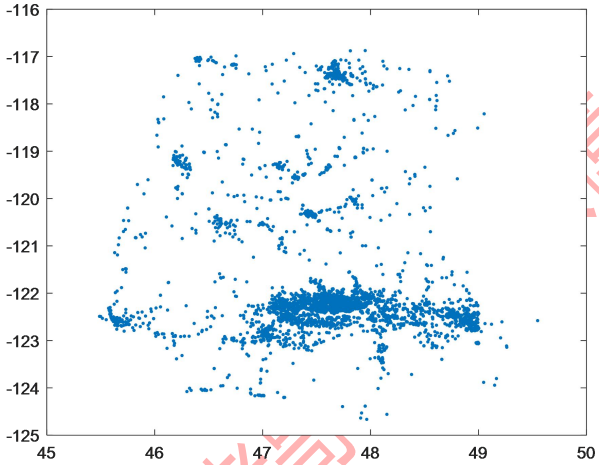
2020 数学建模美赛 C 题参考思路

对 2021MCM_ProblemC_DataSet.xlsx 数据集进行统计发现：Positive ID 意味着这是一只亚洲大黄蜂，Negative ID 意味着它被排除在外，Unverified 意味着缺乏信息而没有做出决定，Unprocessed 意味着还没有做分类。但实际上 Positive ID 只有 14 份，而 Negative ID 有 2069 份，Unverified 有 2342 份，Unprocessed 有 15 份，意味着大部分的报告均不是亚洲大黄蜂。对所有数据集上的经纬度进行可视化得到：



对于这 14 份 Positive ID 的经纬度进行统计发现这些被验证为真的亚洲大黄蜂的样本 Latitude 范围：[48.7775,49.1494]，Longitude 范围：[-123.9431,-122.4186]。所以被证明是亚洲大黄蜂的相对于被报告的仅处于一个较小的范围内。

首先需要解决并讨论这种有害生物在一段时间内的传播是否可以预测，把 14 个 Positive ID 按照时间顺序排列起来，采用 GM 模型、神经网络模型、时间序列模型等等对传播范围进行预测，并对拟合优度进行计算。而且需要记住题目中说一只新蜂后的筑巢距离估计为 30 公里，需要满足这个条件。题目中说大多数报道的目击事件都把其他黄蜂误认为是胡蜂。需要你仅使用提供的数据集文件和（可能）提供的图像文件来创建、分析和讨论一个预测错误分类可能性的模型。可以分别对提供的图像文件和数据集文件上进行训练和预测，我们想要通过目击

事件的报告来研究预测错误的可能性，因此在数据集2021MCM_ProblemC_DataSet.xlsx上，实际上有用的只有Detection Date、Notes、Lab Status、Latitude、Longitude。其中以Detection Date、Notes、Latitude、Longitude作为输入，Lab Status作为输出进行训练，将数据集划分成训练集和预测集来分别验证预测错误的可能性即模型的准确率。因为涉及到分类预测，那常用的SVM就最合适不过了，这里推荐大家可以下载一个台湾大学林智仁(Lin Chih-Jen)教授开发的一个软件包LIBSVM，python，matlab，java都可以支持，通过svmtrain和svmpredict函数就可以方便的实现训练和预测，而且还可以得到预测的准确率以及一些其他参数指标。图像识别的话像深度学习神经网络等等，跑跑网上的模型代码。在拿到目击的报告后，带入构建好的模型里，会得到一个预测的估计概率，根据概率即可制定优先顺序。随着时间的推移，如果有其他新报告应该将其加入到训练集中重新对我们构建的模型进行训练。

之后对华盛顿州已经消灭了这种有害生物的可能性进行探讨，可以对历史数据进行预测，直至预测报告次数为0的时间点仍然没有报告且被验证确切为这种有害生物的，即可认定为华盛顿州已经消灭了这种有害生物。

对于模型优化部分，诸如GM模型可以采用时间序列模型对拟合残差来预测进而提升预测效果，通过误差来作为衡量标准。诸如AR模型的话可以进一步构建ARMA模型提升拟合效果。诸如SVM，可以采用诸如IC-OM、MC-OM之类的方法来进行优化。