# 胡蜂,你在哪？

利用评分系统快速查找并证实胡蜂的存在

（**Confirming the Buzz about Hornets**）

Summary

东方和西方的人们一直都在想象龙。如果有龙生活在我们的世界里，研究它的特征以及与生态系统的相互作用将是非常有意义和有趣的。

首先，研究了龙的基本物理特征。根据生物的生长特性，建立了龙的重量Logistic生长模型。根据这个模型，估计一条成熟的龙的重量在10吨到20吨之间。然后,利用力学知识和类比法,我们龙的类推主干弹性圆柱和估计,成熟的龙的躯干的长度是4米到6米的长度,整个身体成熟的龙是大约20米到35米。其次，我们分析了龙的日常能量摄入和消耗。将龙与其他具有相似特征的生物进行比较，估算出龙的每日能量摄取量和饮食，以及在飞行过程中能量消耗的速率。此外，在给定龙喷火条件下，建立了龙喷火的化学动力学模型，分析了龙喷火的机理和能量消耗。第三，在上述结果的基础上，我们考虑了维持龙的栖息地的最低要求。为了分析龙对生态系统的影响，在Logistic模型的基础上建立了龙-羊-生境模型。更重要的是，比较了干旱地区、暖温带地区和北极地区龙的生存资源的差异，分析了这些地区龙的生存状况。在敏感性分析中，我们着重分析了龙的体重固有增长率、环境因素和环境温度三个因素随其变化对成年龙体重的影响。最后，我们给《冰与火之歌》的作者乔治·r·r·马丁写了一封信，就如何维护故事的现实生态基础提出了一些建议。虽然龙在现实世界中并不存在，但本文所建立的模型也可以用于解决恐龙的生长、外来物种的入侵以及新型火焰喷射器的设计等问题。

the Asian giant hornet 带着它的“侵略计划”，跨过海湾，偷偷地了潜入了华盛顿的地盘。充足的食物，舒服的气温，Vespa mandarinia渴望着将这片土地全部变为自己的“王国”。但是华盛顿政府立刻采取行动号召当地人民进行汇报，但由于资源有限和时间紧迫，需要对最有可能为胡蜂的报告进行调查研究。

为了解决这一问题，我们首先考虑了空间的差异性，研究了害虫栖息地扩张与时间的相关性，在已有的回归模型上，研究了害虫数量与时间的相关性，并做了初步预测

先讨论了害虫在地域上的扩张与时间的关系，其次，在考虑地域差异性的基础上，我们研究了害虫出现事件次数随时间变化的相关性，并进行多项式回归分析，

Content

**1. introduction**

众所周知，外来物种入侵会给当地带来严重的生态环境和经济的破环，2019年9月份，在加拿大不列颠哥伦比亚省的温哥华岛发现了一个维斯帕曼陀林的聚落。巢穴很快被发现并被摧毁，事件的消息传遍了整个地区。之后，在飓风的影响下，在邻近的华盛顿州也确认观察到该害虫，胡蜂的繁殖和对农业的影响引起了当地人民严重的焦虑。华盛顿州为此设立了热线电话和网站，供人们报告目击这些胡蜂的情况。但因为人力和财力资源有限，国家必须决定如何优先安排其有限的资源，以便进行进一步的后续调查。虽然一些报道被确定是胡蜂，但许多其他目击到的昆虫被证明是其他类型的昆虫。

为了提高华盛顿政府的工作效率，我们需要设计一个初步判断信息可靠性的系统，用来初步判断信息可靠性快速查找有效信息，为专业研究人员提供一个查阅公众报告并分配资源进行核实的优先级。

为了达到这样一个目的，我们开发了两个模型。第一个，是在已有的**回归分析模型**中，加入了空间的差异性；第二个，是我们自己设计的**错误分类概率模型**，把一大份报告中的信息归类，并提供了信息出现错误的概率。在本文中，我们分别利用两种模型，预测胡蜂未来在时间和空间上的发展趋势，并且完成了对公众报告的筛选，给出了查阅的优先级。对未经处理的报告进行了有效信息的筛选，并比较了结果。这种双管齐下的方法帮助我们理性分析了胡蜂危机，并且提供了信息筛选的办法，为华盛顿确定了政府干预的策略。

1.1 problem statement

Problem 1: 证明胡蜂的数量和栖息地点随着时间的变化而变化，利用相关性去解释变量间的关系，精度如何？

Problem 2: 因为大多数的公众报告都是错误的，所以需要创建一个模型用来对报告的错误进行分类，using only the data set file provided, and (possibly) the image files provided, 并预报每一类导致结果为negative的可能性。

Problem 3: 利用我们创建的模型要对报告进行一个筛选，得到一个优先度，专业研究人员可以据此优先分配资源对最有可能为positive的报告进行调查。

Problem 4: 随着时间的推移，胡蜂的数量也随之变化，那么我们应该怎么更新自己的模型，以及更新的频率如何？

Problem 5: 利用我们的模型，当满足什么条件时，华盛顿就已经根除了这个害虫？

1.2. Our Model

首先我们建立的了一个回归模型，，错误分类的模型，计算了它的概率，验证了可靠性，建立一个评分机制，给出了更新的方法和频率，并给华盛顿政府提供了一份报告。

受到已有的CE-VADER方法，我们发明了通过文本中特征词组来筛选的方法

**2. background**

2.1 胡蜂外观生理等特性（加入参考范文）

外观

生命周期：移动范围、生殖

2.2 胡蜂的食性

胡蜂可能对本地蜜蜂的影响（比如说杀死蜜蜂，出现尸体）

2.3 washington的地理气候特征（宜居地）

胡蜂生存的环境气候特征，地形等因素

**3. Foundations of the Models**

3.1 Assumptions

1华盛顿整个大洲的气候适宜胡蜂的生存，胡蜂可以在华盛顿任何地区筑巢。

2在高风险地区1，蜜蜂的死亡如果保留着尸体，则都视为被胡蜂所杀。【背景跳跃】

3 报告人在向政府汇报前对胡蜂的特征有一定的基础认识

4 不考虑自然灾害导致的蜂巢迁移

5 不考虑胡蜂对于地形的选择

1.高风险地区是指胡蜂巢穴方圆8公里以内

3.2 Data Preprocessing

1. 清理了2019年前的数据

2. 清理了在加拿大温哥华岛的特殊点

3. 将global ID【2】和图像【3】合并成为一个文件

3.3 Nomenclature

X 月份

Y 黄蜂事件出现次数

rs spearman 相关系数

H1 胡蜂出现事件发生的次数和时间是显著相关的

H0 零假设

N 样本总体

P1*i* 报告的地域类错误概率

P2*i* 报告的时间类错误概率

P3*i* 报告的文本类错误概率

Pi 每个报告的错误概率

Pe 归一化后的错误概率

4.模型的建立

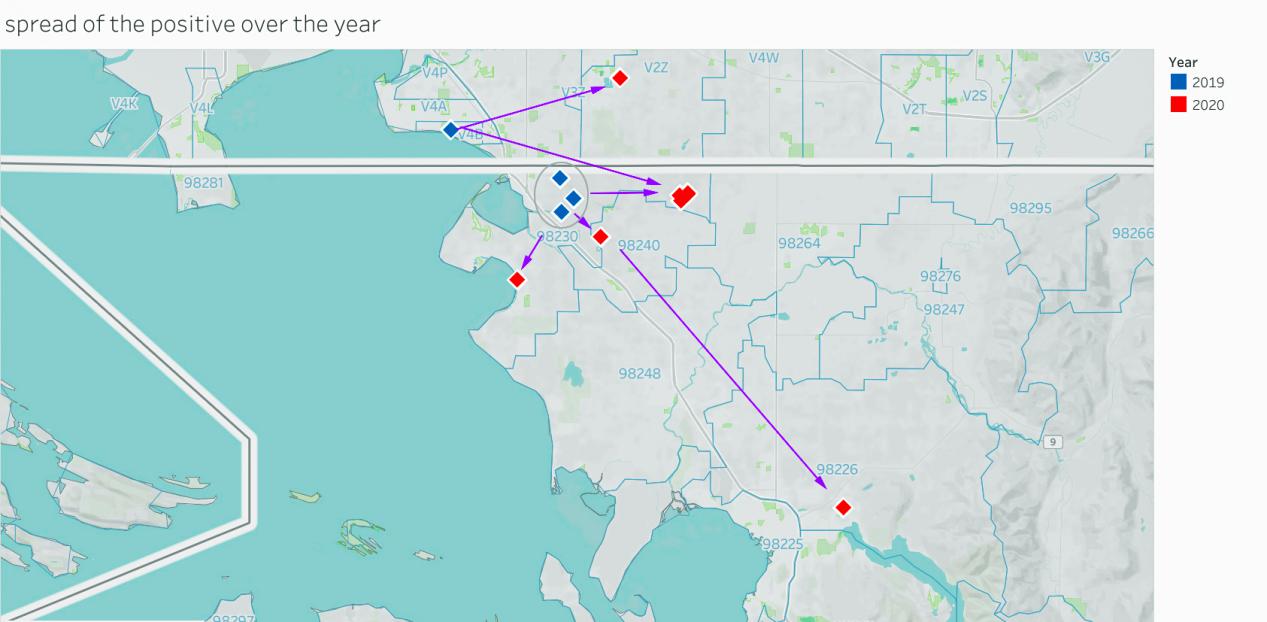
4.1回归模型

为了解决随着时间的推移,这种害虫的传播是否可以预测,考虑到害虫的传播不仅是时间上的数量增加，也可能是地域上地盘的扩张。所以在已有的回归模型上我们考虑了空间的差异性，研究了害虫栖息地和出现次数与时间的相关性，并做了初步预测

4.1.1地域扩散

背景2.1可知，一只新蜂后在30公里的范围内筑巢（这里要填写证据证明胡蜂的移动地点随时间的变化而变化）。在给定的positive数据中，我们绘制了2019年初到2020年末的胡蜂的一个扩散趋势，由图1可知，胡蜂的移动地点随时间的变化而变化。

（解释图像，蓝色，红色分别代表什么）



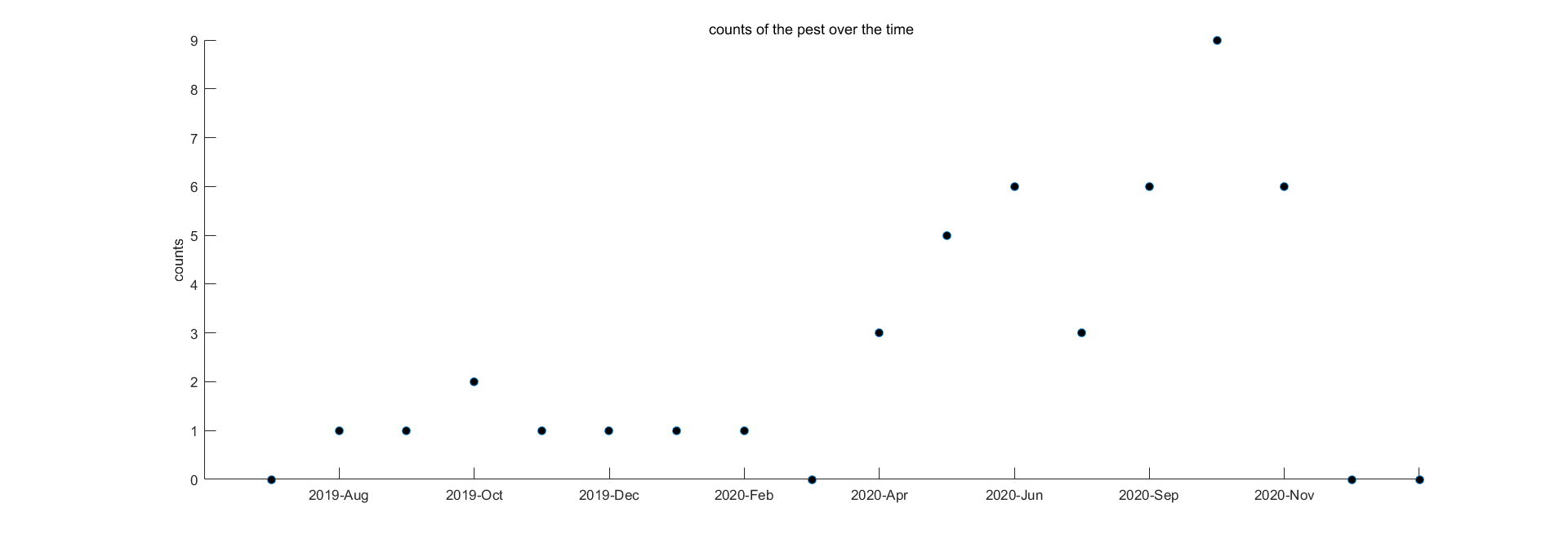
4.1.2 蜂数增加（数据处理）

材料中给出的report呈positive的例子只有14例，已知的胡蜂出现事件的样本太少，不足以说明相关性。为了得到更多的样本，我们采取了如下措施：

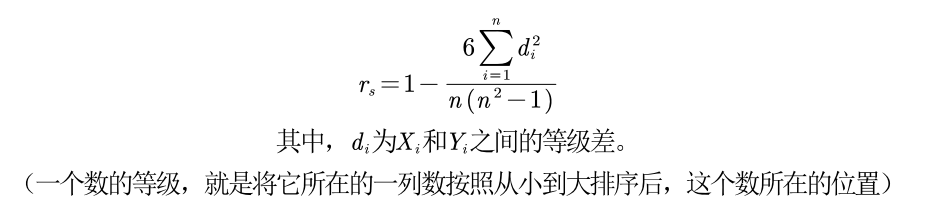
1. 根据图1可知，胡蜂的巢穴会随时间而变化，可以证明其具有地域性，同时，根据材料我们知道胡蜂的活动范围为方圆8公里。综合以上因素，我们认为胡蜂巢穴出现地方圆8公里都应该为高风险地区。
2. 利用matlab进行经纬度距离换算【代码1】，把4440个目击报告【2】进行筛选，把处于高风险地区的报告筛选出来，最终得到255个处于高风险地区的报告
3. 在这些报告中，部分unverified里的文本直接或间接的表诉了胡蜂的存在，【附录A1中表示出了这些文本】所以我们再次人工筛选出unverified和negative文本中直接或间接的反映胡蜂存在的报告，最终得到46份胡蜂存在的报告。

4.1.3相关性分析

将得到的46份报告中的日期进行一个提取,以月份为单位，得到18个月的时间变量X和胡蜂出现事件的发生次数Y。以2019年到2020年的月份X为横坐标，胡蜂出现事件发生的次数Y为纵坐标，画出了一个散点图图2



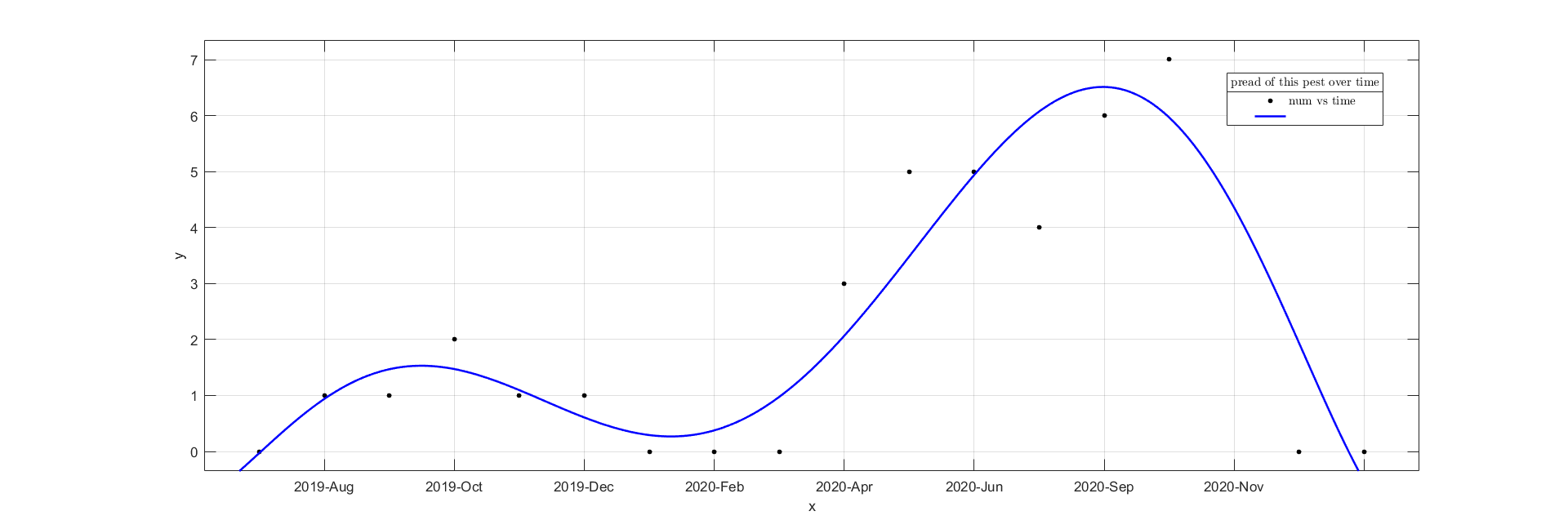
由图可大致判断X与Y不成线性关系，所以我们排除使用Pearson相关系数，采用spearman相关系数。



通过计算得到rs的值是0.46。通过查spearman临界值表【附录A2】，rs>0.41，得出在90%的置信水平下拒绝原假设H0，相信H1，认为胡蜂出现事件发生的次数和时间是显著相关的。

4.1.4数据拟合与元胞自动机

利用matlab对该图进行了多项式回归曲线拟合，拟合优度为83％，由图3可知，胡蜂出现事件的发生次数随着时间的变化而变化。胡蜂的出现次数在夏季达到一个峰值，在冬季几乎减为0，这表明胡蜂的繁殖满足背景2.1，符合预期效果。



通过图1和图3,我们可以判断，随着时间的变化，在保持现有条件的情况下，胡蜂的数量峰值会不断增加，但在冬天会死掉，只剩下蜂后进行冬眠。同时，我们对增加后的样本点，在空间上也进行了一个绘制，利用元胞自动机，模拟了胡蜂在空间上的变化趋势，如图4，随着蜂后的增加，胡蜂的巢穴在地域上也将不断扩大，人们目击到胡蜂的事件个数也将不断增加。

4.2 错误分类概率模型

因为大多数报道的目击事件都将其他黄蜂误认为是野黄蜂，所以为了节约政府资源，我们必须要设计了错误分类概率模型对报告进行了一个预判。

我们将报告人误判类型分成了三大类：地域类，时间类，文本类，并分别探究了不同报道出现误判的概率。

4.2.1 数据处理

在处理的公众报告中有positive、negative、unverified三大类，unverified是指专业人员无法识别图像给出准确的判断，但对于报告者来说，目击到的黄蜂要么是其它蜂（negative），要么就是胡蜂（positive）。所以在对报告者误判为胡蜂的错误类型分类时，我们需要删掉unverified的数据，因为unverified这一类报告不能直接反映目击事件的对错。

所以本模型的建立的数据是，删除unverified数据，只留下positive和negative的数据。

4.2.2 地域类错误概率

从背景2.1中我们知道，胡蜂的活动范围和筑巢范围，所以胡蜂的存在与地域有着紧密的关系。越靠近蜂巢的地域，胡蜂移动越活跃，目击到胡蜂的可能性越大，报告中出现错误的可能性越低。反之，离蜂巢越远的地域，目击者误判的可能性越大。

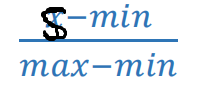
4.2.2.1 K-means++聚类算法

如果一开始就分开计算每一个报告点的位置到14个蜂巢的距离，在将14个距离求平均值，最后得出一个平均距离，采用这样的求均值的方法，计算量很大。为了简化计算，我们先利用**k-means++聚类算法**简化计算，算法流程见图5。

利用spss我们将positive14个数据分为了两类，见表1

4.2.2.2 TOPSIS算法

为了将每个报告点到蜂巢的距离转化为出现报告出现错误的可能性，我们采用了TOPSIS算法，用每个点到蜂巢的距离s除以最远的点到蜂巢的距离得到概率P0.

P0=

Latex小字备注：此时min是0，max是最远的点到蜂巢的距离

P0越大说明报告点位置离蜂巢越远，P0越小说明报告点位置离蜂巢越近

4.2.1.3 均值法

计算每个报告点到两个聚类点的P0，并将其相加求均值得到该报告的地域类错误概率P1*i*（*i=1，2，···，N*）

P1*i*越接近于1，说明该报告越可能出现错误，因为报告点离蜂巢非常远，几乎不可能出现胡蜂，反之，越接近于0，报告出错的可能性越低，因为报告点离蜂巢很近，胡蜂经常出没，发现胡蜂的可能性越大。

4.2.3 时间类错误概率

由positive样本数据中，胡蜂的出现与季节有关系，如图6，在夏季达到一个峰值，冬季几乎不可能出现胡蜂。

在其它条件都相同的情况下，我们计算出报告在每个季度出现错误的可能性，表2，夏季气温适宜，胡蜂大量繁殖，报告人目击到胡蜂的可能性越大，误判的概率越小，反之，冬季只剩下蜂后冬眠，报告人目击到胡蜂的可能性几乎不可能，出错的概率极大。

对于每一份报告，我们都可以按季节归类，给出一个时间类错误概率P2*i*（*i=1，2，···，N*）

4.2.4 文本类错误概率

对于报告人本身，如果对于胡蜂有一个清楚的认识，那么给政府的报告准确率越高，出现错误的可能性越低，但是如果报告人不太了解胡蜂，只是初步的把反映了部分甚至没有反映胡蜂的特征，那么报告的正确率性就低，出现错误的可能性就很高。

受到已有的CE-VADER方法【2020美赛C题】，我们采用了通过文本中特征词组来筛选的办法，对数据中的关键词提取，找到出现频率最多的10组描绘胡蜂特征的单词，对于每一份报告，每个词组出现的次数越多，那么报告人目击到的黄蜂越有可能是胡蜂，错误的可能性越小，反之，如果报告中提到的关键词很少甚至没有关键词，那么报告人目击到的黄蜂用以证明它为胡蜂的依据越少，越不可能是胡蜂，错误的可能性越大。

体型：Large、big、 huge、 giant、largest

长度：Inches、feet、long、inch、size

颜色：Orange、blank、 yellow、red、 brown、bright

肢体：Singer、leg、stripes

行为：Hit、 stung、 kill、 killer

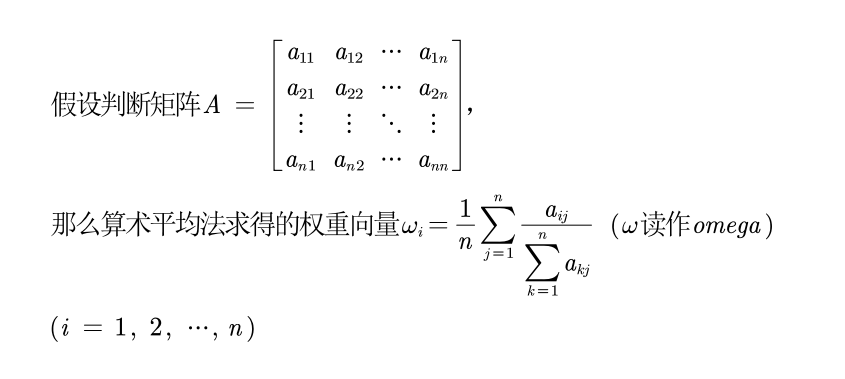
捕食：Dead、 decapitated、 loss

巢穴：Nest、ground、 holes

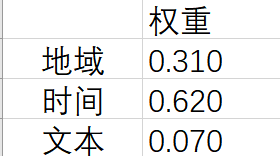
对于每一份报告，我们都可以通过关键词提取法，给出一个时间类错误概率P3*i*（*i=1，2，···，N*）

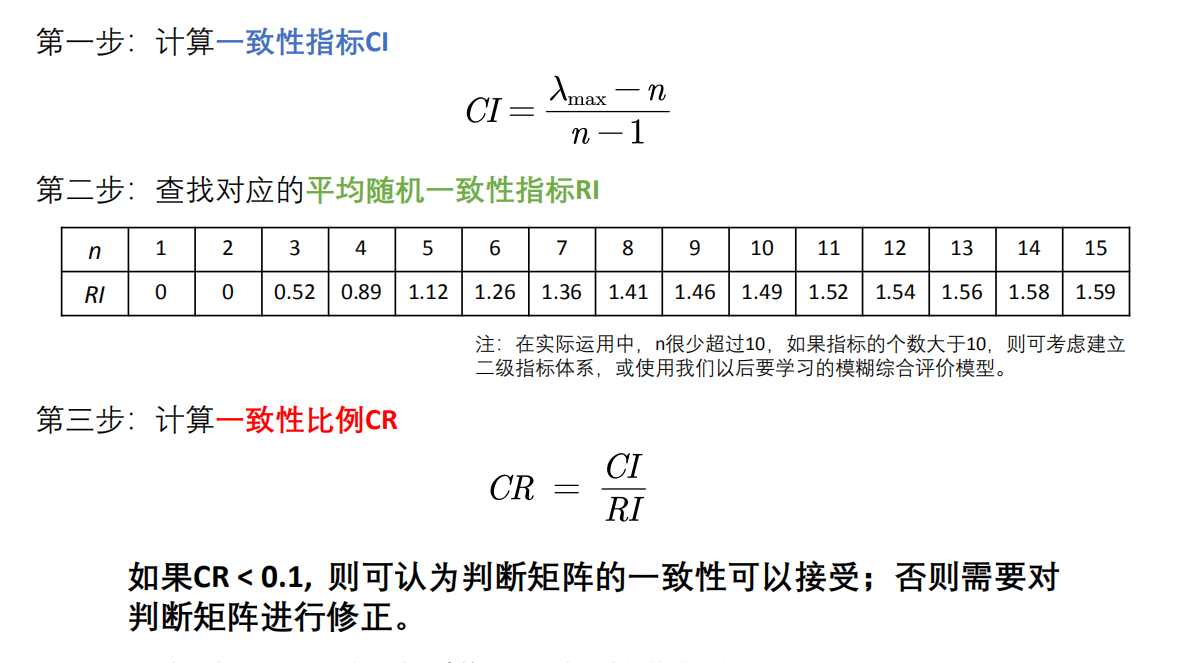
4.2.5 报告中的错误概率

在得出report的三个错误类型概率后，我们利用层次分析法求和给出了判断矩阵，一致检验后CR为 ，CR < 0.1, 则可认为判断矩阵的一致性可以接受，最后我们给出权重指标，地域为 ，时间为 ，文本为 。









对所有的报告错误类型概率进行加权，最后求出每一个报告总的错误概率p，p值越大，错误的可能性越大，结果越不可能为positive，反之，p值越小，错误可能性越小，报告人越有可能目击到胡蜂。

4.2.6 检验

为了检验错误分类概率的合理性，我们利用反证法，如果在错误可能性（p值）很大的情况下，报告为positive的案例很少，那么说明模型可靠。

我们首先通过TOPSIS方法再一次将数据中所有的报告的错误概率归一化Pe，范围0~1.然后作出了positive案例与错误概率的图象

如图6，在错误概率靠近1时，positive的案例为0，所有的positive案例全部集中在小于25%的错误概率内，其中有6例都集中在出错概率为5%左右，这一方面说明我们模型的可靠性；另一方面说明模型具有一定的筛选性，在案例错误概率大于25%时，基本可以判断为negative。

4.3 模型的实际用途

通过建立错误概率分类模型，我们可以对之后发来的报告通过不同类别的筛选，得到报告误判的可能性大小，把错判的可能性从小到大进行一个排列，为研究人员提供一个查阅的优先级。

基于原模型，我们对于任何一个报告，都可以给出它的错误概率，如果错误概率低，那么专家科研人员便可以优先对图片进行审查，不能识别则unverified，图像识别之后，如果为胡蜂就为positive，不能识别就为negative。

（给出流程图）

5 模型优化

5.1模型的更新

为了帮助政府更快的控制住胡蜂的传播，我们的模型要按照胡蜂的传播规律进行更新。

针对地域类的错误概率，如果在华盛顿政府还没有采取措施的情况下，随着时间的推移，胡蜂在空间上会进行一个扩散（跳转到元胞自动机），随着蜂巢的增多，k-means++算法需要聚的类就不只2类了，类别逐年增多，那么报告点到这些蜂巢的距离也会逐渐缩短，那报告出现错误的可能性也就降低了许多。

针对时间类的错误概率，随着positive样本数量的增加，我们便可以更精准的预知每个季度出现胡蜂的概率。当样本数量足够大时，我们甚至可以预知出胡风在每个月甚至每一周的出现概率，根据报告日期，我们会有一个更精确的胡蜂出现事件的概率。

对于文本类的错误概率，随着科技的发展，研究人员观察到的胡蜂新的习性或特征，也将会更改胡蜂出现事件的概率。

根据实际情况，在空间上，至少半年应该更新一次，每半年按照胡蜂在空间上的传播，人为增加类别；在时间上，每年要对季度进行更新，当positive数量超过100时，可以对每一月求事件发生概率；对于文本，一旦发现了新的胡蜂特征，就立即跟新文本错误类型。

5.2 根除胡蜂

在华盛顿政府采取严厉防控措施的情况下，胡蜂的数量应该会得到控制，出现positive的案例也会逐渐减少，所以我们的模型也要随之更新。

在地域方面，胡蜂在地域上的聚类也会减少，最后甚至不需要聚类，如果最后，蜂巢只剩下一个了，甚至没有了，说明胡蜂已经快被根除了；在时间方面，由于亚洲大黄蜂寿命为一年，如果在一年的周期中，经过优化后的模型筛选，对优先级高的报告进行筛选后，没有发现一例为positive的报告，则可以认为亚洲大黄蜂已经被根除。

5.2 Eradicate Hornet

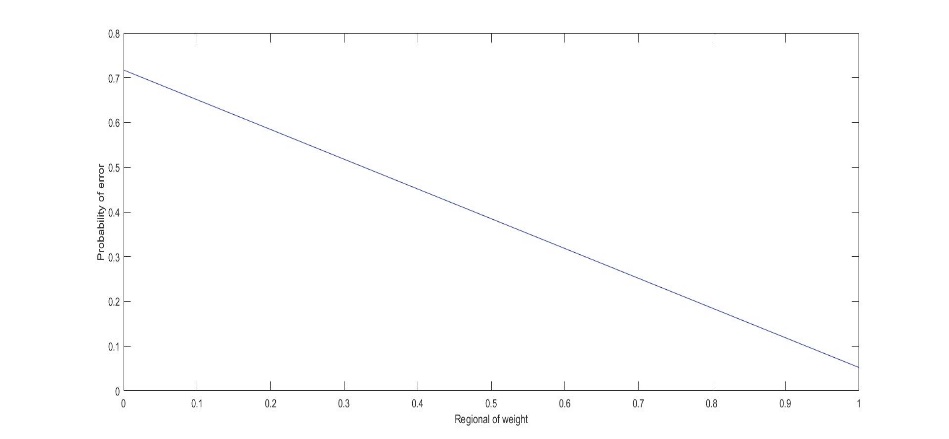
With the strict measures taken by the Washington government, the number of Asian Giant Hornet should be under control. The case of ‘positive’ also gradually reduce. So our model need be updated as well.

In terms of geography, The clustering of Asian Giant Hornet gradually reduce, finally, clustering is not even needed. In the end, there was only one nest left, not even one.it means that the pest has been eradicated. In terms of time, Because of the life of the Asian Giant Hornet only 1 year. So, after using our model to filtrate the report and judge, if in one year, the number of ‘positive’ is 0, we can thing that the pest has been eradicated in Washington State.

6.模型分析

6.1 Sensitivity Analysis

We chose one of the “positive”, witch is the 1296th of all specimens, his probability of error is 18.9%(after normalization),we take his value as the dependent variable, change the value of regional weight, meanwhile temporal weight also changes. Observe the change of the result, and analysis

****

It can be seen from the image，the change of the weight has a huge impact on the results, and the influence is uniform.

Therefore，In the actual evaluation，need to be evaluated by many experts and pass the consistency test, Take different values and find the most suitable weight. So that can adapt to the actual situation.

6.2 Strengths and Weaknesses

6.2.1 Strengths

1. 既有已存在的回归模型，又有自行创立的错误分类模型

2. 模型具有实用性，回归模型可以预测胡蜂的传播随时间的变化，the Information Evaluating Model can lead to prioritizing investigation of the reports

3. 随着样本的增多，模型可以不断更新，精准度提高

4. 模型具有合理性，所有的模型都是建立在一定的可能性上的。

5. 模型都是基于数据和数学算法建立，量化分析问题，具有数学依据，并且能够代码实现

6.2.2 Weaknesses

1. 胡蜂在地域上的移动，会受到气候、地形等等因素的干扰，这些都有可能导致胡蜂的实际传播并不一定符合模型。

2. 对于每一种错误概率类型，我们并不能给出最合适的权重。

1）．数学模型的缺点之一是模型可能过分简化,因而不能正确反映实际情况.（2）．模型受设计人员的水平的限制,模型无法超越设计人员对问题的理解.（3）．创造模型有时需要付出较高的代价

**7.Conclusion**

**8.Memo to the Chief Administrator（模板）**

**时间：January 29, 2019**

**To: Uttam Dhillon – Acting Administrator, Drug Enforcement Agency**

**cc: Dr. Nora D. Volkow – Director, National Institute on Drug Abuse**

**From: MCM Team #1901679**

**Subject: Strategies to Combat the Opioid Epidemic**

**9. Reference**

**Appendices**

**A Appendix**

**B Appendix**

**C Appendix**