW_j | W_j > 0$$
$$min_w = minR_j | R_j > 0$$

$$W_j = \begin{cases} W_j + max_w & \{j \in 1, 2, 12\} \\ W_j + min_w & \{j \in 3, 4, 5, 6, 7, 8, 9, 10, 11\} \end{cases} \tag{3}$$

$$R_j = \begin{cases} R_j + max_r & \{j \in 1, 2, 12\} \\ R_j + min_r & \{j \in 3, 4, 5, 6, 7, 8, 9, 10, 11\} \end{cases} \tag{4}$$

$$T_j = W_j + R_j$$

1 detection time
2 of
3 negative
4 positive
5 take P_w.
6 negative reports
7 语句不通，
8 pointing
9 so
10 语句不通
1 不通顺
2 不通顺
3 不通顺
4 不通顺
5 a negative report
6 这里建议把花括号加上不然很别扭

results are as follows:

According to the information provided, we can know that the probability of prediction error should approximately decrease, and then increase with time, which corresponds to the first reproduction of bumblebee population in its life cycle (As the number of individuals in the population increases, the probability of being found increases), And population extinction (the number of individuals decreases, and the probability of being found decreases). Here, we choose a polynomial function to fit, using the least square method. After testing the 1-5 degree polynomial function, we finally decided to use the cubic function to fit the curve of the change of error probability over time.

$$P_w(m) = 0.0011m^4 - 0.0306m^3 + 0.2793m^2 - 0.9373m + 1.7591 \quad (5)$$

The two minimum points of the curve are March and November, respectively, which correspond to the time when bees are awake and when they are out looking for nests, respectively, and tend to produce the most credible reports.

To better fit the curve, an extra abscissa is set, and the corresponding ordinate is the January value.It represents the estimated value on December 31, which should be close to the estimated value on January 1. In addition, it is observed that there are parts of the curve greater than 1, which are replaced by a probability of 0.99.

4.2.2 Notes

To deal with Note, the first thing is converting each the note sentences to a numerical fixed-length vector. To get these vector, we pay attention to the importance to each words, using TF-IDF (Term Frequency / Inverse Document Frequency) to get calculate each important of words. Finally _

TF represents the frequency of occurrence of word w in sentence D_i. the $count(w)$ representes the the numver of word w in sentence D_i and $|D_i|$ represents the number of words in sentence D_i.

$$TF(w, D_i) = \frac{count(w)}{|D_i|} \tag{6}$$

IDF reflects the prevalence of words. When a word is more common, this word has lower IDF value.

$$IDF(w) = \ln \frac{N}{1 + \sum_{i=1}^{N} I(w, D_i)} \tag{7}$$

Then for any notes, we could calculate a TF-IDF vector to represent this notes. we use this value to measure the possibility in the note part in the following method: find the closest positive in Eular distance d

$$d(x, y) = \tag{8}$$

We use the k-top frequent words to build a dictionary bu the statistics of all notes. After suspending of some meaningless words, we use k as 100 here and part of the dictonary are as below

In all the already curtain note, which already has the positive or negtive lab status, we pick up the cloest positive one and calculate the distance d_{min}, putting this value in the function below:

The above matrix is a positive reciprocal matrix. According to the Perror theorem, it must have a maximum eigenvalue λ_max, and the eigenvector $\vec{X}$ corresponding to λ_max is a positive vector. There is $AX = \lambda_max\vec{X}$, All the components of $\vec{X}$ are unified for normalization, and the vector is obtained as the weight vector $\vec{W}$. After calculation, get $\lambda_max = 4.05111$ and

Then, the [redacted] calculation formula is obtained.

$$P(Positive) = \vec{W} \cdot (RP, MN, NT, PT)^T \tag{10}$$

4.3 the Determination of Weights

Using analytic hierarchy process (AHP) to assign weights, we need to consider the impact of month, feature description, and geographic location of the report. We compare the indicators with each other to build A comparative judgment matrix to represent the relative importance of the indicators, which is represented by A below. After discussion, we think the importance is: photo, report location, month, and feature description.

a

$$A = \begin{bmatrix} 1 & 2 & 4 & \frac{1}{2} \\ \frac{1}{2} & 1 & 3 & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{3} & 1 & \frac{1}{5} \\ 2 & 3 & 5 & 1 \end{bmatrix}$$

The larger the value of elements a_{ij}, the more important it is. a_{ij} value greater than 1 means that the attribute represented by the i-th row is more important than the attribute represented by the j-th column.

The above matrix is a positive reciprocal matrix. According to the Perror theorem, it must have a maximum eigenvalue λ_max, and the eigenvector $\vec{X}$ corresponding to λ_max is a positive vector. There is $AX = \lambda_max\vec{X}$, All the components of $\vec{X}$ are unified for normalization, and the vector is obtained as the weight vector $\vec{W}$. After calculation, get $\lambda_max = 4.05111$ and

According to the Report credibility evaluation system we established, we used our index analysis on the original Report data.It can be seen that most of the positive samples rank in the front after the indicators of each Report are estimated and ranked by our evaluation system.

Reliability

reliability

5.2 Task 2

According to the Report credibility evaluation system we established, we used our index analysis on the original Report data.It can be seen that most of the positive samples rank in the front after the indicators of each Report are estimated and ranked by our evaluation system.

reliabilty

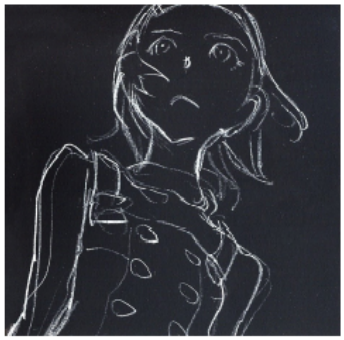

Figure 5: Pending

It can be seen from observation that a small number of positive samples are not in the list because of the absence of photos and other elements, which are within the range of explanation.

For this problem, due to the limited manpower and material resources, it is a rational decision to prioritize the areas with high reliability of reports as the survey target. If there are multiple areas with high reliability, we choose the areas with more intensive distribution of reports as our survey target.All we need to do is find those areas where reliable reports are concentrated and prioritize them.We chose the lower quartile of the indicators of the positive sample as the threshold, and reports above this threshold were considered to have high reliability.For all reports whose reliability indicators are above the threshold regardless of their Status, we make their distribution on the map:

3 can see

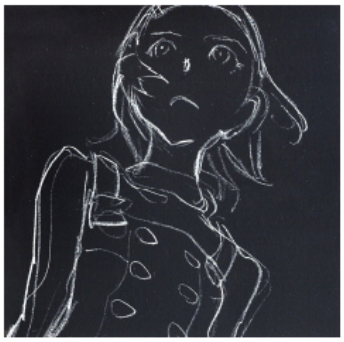

Figure 6: Pending

The specific ranking of regions is as follows: 1 不要了

Therefore, give priority to the investigation of XX area.

2 这里还没回答

We assume that the new report we get has been confirmed by the laboratory whether it is Asian bumblebee or not. If not, we use the current model for preliminary confirmation (if the index value is greater than the positive report, otherwise it is considered negative report).We adopt the following update strategy:

1. Maintain a queue for a new report, which is initially empty.
2. If no new positive reports appear, then we update our model about

1 名字错了

2 reliability

3 是大于正报告的下四分数，不是大于正面报告

once a month.Update model including put the pictures of the new report added to the image recognition of neural network in the training set and the training of neural network further, recount each point to the nearest k report point distance and update the RP value, recount positive report along with the change of in relation to a new curve fitting, and the influence of back calculation note value.

4 不通

3. If new, more positive reports emerge, we immediately update the model.
4. When the update is complete, clear the queue and wait for the next update.

3. If new, more positive reports emerge, we immediately update the model.

1 状语位置?

5.5 Task 5

1 名字错误

Considering that Asian bumblebees have a one-year propagation cycle, we chose one year as the time window for observation.

2 as the length of time window