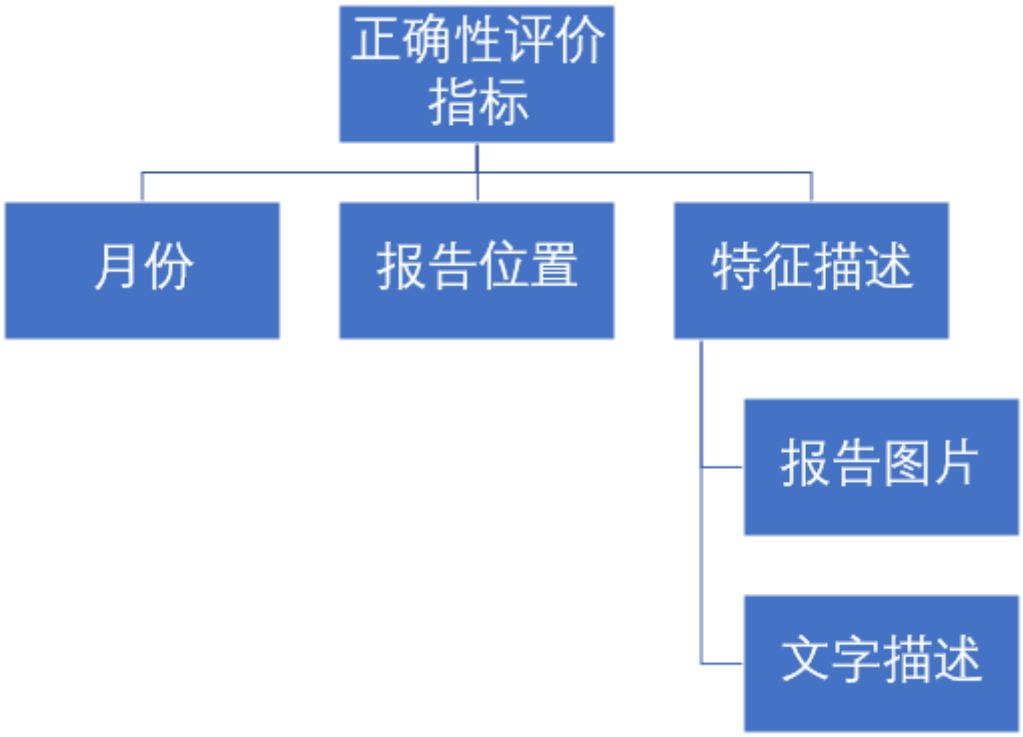
Task B and so on，评价模型的搭建

主要方法

层次分析法。层次结构如下



思路

月份（MN）

对于月份，主要考虑每个月份出现正确目击报告的可能性。这里从反面出发，计算出现错误目击报告的概率P_w，正确的概率为$P_r = 1 - P_w$

我们统计近两年来每一年错误的报告，关于月份的频率分布，（这是由于正确的样本太少，而可以利用的错误的样本却很多，如果研究），并将频率作为错误的概率，作为总的错误概率的估计。但注意到C题中的pdf附件内指出，种群数量大约在八月达到高峰，从逻辑上讲，大黄蜂被发现的概率在该时期内比较高，然而2019-2020年的正确目击报告中并没有7，8月的数据，此外，也有一例于冬季的正确目击报告（Global ID 5AC打头的那个）。仅仅单纯地通过原本的数据进行估计是不准确的，因此，需要进行数据平滑。考虑到大黄蜂的生命周期，在冬季和2人们目击到的概率较小而在其他季节目击到的概率较大，因此有如下的数据平滑策略。

记两年内的第j月的错误样本数为W_j，而正确的样本数为R_j，总样本数为T_j，报告错误的概率为P_j。

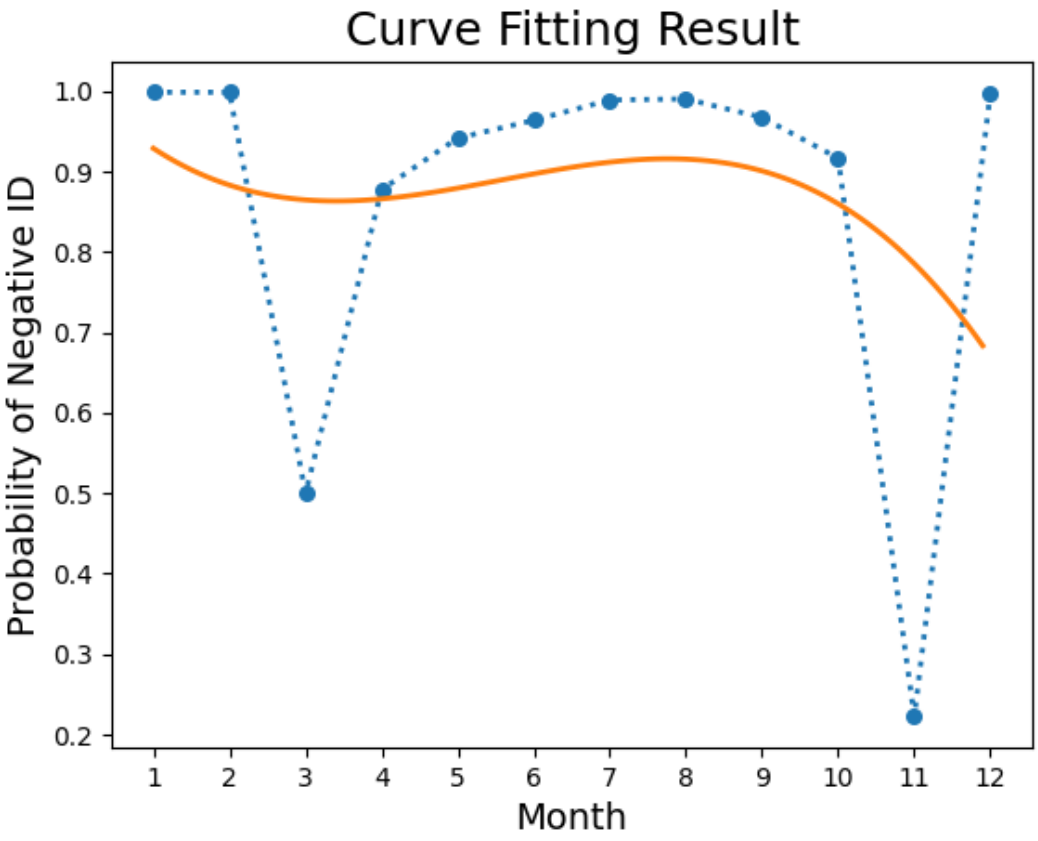
$$
\begin{aligned}
max_w &= \max\{W_j\} \\
max_r &= \max\{R_j\} \\
min_w &= \min\{W_j | W_j > 0\} \\
max_w &= \min\{R_j | R_j > 0\} \\
W_j &= \begin{cases} W_j + max_w & j \in \{12, 1, 2\} \\ W_j + min_w & j \in \{3, 4, 5, 6, 7, 8, 9, 10, 11\} \end{cases} \\
R_j &= \begin{cases} R_j + min_r & j \in \{12, 1, 2\} \\ R_j + max_r & j \in \{3, 4, 5, 6, 7, 8, 9, 10, 11\} \end{cases} \\
T_j &= W_j + R_j
\end{aligned}
$$

j	W_j	R_j	T_j	P_j
1	715	1	716	0.99860
2	715	1	716	0.99860
3	6	6	12	0.50000
4	43	6	49	0.87755
5	129	8	137	0.94161
6	186	7	193	0.96373
7	539	6	545	0.98899
8	715	7	722	0.99030
9	354	12	366	0.96721
10	88	8	96	0.91667
11	2	7	9	0.22222
12	714	2	716	0.99721

依据提供的信息，我们可以知道预测错误的概率应该大致呈现出随时间先减小再增大的趋势，对应大黄蜂种群生命周期的先繁衍（种群中的个体数量增加，导致被发现的概率增大），以及种群消亡（个体数减少，被发现的概率减小），这里我们选择多项式函数来拟合，使用的方法为最小二乘法（ordinary least squares），在测试了1~5次的多项式函数后，最终决定采用三次函数来拟合报告错误概率随时间变化的曲线。

拟合结果如下：

$$P_w(m) = -0.00127m^3 + 0.02127m^2 - 0.10062m + 1.00951$$



预测正确的概率：

$$P_r(m) = 1 - P_w(m) = 0.00127m^3 - 0.02127m^2 + 0.10062m - 0.00951$$

我们以预测正确的概率来衡量报告月份对于评价的影响。

报告位置（RP）

特征描述（FS）

权重分配

使用层次分析法来分配权重，我们需要考虑的指标有月份、特征描述、以及报告的地理位置的影响。我们将指标相互比较，构建一个比较判断矩阵来表示指标之间的相对重要程度，下面用A表示。经过研究，我们认为重要程度依次为：报告位置、月份、特征描述。

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

上述矩阵为一个正互反矩阵，由佩罗（Perror）定理知道，它一定存在一个最大的特征值λ_{max}而且λ_{max}对应的特征向量$\vec{X}$为正向量，有$AX = \lambda_{max}\vec{X}$，将$\vec{X}$的各分量统一起来，进行归一化，得到向量$\vec{W}$作为权值向量。

计算可得，$\lambda_{max} = 3.01829$

权值向量为：

$$\vec{W} = [0.55842 \;\; 0.31962 \;\; 0.12196]$$

所以说预测正确的指标计算公式如下：

$$likehood(Postive) = \vec{W} \cdot (RP, MN, FS)^T$$