2021美赛C题

Task A 虎头蜂的传播的预测

- 验证传播与时间的相关性

  首要的问题是怎么反映大黄蜂的传播。

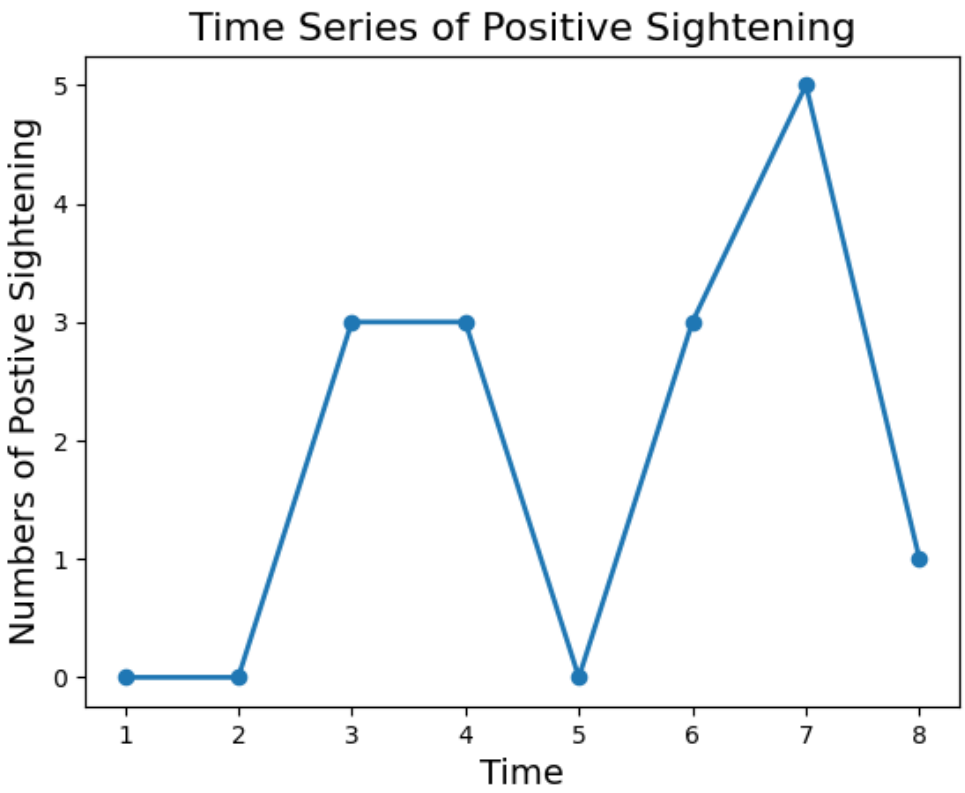
  我们采取的方案是提供统计大黄蜂的目击数量来间接说明大黄蜂的增长，从而反映大黄蜂的数量随时间变化情况。

  由于正向样本数据过少，很多月不是1就是0，不适合以月份为单位进行验证，但若以年为单位进行验证，存在数据不足的问题，故这里选取以三个月作为时间窗口，进行有效目击的结果统计。

	ID	Date	Latitude	Longitude
3	{124B9BFA-7F7B-4B8E-8A56-42E067F0F72E}	2019-09-19 00:00:00	49.149394	-123.943134
1069	{7F3B6DB6-2ED4-4415-8DC2-3F03EC88F353}	2019-09-30 00:00:00	48.993892	-122.702242
1	{5EAD3364-2CA7-4A39-9A53-7F9DCF5D2041}	2019-10-30 00:00:00	48.971949	-122.700941
924	{F1864CC3-508C-4E60-9098-B158AB413B03}	2019-11-13 00:00:00	49.025831	-122.810653
0	{5AC8034E-5B46-4294-85F0-5B13117EBEFE}	2019-12-08 00:00:00	48.980994	-122.688503
956	{1C6D0EAB-F68D-411D-974E-1233618854CC}	2020-05-15 00:00:00	49.060215	-122.641648
834	{AD56E8D0-CC43-45B5-B042-94D1712322B9}	2020-05-27 00:00:00	48.955587	-122.661037
1011	{FC6E894B-F6DF-4FDC-853A-D7372D253988}	2020-06-07 00:00:00	48.777534	-122.418612
3279	{A717D86F-23E9-4C8C-9F12-198A71113E93}	2020-08-17 00:00:00	48.927519	-122.745016
4127	{2138197A-F5CF-4308-93E2-62EA6F84D098}	2020-09-21 00:00:00	48.984269	-122.574809
4338	{AA461F47-1B2B-4EA1-8154-ECF70B55A334}	2020-09-28 00:00:00	48.98422	-122.574726
4208	{DEF5D82B-E326-41A5-9B6C-D46DCD86950C}	2020-09-29 00:00:00	48.984172	-122.57472
4206	{0FAC3767-EAC4-477A-B5F0-24AF8A40BD09}	2020-09-30 00:00:00	48.979497	-122.581335
4207	{BEAC832C-0783-414A-9354-C297F38570AD}	2020-10-01 00:00:00	48.983375	-122.582465

  统计结果如下（不删除重复观测）：

标号	时间	目击次数
1	2019.1-2019.3	0
2	2019.4-2019.6	0
3	2019.7-2019.9	3
4	2019.10-2019.12	3
5	2020.1-2020.3	0
6	2020.4-2020.6	3
7	2020.7-2020.9	5
8	2020.10-2020.12	1



弱平稳数据的概念：指固定时间和位置的概率分布与所有时间和位置的概率分布相同的随机过程。其数学期望和方差这些参数也不随时间和位置变化。换言之数据没有随时间变化的趋势，也就无从预测。

弱平稳

与强平稳对应的是弱平稳，其应用比较广泛，它有三个要求：

- 对于任意时期的时间序列X_t，有$E(X_t)=\mu$，即序列均值为常数；
- 对任意时期的时间序列X_t，其二阶矩存在（二阶矩即方差）；
- 对任意时期的时间序列X_t、任意的整数h、任意的阶数l，都有$\gamma_l(X_t)=\gamma_l(X_{t+h})$，也就是说，当指定了两个时间点的距离$h$后，两组数据的$l$阶自协方差不会随着时间波动，仅与阶数有关。

简而言之，一组时间序列数据的**均值恒定、方差始终存在、自协方差不随时间波动**，即可认定其为弱平稳序列。

相关性检验：

- 单位根检验（ADF）（自相关性检验）。

  对于一个普通的自回归（AR）过程。

  t时刻的输出可以表示为$y_t=\sum_{i=1}^{p-1}\zeta_i\Delta y_{t-i}=x_t^{'}R+\epsilon t$，其中$R=(\zeta_1,\cdots,\zeta_{p-1},\rho)$。

  假设：$H_0:|\rho|\geq 1, H1:|\rho|<1$其中$H_0$表示序列平稳，$H_1$表示序列不平稳。

  检验统计量为：$t=\frac{\hat{\rho}-1}{\hat{\sigma_p}}$

  这里的样本数据为统计次数，$Y=(y_{p+1},\cdots,y_8)^{'}$

$$X=\begin{bmatrix}\Delta y_2 & \Delta y_3 & \cdots & \Delta y_p & y_p\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \Delta y_{8-p+1} & \Delta y_{8-p+2} & \cdots & \Delta y_7 & y_7\end{bmatrix}$$

则有$\hat{R}=（X^{'}X）^{-1}X^{'}Y$

$\hat{\sigma_p} = \sqrt{s_t e_i^{'}(X^{'}X)^{-1}e_i}$

其中

$$e_i = [0 \quad \cdots \quad 1]^{'}$$

$$s_t = \frac{1}{8-2p}(Y-\hat{Y})^{'}(Y-\hat{Y})$$

通过计算检验统计值，选定置信度后，查表可以得到检验统计量在置信区间内的概率，还有一种方式是计算统计值不在置信区间里的概率，记为$p-value$。若其远小于置信度，则有充分的把握可以认为H_0成立，否则认为H_1成立。

选取置信度0.05，计算得$p-value=0.632$，远大于置信度，因而认为原序列是不平稳的，即原序列是和自身历史有关，也就和时间有关了。

- 计算Spearman相关系数（直接检验和时间的函数关系）

  将时间看为自变量，目击次数看为因变量，直接计算其相关性大小。由于它们之间的关系不一定是线性关系，故这里没有选择pearson等计算线性相关系数的方法。

  > 介绍不贴了。。写着太慢了。大部分参考的百度百科，也可以随便找一篇英文文献水一水。

  计算可得：

  有个$p-value=0.206$，远大于置信度水平0.05，因此拒绝没有相关性的原假设。(这里假设来得很迷，没有找到具体资料，但是有个假设检验的过程)

  相关系数$r=0.500$。证明二者之间呈现中度正相关。

- 基本假设
  - 亚洲蜂最早于3月1日开始出现（北半球春季开始），除了蜂后，其他蜜蜂最晚死于12月1日（北半球冬季开始）（根据题目所提供的pdf材料，该蜜蜂一年生，春季蜂后苏醒后开始找地方筑巢，种群数量缓慢成长并于八月达到顶峰），没有受精的蜂巢当年冬天全部死亡。
  - 九月蜂后开始进入产卵期，约0.35的雌蜂会被受精，然后来年会外出建立新巢。（C题附带的PDF材料里有写）受精雌蜂都能顺利存活并且建立蜂巢。
  - 人们没有对该物种采取非常强硬的防治措施。在模型预测范围内，亚洲大黄蜂的生存环境没有受到太大影响。同时，作为入侵物种，亚洲大黄蜂暂时不会遇到天敌（自由繁殖）。
  - 新生受精雌蜂新建曹距离遵循泊松分布。
  - 由于水域面积过小，含有水域的地方仍然视为陆地。

- 研究区域
  - 截取北纬45度到50度，西经115度到125度的矩形区域作为研究对象。
  - 经度纬度都每隔0.2度画线，把图分割成一个个小矩形。（根据球面距离HaverSineg公式$haversin(\frac{d}{R}) = haversin(\varphi_2-\varphi_1)+\cos(\varphi_1)\cos(\varphi_2)haversin(\Delta\lambda)$，其中$haversin(\theta)=\sin^2(\frac{\theta}{2})=\frac{(1-\cos\theta)}{2}$，$R$为地球半径，（$\varphi_1,\varphi_2$为两点的维度，$\Delta\lambda$是两点精

度的差值）这样划分，可以保证一个矩形的中心距离相邻的八个矩形大约在20~30km（C题附带的PDF材料里），为一个受精雌蜂飞行的最大距离，可以简化后面的计算。

- 基本想法

使用元胞自动机（CA），将地图划分成一个个元胞，由于其繁殖周期为一年，故以年为单位，进行大黄蜂传播的仿真。

> 这里可以贴一些关于元胞自动机的基础知识。最好是英文文献混一混。

元胞自动机由以下要素构成。

- 空间：这里的空间是二维空间，即地图上划分出来的网格。
- 状态集：定义每个网格的状态为$s_{i,j}$，代表每个网格内亚洲大黄蜂活动巢穴的数量。
- 邻居：某一元胞的邻居定义为与它相邻的8个元胞。
- 演化规则：

  演化规则可以概括为$s_{i,j}^{t+1} = f(s_{i,j}^t, s_{neighbor_{i,j}}^t)$

  假设服从雌蜂飞行距离服从泊松分布，飞行过程的8个取向等概率。由于泊松分布的概率会在参数λ附近达到最大，故简单地取$\lambda = \frac{30+0}{2} = 15$。同时，一个网格内的亚洲大黄蜂的种群数量限制limit主要和纬度有关，可以简单地假设成线性负相关。（根据文献【1】（https://www.nature.com/articles/s41598-017-06212-0#citeas）。同时设一只亚洲大黄蜂繁殖期内共产生了egg只卵，每只卵成功孵化的概率为$alive_prob$，这里取$egg = 20000, alive_prob = 0.00001$。根据C题附赠的PDF材料，其中有$fert_prob = 0.35$的雌蜂被受精。

  主要有以下演化规则：

  1. $0 \leq s_{i,j} \leq limit$
  2. 仅考虑一个元胞周围8个元胞以及自身状态对下一状态的影响，由于东南西北的邻居元胞和东南、西北、东北、西南的邻居元胞到中心元胞的距离稍微有所不同，因此分开考虑其贡献。记雌蜂沿东南西北的方向飞而不飞出目前网格的概率为p_1，在这四个方向上的邻居的集合为S_1，沿东南、西南、东北、西北方向飞而不飞出目前网格的概率为p_2，在这四个方向上的邻居的集合为S_2。有

$$s_{i,j}^{t+1} = fert\_prob * eggs * alive\_prob * (s_{i,j}^t num(s_{i,j}) * (p_1 + p_2) + 0.25 * \sum_{k=1}^{2}(1 - p_k) * \sum_{(i,j)\in S_k} s_{i,j}^t)$$

初始数据采用2019年的几个目击报告作为起始，报告处为大黄蜂的入侵点，最初只有一个活动的蜂巢。

这里可以贴个表格把2019年的数据列出来	

	ID	Date	Latitude	Longitude
0	{5AC8034E-5B46-4294-85F0-5B13117EBEFE}	2019-12-08	48.980994	-122.688503
1	{5EAD3364-2CA7-4A39-9A53-7F9DCF5D2041}	2019-10-30	48.971949	-122.700941
3	{124B9BFA-7F7B-4B8E-8A56-42E067F0F72E}	2019-09-19	49.149394	-123.943134
924	{F1864CC3-508C-4E60-9098-B158AB413B03}	2019-11-13	49.025831	-122.810653
1069	{7F3B6DB6-2ED4-4415-8DC2-3F03EC88F353}	2019-09-30	48.993892	-122.702242