北京師範大學

本科生毕业论文（设计）

毕业论文（设计）题目：

基于转移概率的地震预测模型

部　　院　系：　数学科学学院

专　　　业：　数学与应用数学

学　　　号：　201711130137

学 生 姓 名：　李子文

指 导 教 师：　何辉

指导教师职称：　教授

指导教师单位：　数学科学学院

2021 年 4 月 13 日

北京师范大学本科生毕业论文（设计）诚信承诺书

本人郑重声明： 所呈交的毕业论文（设计），是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

本人签名: 年 月 日

北京师范大学本科生毕业论文（设计）使用授权书

本人完全了解北京师范大学有关收集、保留和使用毕业论文（设计）的规定，即：本科生毕业论文（设计）工作的知识产权单位属北京师范大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许毕业论文（设计）被查阅和借阅；学校可以公布毕业论文（设计）的全部或部分内容，可以采用影印、缩印或扫描等复制手段保存、汇编毕业论文（设计）。保密的毕业论文（设计）在解密后遵守此规定。

本论文（是、否）保密论文。

保密论文在 * 年 * 月解密后适用本授权书。

本人签名: 年 月 日

导师签字: 年 月 日

基于转移概率的地震预测模型

摘要

本篇文章利用2009年至今全国5级以上地震数据，利用K-means聚类算法将全国分成了四个地震区域，把地震当作在这些地震区域间随机游动的质点，且其转移不依赖于两次以前发生的地震。统计各个区域地震发生的频率，计算出地震在各个区域之间的转移概率。利用地震的一阶以及二阶转移概率，建立预报测度用以预测地震发生在各个区域的可能性。利用此预报测度，本文预测了国内 2020 年 7 月 26 日至 2021 年 3 月 24 日发生的 11 处 5 级以上地震发生的区域，其中有 8 次发生在预测的第一可能区域，2 次发生在预测的第二可能区域，仅有 1 次错报，预测结果较好。

关键词：马尔可夫链；地震预测；K-means 聚类；地震带划分

ABSTRACT

This article uses the national earthquake data of magnitude 5 and above from 2009 to the present, and uses the K-means cluster algorithm to divide the country into four earthquake regions. The earthquake is regarded as the mass point, randomly walking between these earthquake regions, and their transfer does not depend on earthquakes occurred twice before. We count the frequency of earthquakes in the divided regions and calculate the transition probability of earthquakes between regions. Using the first-order and second-order transition probabilities of earthquakes, a prediction measure is established to compare the possibility of earthquakes in various regions. Using this prediction measure, this article predicts 11 areas where earthquakes of magnitude 5 or higher occurred in China from July 26, 2020 to March 24, 2021, of which 8 occurred in the most possible area and 2 occurred in the second possible area, there was only one error.

Key Word: Markov chain; Earthquake Prediction; K-means Cluster; Earthquake Region

目录

1. 研究背景

地震又名地动，是一种由于地壳释放能量时产生震动，并会产生地震波的自然现象。全球每年发生约 550 万次地震，造成大量人员伤亡和财产损失，被称之为群害之首。

由于身处两大地震带（环太平洋地震带和亚欧地震带）之间，中国是遭受地震灾害最为严重的国家之一。20 世纪以来，全球有约 120 万人因地震而失去生命，我国遇难人数占全球的二分之一。中国地震活动频率高、强度大、震源浅、分布广，且随着城市化的迅速推进，地震致灾日益严重。[1]所以为了降低地震带来的损失，地震预测技术十分重要。

针对地震预测，知名地质学家张国民指出：“地震预报现阶段仍未能走出经验科学的不确定性。”地震难以预测原因有三：

一, 地震过程的复杂。地震是地壳运动的产物，但现在人们对地壳的认知少之又少，所以我们现阶段缺少对地震发生规律和机理的认知，这很大程度上限制了我们对地震的预测能力。

二, 现有技术无法深入地壳内部。地震普遍发生在地下十多公里的地方，以人类现阶段的技术，无法把仪器深入地下探究其源头。人们现在能做的只有在地面上设置数量有限的台站，对于地下世界的认知只是经验性的推测。用这种推测去揣摩地震的过程，无疑是不可靠的。

三, 地震发生的低频性。虽然在全球范围内来看地震发生次数很多，但是大部分都发生在海洋和人迹罕至的陆地，人们所能确切记录的地震数量其实是十分有限的。研究对象的稀少，将地震预测这门科学锁死在了一个较低的水平。

所以综上所述，想要从机理上预测地震成为了一项近乎不可完成的任务。目前，地质学家们期望找到地震发生的先兆或识别地震发生前的地球运动趋势。成功的案例有于 1975 年 2 月 4 日发生的海城地震。在海城地震前，我国地震部门做出了中期预报和短临预报，当地政府做出了正确而及时的防震措施并且在地震发生前转移了重要的物资设备，最大程度上的减少了当地的人员伤亡和财产损失。这次成功的预测震动了地震学界，这是人类在地震面前由被动向主动迈出的重要一步，开创了人类地震短临预报的先河，给予人们能成功预测地震的希望。

但是，海城地震的成功预测也是不可复制的。最为关键的是，海城地震的前震序列非常明显，依据当时专家的“小震闹，大震到”的经验，地震专家把地震发生前的异常现象定为地震前兆。但仅仅是一年之后发生的唐山大地震，就否决了海城地震的成功经验——唐山地震没有前震！所以从根本上来讲，通过前兆进行预报是经验性的，现阶段我们对地震前兆的认知也非常局限，没有普遍试用的

意义。

所以，我们考虑建立一个地震预测模型，能够程序化的预测地震的发生。我们把地震视作一个做随机游动的质点，在不同地震区域以某种概率有规律的转移着。我们在这里做出贯穿本文的假设：地震转移的随机性是二重相依的，即如果已知最近两次地震发生的地点，那么下次地震发生的地点与更早发生的地震无关。我们做出这个假定，一方面是以往的数据表明下次地震与两次以前发生的地震的相依性非常小。[2]另一方面是地震资料相对较少，若考虑更复杂的转移，则可能会导致过拟合现象的发生，使模型丧失预测的意义。

我们的研究思路如下：首先利用 K-means 聚类的手段将自 2009 年以来全国发生的 5 级以上地震分类，并用轮廓系数判断其划分的好坏。列出地震转移频次表，用频率近似地代替概率，计算出一阶转移概率P_{jk}以及二阶转移概率P_{ijk}，用一阶、二阶转移概率的线性组合构建预报测度$M(k) = AP_{jk} + BP_{ijk}$，其系数$A$，$B$的确定将由我们的评分系统给出，预报测度$M(k)$越大意味着地震发生在$k$区的可能性越高。通过预报测度，给出下次地震发生的第一可能区域和第二可能区域。

2. 地震带划分

2.1 状态的划分

为了构建模型并进行地震预测，我们首先要完成的工作就是划分地震区域。

著名的地质学家李四光先生曾将我国划分为四大地震带，即：1、东南部的台湾和福建沿海；2、华北的太行山沿线和京津唐地区；3、西南青藏高原和它边缘的四川，云南两省西部；4、西部的新疆，甘肃和宁夏。现代普遍认为我国地震活动主要发生在五大区域，23 条小地震带上。国家地震局于 1981 年采用三级划分方案，将我国划分成 10 个地震区，13 个地震亚区和 30 个地震带。《中国地震烈度区划图》[3]将中国地震按危险性划分出了 5 个地震区和 25 个地震带。《中国地震动参数划分图》（2015）综合之前的经验，划分了 8 个地震区和 24 个地震带。

在本篇文章中，我将结合以往的经验，在令我们模型简洁有效的基础上，利用 K-means 聚类方法划分地震带。

2.2 K-means 聚类

2.2.1 K-means 聚类算法

K-means 聚类是一种经典的聚类方法[4]，其目标是将一个数据集分成 K 个类，

使得每个类中至少有一个对象，且每个对象仅属于一个类。

K-means 聚类是通过“相似性”将不同的对象放到同一个类里，而在我们的文章中，相似性为两次地震发生的距离，距离越大则相似性越小。我们可以从国家地震科学数据中心（https://data.earthquake.cn/）获取我国自 2009 年 2 月 20 日至 2021 年 3 月 24 日的 5 级以上地震数据（共计 361 条），数据集中包括地震发生的经度（lon），纬度(lat)。我们以欧氏距离（d）作为相似性的度量。在这里，我们规定：

$$d = \sqrt{lon^2 + lat^2}.$$

实现 K-means 聚类的方法步骤为：

1. 随机选取K个初始点$\alpha_1, \alpha_2, \dots, \alpha_K$，计算每个数据（向量）$x$，到这 K 个初始点的距离$d_j(x)$，j=1,2,…, K, 那么，一定存在一个整数i，s.t. $1 \le i \le K$且$d_i(x) = \min_j d_j(x)$。于是我们将数据x归于第i类（记为c_i）。
2. 将所有数据归类后，我们计算每个类的中心:

$$\alpha_i = \frac{1}{|c_i|}\sum_{x \in c_i} x.$$

重新计算每个数据到类中心的距离，并根据相似性将所有数据再次归类。

3. 多次重复上述步骤，直到聚类不再发生变化。

2.2.2 K 值的选取

K-means 聚类的 K 值选取十分关键，可以说是决定模型是否可靠的最关键的一步。传统的选择方法有按需选择，观察法，手肘法以及 Gap Statistics 方法。这些传统方法都具有一定的主观性，在聚类比较明显的情况下较为好用，但由于本文中的数据很难通过经验给出分类，且本文力求寻求一种解决地震预测问题的程序化方法，所以在本篇文章中我们将利用轮廓系数这一指标来选择 K 值。（注：我也用手肘法对 K 值进行了选取，但由于在构建模型时我们用的是后文介绍的轮廓系数选取 K 值，所以我将用手肘法选取 K 值的结论放到了附录三中。）

轮廓系数是评价聚类效果好坏的一种指标，由 Peter J. Rousseeuw 提出[5]。该指标综合考虑了一个聚类的簇内相似度和簇外相似度。一个好的聚类要满足簇内相似度尽可能的大（同一个簇内的数据距离小），不同的簇之间的间隔尽可能的大。基于这种考虑，轮廓系数应运而生。

轮廓系数S计算公式如下：

$$S(i) = \frac{b(i)-a(i)}{max\{b(i),a(i)\}}.$$

其中，$a(i)$代表内聚度，其计算公式如下：

$$a(i) = \frac{1}{n-1}\sum_{i\neq j} d(i,j).$$

其计算的是数据i与其类内的其他点的距离的平均值，n为数据i所在类内元素个数。

$b(i)$代表的是数据i与其他簇的平均距离（也可理解为不相似度）的最小值，其计算公式为：

$$b(i) = \min_{1\leq k\leq K} \frac{1}{|c_k|}\sum_{j\in c_k} d(i,j).$$

在这里，c_k指的是不包括数据i的其他簇，$|c_k|$指该簇中元素个数。

由轮廓系数的计算公式，我们可知轮廓系数S的取值范围为[-1,1]，越接近 1 表示聚类效果越好。

根据地质学的知识，我们知道我国可划分为 2 到 5 个地震区域。我们分别计算了$K = 2,3,4,5$情况下的轮廓系数S，结果如图所示：

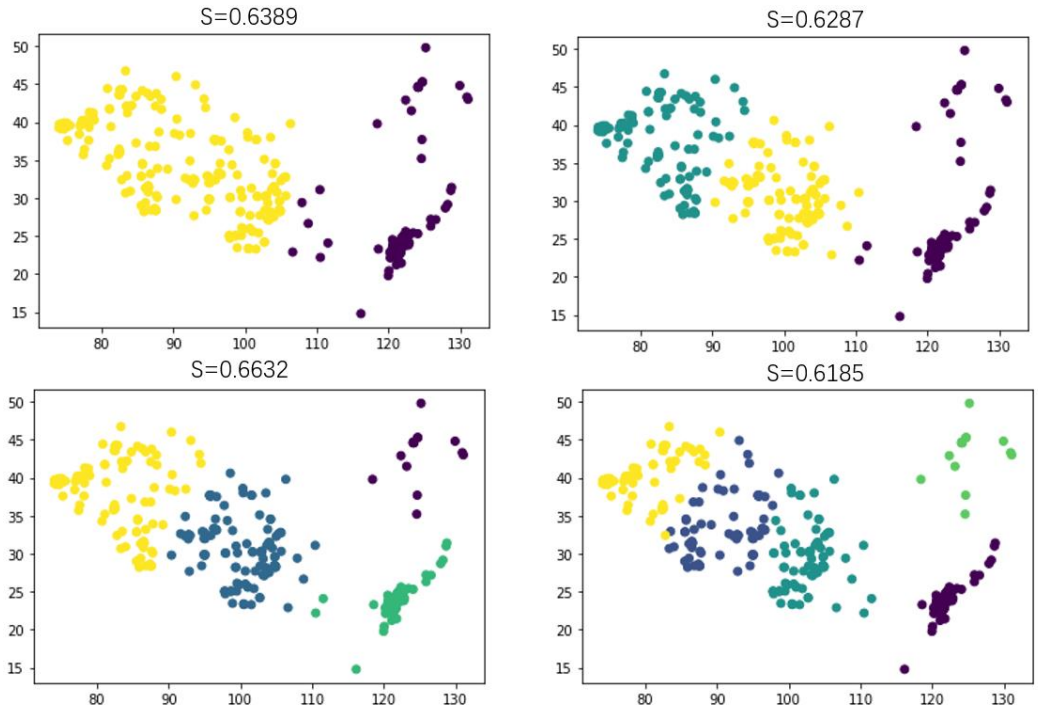

图 1 轮廓系数

由图 1 可知，当取$K = 4$时，即将我国划分成四个地震区域时，聚类效果最好，此时轮廓系数为$S = 0.6632$。

3. 模型的建立

3.1 基础知识

为了让读者方便理解后续的内容，在本节我将简要的介绍和转移概率相关的

背景知识。

定义一：设$\{X_n\}$为随机变量序列，状态空间为$E=\{i\}$。若对任意正整数m，k和任意的非负整数$n_1<n_2<\cdots<n_r<m$，以及任意的$i_1,i_2,\ldots,i_r,i,j\in E$，我们有如下式子成立：

$$P\{X_{m+k}=j|X_{n_1}=i_1,\ldots,X_{n_r}=i_r,X_m=i\}=P\{X_{m+k}=j|X_m=i\}.$$

那么我们得到$\{X_n\}$为Markov链。若E为有限集，$\{X_n\}$为有限状态Markov链；若E为可数集，则$\{X_n\}$为可数状态 Markov 链。

定义二：设$\{X_n\}$是 Markov 链，若记：

$$p_{ij}^k(m)=P\{X_{m+k}=j|X_m=i\},$$

其中k为正整数，m为非负整数，那么我们称$p_{ij}^k(m)$为$\{X_n\}$在时刻m从状态i出发经k步到达j的转移概率，称矩阵：

$$P^k(m)=\left(p_{ij}^k(m)\right),\qquad i\in E,j\in E,$$

为$\{X_n\}$从时刻m出发的k步转移概率矩阵。

转移概率矩阵$P^k(m)$有如下的性质：

（1） 每一个元素$p_{ij}^k(m)>0$

（2） $\sum_{j\in E}p_{ij}^k(m)=1$

定义三：设$\{X_n\}$的转移概率矩阵为$P^k(m)=\left(p_{ij}^k(m)\right)$，如果$\forall i,j\in E$，任意正整数$k$，$p_{ij}^k(m)$与$m$无关，那么我们称$\{X_n\}$是齐次 Markov 链，并记$p_{ij}^k(m)$为$p_{ij}^k$，称他为$\{X_n\}$由状态$i$出发经过$k$步到达状态$j$的转移概率。

下面，我们将在 Markov 链的基础上，扩展地给出（步长为 1 的）二阶 Markov 链的概念。

定义四：若$\{X_n\}$为一列随机变量，状态空间为 $E=\{i\}$。对于正整数m以及$i_1,i_2,\ldots,i_m,i_{m+1},i_{m+2}\in E$，我们若有下式成立：

$$\begin{aligned}&P\{X_{m+2}=i_{m+2}|X_1=i_1,\ldots,X_m=i_m,X_{m+1}=i_{m+1}\}\\&=P\{X_{m+2}=i_{m+2}|X_m=i_m,X_{m+1}=i_{m+1}\}.\end{aligned}$$

则$\{X_n\}$为二阶 Markov 链。

我们记$p_{ijk}(m)=P\{X_{m+2}=k|X_{m+1}=j,X_m=i\}$为二阶转移概率，若$p_{ijk}(m)$与$m$无关，那么我们称这个二阶 Markov 链是齐次的。

由我们的假设，地震被视作随机转移的质点，且其转移与两次以前发生的地震无关，所以我们可以把地震视作一条二阶 Markov 链。

3.2 模型的构造思路

根据上文，我们利用 K-means 聚类技术将中国划分成四个地震区域（以下记为 0 区，1 区，2 区，3 区），且每一次地震仅属于一个地震区。由于我们假定

了地震的转移是二重相依的，为建立模型，我们需要计算出地震转移的一阶、二阶概率。本文中，我们将用转移频率近似地替代转移概率，例如：若在i区发生了17次地震，其中有9次地震转移到了j区，则我们令$p_{ij}=9/17$，同理我们可计算二阶转移概率，例如，若有11次地震时从i区转移到j区，其中有4次从j区转移到了k区，则我们令二阶转移概率$p_{ijk}=4/11$。在此之后，我们将利用一阶、二阶转移概率的线性组合构建预报测度，其系数将由我们的评分系统给出。

为了令我们的模型更具有说服力，我将数据集划分成训练集（2009年至2020年7月23号全国5级以上地震数据，共计350条）和测试集（2020年7月26号至今的全国5级以上地震数据，共计11条），通过观察模型在测试集上的准确率，考察模型的效果。

3.3 一阶转移概率

为了计算一阶转移概率，我们先在训练集上统计地震从状态i转移到状态j的频率，结果见下表：

i \ j	0	1	2	3
0	4	4	5	3
1	7	45	33	28
2	2	40	51	28
3	3	24	32	40

表 1 一阶转移频率

我们若想知道从状态1转移到状态2的概率，我们需要先查阅表1，然后计算由状态1转移到状态0，1，2，3的总次数，即7+45+33+28=113，计算得到$P_{12}=33/133$。

3.4 二阶转移概率

与求一阶转移概率类似，为了得到二阶转移概率，我们首先需要统计由状态i转移到状态j再转移到状态k的事件频数，结果见下表。

k / i, j	0	1	2	3
0, 0	2	0	1	1
0, 1	1	0	0	3
0, 2	1	2	2	0
0, 3	0	0	0	3
1, 0	1	4	1	1
1, 1	2	22	11	10
1, 2	1	12	12	8
1, 3	0	6	9	13
2, 0	1	0	1	0
2, 1	3	14	14	9
2, 2	0	16	25	10
2, 3	1	8	10	9
3, 0	0	0	2	1
3, 1	1	9	8	6
3, 2	0	10	12	10
3, 3	2	10	13	14

表 2 二阶转移频率

我们若想知道地震由状态 1 转移到状态 3 再转移到状态 2 的概率，我们首先计算状态 1 转移到状态 3 的总次数 0+6+9+13 = 28，然后计算得到$P_{132} = 9/28$。

3.5 预报测度

有了上文中得到的一阶转移概率和二阶转移概率，接下来我们将要开始构建预报测度。根据我们的核心假设：地震的转移是二重相依的，我们可以构建预报测度M。$M(k)$表示下一次地震发生在区域k的可信性，$M(k)$的大小仅依赖于前两次地震。基于这种朴素的想法，我们构建的预报测度计算公式如下：

$$M(k) = AP_{jk} + BP_{ijk}.$$

其中i，j表示前两次地震所在区域。

下面，我们将采取一种计分策略来制定模型中参数A，B的值。

对于一阶转移概率前面的系数A，这个参数体现了一阶转移概率在本模型中的权重，换句话说，也就是体现了我们利用一阶转移概率预测下次地震的信心。我们将通过以下这个计分系统来计算一阶转移概率的权重A。例如，现在有一个地震发生在 2 区，我们若想根据一阶转移概率来预测下一次地震发生在什么区域，我们根据以往的经验（表 1），发现以往有 2 次转移到了 0 区，有 40 次转移到了 1 区，有 51 次转移到了 2 区，有 28 次转移到了 3 区。那么我们认为地震由 2 区转移到 2 区的概率最大，做出预测下一次地震将发生在 2 区。若我们的预测结果正确，那么我们记 1 分，若发生在第二可能区域（在本例中 1 区是第二可

能区域）记 0.5 分。遍历整个测试集，我们将每一次得分累加后得到一个总分，我们令这个总分作为权重A。

根据同样的想法，我们依旧可以通过计分的手段得到二阶转移概率前面的系数B。例如，当前地震从 1 区转移到 3 区，我们要预测下一次地震发生在哪个区域，我们观察既往发生的频次（见表 2），发现有 0 次转移到 0 区，6 次转移到 1 区，9 次转移到 2 区，13 次转移到 3 区，那么下次地震发生的第一可能区域为 3 区，若地震确实发生在这个区域那么我们给B加上 1 分，同样的，若地震发生在第二可能区域即 2 区，那么我们给B加上 0.5 分。遍历测试集，我们得到的总分即为B的值。

详见 https://github.com/Albertlziwen/dissertation/blob/main/coefficient.py 中的 calA(), calB()函数。

计算得到$A = 197.5$，$B = 223.5$。

综上，我们的预报测度为：

$$M(k) = 197.5P_{j,k} + 223.5P_{i,j,k}.$$

4. 模型检验

有了预报测度这个工具，我们便可以对地震进行预测。为了验证我们模型的准确性，我们将在测试集上进行模型验证。

我们的验证思路如下：

为了预测本次地震发生的区域，我们先通过查表的方式，得到前两次地震发生的地区，以及一阶转移概率，二阶转移概率。通过预报测度公式计算得到一列预报测度：$[M(0), M(1), M(2), M(3)]$，计算其测度和$S = M(0) + M(1) + M(2) + M(3)$，那么发生在四个区域的概率（其实不是概率而是一种可能性，为了方便理解简写作概率）为：

$$\left[\frac{M(0)}{S}, \frac{M(1)}{S}, \frac{M(2)}{S}, \frac{M(3)}{S}\right].$$

我们将发生概率最大的区域称为第一可能区域，并用 1 标记；将发生概率第二大的区域称为第二可能区域，并用 0.5 标记；若地震发生在其他区域则视为一次错报，用 0 标记。

细节请见 https://github.com/Albertlziwen/dissertation/blob/main/mark.py。

预报结果如下表所示：

序号	发震时刻(UTC+8)	经度(°)	纬度(°)	参考位置	mark
1	2021/3/24 5:14	81.11	41.7	新疆阿克苏地区拜城县	0
2	2021/3/19 14:11	92.74	31.94	西藏那曲市比如县	0.5
3	2021/3/2 17:23	121.17	21.92	台湾屏东县海域	1
4	2021/2/9 0:58	122.12	24.32	台湾宜兰县海域	1
5	2021/2/7 1:36	122.5	24.65	台湾宜兰县海域	1
6	2021/1/17 7:10	121.44	22.44	台湾台东县海域	1
7	2021/1/9 19:35	122.21	24.7	台湾宜兰县海域	1
8	2020/12/10 21:19	121.99	24.74	台湾宜兰县海域	1
9	2020/9/30 12:37	122.14	24.85	台湾宜兰县海域	1
10	2020/9/29 4:50	121.1	22.29	台湾台东县海域	1
11	2020/7/26 20:52	122.48	24.27	台湾花莲县海域	0.5

表 3 测试结果

由表 3 知，11 次最新发生的地震中，有 8 次地震发生在第一可能区域，2 次地震发生在第二可能区域，有 1 次错报的情况。模型表现较好。

5. 模型评价

虽然我们的模型在测试集上表现较好（仅有一次错报，且地震发生在第一可能区域概率较高），但是值得注意的是，由于我们测试集数据较少，所以结果是有一定偶然性的。下面我将分析以下本模型可能存在的问题：

其一，采用聚类方法划分地震带，可能会因为初始点的选择和迭代次数的不同出现不同的结果。在本篇文章我们划分的地震带只是统计意义上的地震带，而非物理层面上的地震带。

其二，我们构建的预报测度$M(k) = AP_{j,k} + BP_{i,j,k}$前面的系数$A$，$B$是固定的，与状态$i$，$j$无关，若将此系数与状态结合起来，可能会进一步提高预测精度。这也是我们进行模型改进的重点所在。

其三，系数A，B的选取，我们采取的是一种打分制度，我们设计的打分规则是否合理也有待商榷，我们在本文的剩余部分也会对打分规则进行灵敏度检验。

其四，我们的模型目前只能预测下一次发生地震的大致区域，而无法预测下一次地震发生的时间，在未来的改进中我们或许可以在模型中加入时间项（而不仅仅只考虑地震发生的次序）来给下次地震发生的时间一个大致的推测。

其五，测试集中的 11 次地震有 9 次都发生在同一个地震区，这提示我们或许可以更加细致的考虑地震区的区内转移，即将地震大区再细化成地震小区。这种划分有助于帮我们区分大震后的余震，提升模型精度。

其六，若在两个地震区域上发生的地震次数相同，我们的模型很难判断出谁是第一可能区域，谁是第二可能区域。例如，在表 2 中，$i = 1$，$j = 2$的情况，转移到四个区的次数分别是 1，12，12，8。我们的模型就会把 1 区认定为第一

可能区域，把 2 区认定为第二可能区域，但实际情况也有可能是相反的。（原因请参考 https://github.com/Albertlziwen/dissertation/blob/main/coefficient.py 中 max2()函数的写法）

6. 模型改进

6.1 评分模型的改进

首先我们要指出的是，我们根据一阶、二阶转移概率构建的预报测度一定是线性的，即形如$M(k) = AP_{j,k} + BP_{i,j,k}$，而不可能出现诸如$P_{j,k} * P_{i,j,k}$或者$P_{i,j,k}^2$的形式，其原因由以下定理给出：

定理一：设M_1，M_2，…，M_n是测度空间(R, B)上的有限测度（B是$Borel$集），$f(x_1, x_2, \dots, x_n)$是R^n上的连续处处可导实函数，则对于$A \in B$，$M(A) = f\big(M_1(A), \dots, M_n(\mathrm{A})\big)$是$(R, B)$上的测度的充分必要条件是$f$是线性的且系数非负。

证明：

先证必要性。即我们要证明M是测度，则函数f的形式一定是线性的。这也是本文中要用到的部分。我们先证$n = 1$的情况。

$\forall A, B \in B$且$A \subseteq B$，由于M是测度，所以我们有：

$$M(B \backslash A) = M(B) - M(A) = \mathrm{f}\big(\mathrm{M}_1(\mathrm{A})\big) - \mathrm{f}\big(\mathrm{M}_1(\mathrm{B})\big). \qquad (1)$$

另一方面，由于M_1也是测度，所以我们有：

$$f\big(M_1(A - B)\big) = f\big(M_1(A) - M_1(B)\big). \qquad (2)$$

由于（1）=（2），所以我们得到：

$$f\big(M_1(A) - M_1(B)\big) = f\big(M_1(A)\big) - f\big(M_1(B)\big).$$

为了证明f是线性函数，我们先证明下面这个引理：

引理一：若f是定义在 R 上的连续可微函数，若$\forall a, b \in R,\ f(a - b) = f(a) - f(b)$，那么一定$\exists t \in R,\ s.t.\ f(x) = tx$。

证明：

令$a = b + \frac{1}{k}$，那么

$$f(a - b) = f\left(\frac{1}{k}\right) = f(a) - f(b) = f\left(b + \frac{1}{k}\right) - f(b).$$

同时除以$\frac{1}{k}$并取极限，得到：

$$\lim_{\mathrm{k}\to\infty} \frac{f\left(\frac{1}{k}\right)}{\frac{1}{k}} = \lim_{\mathrm{k}\to\infty} \frac{f\left(b + \frac{1}{k}\right) - f(b)}{\frac{1}{k}}.$$

由于f是处处可微的，所以我们得到：

$$f'(0) = f'(b), \qquad \forall b \in R.$$

且由于：$f(0-0) = f(0) - f(0) = 0$.

所以f形如$f(x) = tx$，证毕。

由引理一，我们证明了当f是一元函数的时候，f为线性的。下面我们考虑高维的情况。

实际上，若f形如$f(x_1, x_2, \dots, x_n)$，要说明f是线性的，我们只需要证明对$\forall 1 \leq i \leq n$，我们有：

$$\frac{f(0,\dots,a_i,\dots,0) - f(0,\dots,b_i,\dots,0)}{a_i - b_i} = k_i,$$

其中k_i是和i有关的常量。此时再利用我们已经证明的在f是一元函数时的结论，我们就证明了必要性。

下面我们证明充分性。若：

$$M(A) = k_1 M_1(A) + k_2 M_2(A) + \cdots + k_n M_n(A),$$

且$k_i \geq 0, \forall 1 \leq i \leq n$。

显然，$M(\emptyset) = k_1 M_1(\emptyset) + \cdots + k_n M_n(\emptyset) = 0$.

且对于$\{A_m, m \in Z\}, A_m$两两不交，有：

$$M(\cup_m A_m) = k_1 M_1(\cup_m A_m) + \cdots + k_n M_n(\cup_m A_m),$$

由于$M_1, \dots, M_n$是测度，所以由测度的σ可加性，我们有：

$$M(\cup_m A_m) = k_1 \sum_m M_1(A_m) + \cdots + k_n \sum_m M_n(A_m),$$

由于测度有限，所以求和可交换次序，我们得到：

$$M(\cup_m A_m) = \sum_{i=1}^{n} \sum_m k_i M_k(A_m) = \sum_m \sum_{i=1}^{n} k_i M_k(A_m) = \sum_m M(A_m).$$

充分性证毕。

在本文中，由于P_{jk}和P_{ijk}都是概率测度（概率测度都是有限测度），所以由定理一，M若想为测度则一定是线性的。

那么我们对模型改进只能在系数A, B上做文章。我们考虑将A，B与地震状态i，j结合起来。此时我们的预报测度为：

$$M(k) = A_j P_{jk} + B_{ij} P_{ijk}.$$

我们同样使用计分策略，只不过我们的A_j的算法是若当前地震在j区而下次地震发生在第一可能区域则记 1 分，发生在第二可能区域则记 0.5 分。B_{ij}同理，若前一次地震发生在i区，当前地震发生在j区而下一次地震发生在第一可能区域记 1 分，发生在第二可能区域记 0.5 分。

见 https://github.com/Albertlziwen/dissertation/blob/main/coefficient.py 中的 calA_j()及 calB_ij()函数。

新的系数矩阵如下：

j	0	1	2	3
Aj	5.5	44.5	53.5	56

表 4 与状态有关的系数 A 的值

B_{ij}	0	1	2	3
0	1.5	3	1	3
1	1.5	15.5	14	17.5
2	0.5	16	22.5	14
3	2	10	16	20.5

表 5 与状态有关的系数 B 的值

新的预测结果如下：

序号	发震时刻(UTC+8)	经度(°)	纬度(°)	参考位置	mark
1	2021/3/24 5:14	81.11	41.7	新疆阿克苏地区拜城县	0
2	2021/3/19 14:11	92.74	31.94	西藏那曲市比如县	0.5
3	2021/3/2 17:23	121.17	21.92	台湾屏东县海域	1
4	2021/2/9 0:58	122.12	24.32	台湾宜兰县海域	1
5	2021/2/7 1:36	122.5	24.65	台湾宜兰县海域	1
6	2021/1/17 7:10	121.44	22.44	台湾台东县海域	1
7	2021/1/9 19:35	122.21	24.7	台湾宜兰县海域	1
8	2020/12/10 21:19	121.99	24.74	台湾宜兰县海域	1
9	2020/9/30 12:37	122.14	24.85	台湾宜兰县海域	1
10	2020/9/29 4:50	121.1	22.29	台湾台东县海域	1
11	2020/7/26 20:52	122.48	24.27	台湾花莲县海域	0.5

表 6 改进后的结果

可以看出预测结果与之前没有考虑状态时相同，所以将系数A，B与当前状态结合起来并没有提高我们模型的预测能力。

6.2 针对评分系统的灵敏度分析

在上文我们的评分系统中，我们假定预测值落在第一可能区间记 1 分，落在第二可能区间记 0.5 分。但实际上，我们并不知道这个打分制度是否会影响我们模型的预测结果，下面我们来探究一下。探究思路如下：

由于打分的绝对值对我们模型没有影响，有影响的是打分的相对值（比如落在第一可能区间加 10 分，落在第二可能区间加 5 分这种打分策略和我们正在使用的策略效果相同）。所以，我们将额外设置如下两种打分策略：a.落在第一可能区间加 1 分，其余不加分。b.落在第一、第二可能区间均加 1 分，其余不加分。（细节见 https://github.com/Albertlziwen/dissertation/blob/main/coefficient.py 中的 calA_a(), calA_b(), calB_a(), calB_b()函数。）

采用两种策略后的计算结果如下：

策略 a: 此时 A=141，B=177，预测结果无变化。

策略 b: 此时 A=254，B=270，预测结果无变化。

可以看出，无论我们采取何种计分策略，预测结果均无变化。所以我们得出结论，我们的模型对评分系统不敏感。

7. 展望：机器学习在地震预测中的应用前景

关于地震是否可以预测，直到现在依旧没有定论，但人们对于地震预测这门技术的探索依旧持续着。尤其是近几年，深度学习技术掀起了运用机器学习技术来预测地震的热潮。由于在处理大量的复杂数据以及不确定性上具有天然的优势，人们对于机器学习解决地震预测这一世界性难题寄予厚望。

下面我将简单列举几个近年来机器学习技术在地震预测领域产生的成果。

Alves 等[6]将地震活动与经济活动相类比，用人工神经网络模型（ANN）来进行地震预测。Alves 将地震发生的时间、地点、震级输入到人工神经网络中，该模型输出的为预测的时间、地点、震级，实现了输入和输出的一致。此模型在合理的误差范围内，成功地预测了葡萄牙亚速尔群岛于 1998 年 7 月和 2004 年 1 月发生的两次地震。

Panakka 等[7]将地震预测视作一个分类问题。他用了多种不同的模型对美国发生的地震进行了建模。综合所有模型，Panakka 发现递归神经网络（RNN）结果较好。之后，他又自创了概率神经网络模型（PNN）在南加州地震数据集上进行测试，结果表明，该模型对 4.5 级到 6 级的地震有着较高的预测精度。

Cortes 等[8]等研究了人工神经网络模型（ANN）在预测东京及其周边地区的预测准确度（时间范围设置在了 7 天，总计输入了 16 个参数）。研究发现，相比于其他模型，ANN 更具优越性。更进一步地，他又应用四种回归器及组合来进行震级预测，研究结果表明，回归器越复杂，预测结果越准确。

Reyes 等[9]利用智利四个地震区的数据，将 7 个地震指标输入前馈反向 ANN，当预测的地震发生概率高于给定的阈值时，预测为有地震。他利用灵敏度、精确度和特异性指标对这四个震区进行比较分析，结果发现 ANN 的准确率和所处区域有关。

马士振等[10]利用地脉数据，用回归与分类树（CART），梯度提升决策树（GBDT）和支持向量机(SVM)三种算法来预测地震。用 k 折交叉验证方法进行评估，发现 SVM 效果最佳且地震活动与地脉变化有关。

当前，地震预测届从以往的统计模型向着机器学习模型的转变，或许为未来指出了一条明路。机器学习模型可能通过深入的数据挖掘，找出隐藏在数据背后更深刻更基础的地震发生的原理，这或许能帮助我们从根本上解决地震预测这个

“世界性难题”。

8. 参考文献

[1] 赵晓燕. *地震概论*. 北京: 清华大学出版社, 2013. Print. 防灾减灾系列教材 Fang Zai Jian Zai Xi Lie Jiao Cai.

[2] 王梓坤,朱成熹,李彰南,王启鸣,孙惠文,徐道一.地震迁移的统计预报[J].地质科学,1973(04):294-306.

[3] "中国地震烈度区划图（1990）及其说明." 中国地震 8.4 (1992): 1-11. Web.

[4] Alpaydin, Ethem. *Introduction to Machine Learning*. 2nd ed. Cambridge, Mass.: MIT, 2010. Print. Adaptive Computation and Machine Learning.

[5] Silhouettes: A graphical aid to the interpretation and validation of cluster analysis

[6] Alves, E Ivo. "Earthquake Forecasting Using Neural Networks: Results and Future Work." *Nonlinear Dynamics* 44.1 (2006): 341-49. Web.

[7] Panakkat, Ashif, and Adeli, Hojjat. "NEURAL NETWORK MODELS FOR EARTHQUAKE MAGNITUDE PREDICTION USING MULTIPLE SEISMICITY INDICATORS." *International Journal of Neural Systems* 17.1 (2007): 13-33. Web.

[8] Asencio‐Cortés, G, Scitovski, S, Scitovski, R, and Martínez‐Álvarez, F. "Temporal Analysis of Croatian Seismogenic Zones to Improve Earthquake Magnitude Prediction." *Earth Science Informatics* 10.3 (2017): 303-20. Web.

[9] Reyes, J, Morales-Esteban, A, and Martínez-Álvarez, F. "Neural Networks to Predict Earthquakes in Chile." *Applied Soft Computing* 13.2 (2013): 1314-328. Web.

[10] 马士振,刘宏志,牟磊育.在地脉动数据上应用分类算法的地震预测实验[J].地震,2020,40(01):159-171.

9. 附录：

附录一：本文用到的程序

为了方便读者利用或者学习本文的模型我将本文用到的所有代码以及数据放到了 GitHub 上，详情请见 https://github.com/Albertlziwen/dissertation。欢迎交流！

附录二：全国五级以上地震分布散点图

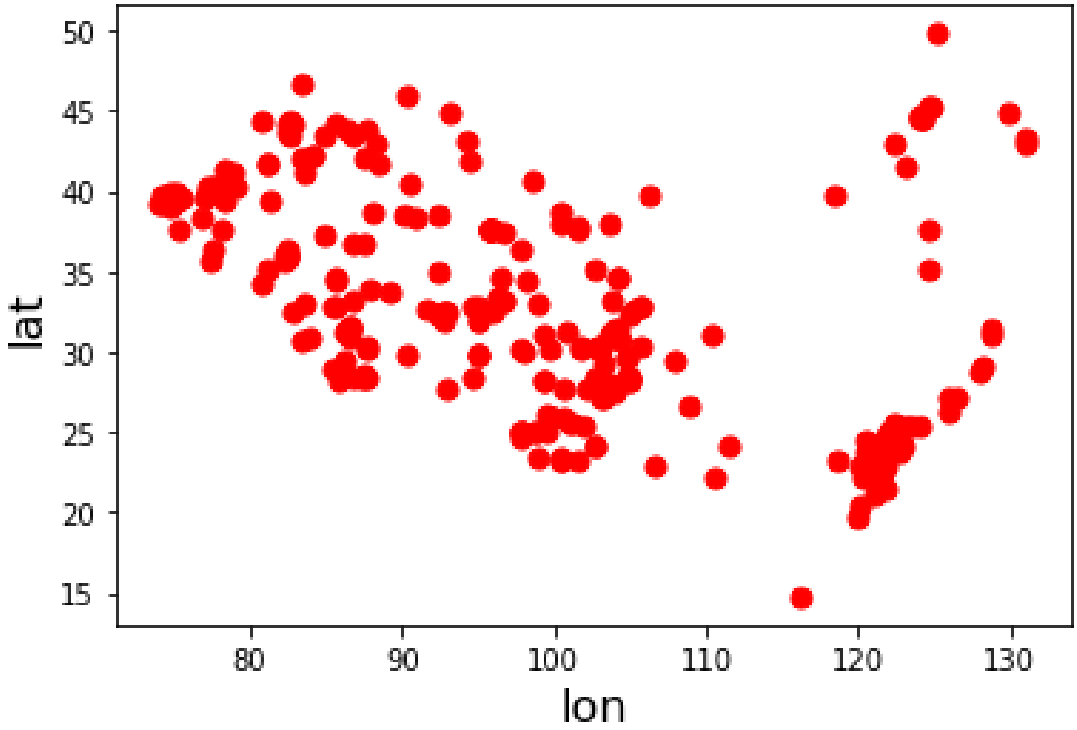

图 1 全国 5 级以上地震分布散点图

附录三：手肘法选取 K 值

手肘法是一种常用的选取聚类数量的方法，此种方法考虑误差平方和（SSE）与聚类数量的关系。SSE 的计算方法如下：

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2.$$

其中，C_i是第i个聚类，m_i是该聚类的中心，SSE 表示的是数据集中所有点到他们所属聚类中心的误差。这么说有点让读者“意会”的感觉，我在此举一个简单的例子，如果一个数据集只有 0，1 两个数据，我们若自然的把他们分成两类，中心自然也是 0，1 两点，此时 SSE 为 0，若我们一定要强行把他们归到一类，则聚类中心为 0.5，此时 SSE=0.5。

显然地，K 值越大，意味着对数据集的划分越细致，所以 SSE 的值必然会随着 K 值的增大而减小。而当 K 值增大到某一值时，SSE 的值随着 K 值增加而减小的幅度会减小，这意味着增大 K 值的收益减小。手肘法要找的 K 值就是 SSE 曲线斜率变化最大的拐点。

将手肘法应用到我们的地震数据集，SSE 曲线如图：

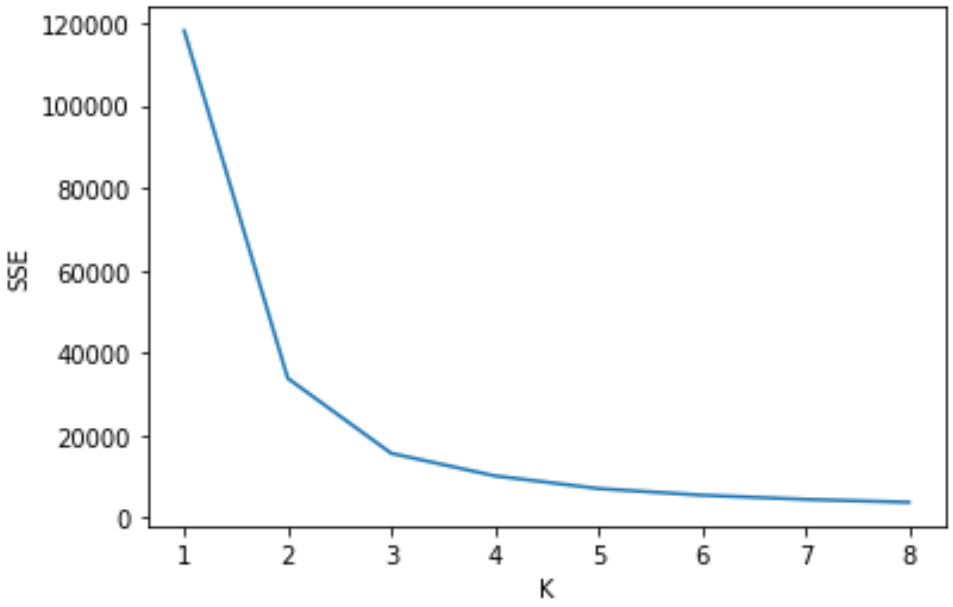

图 2 SSE 曲线

由图 2 知，用手肘法选取的 K 值为 3。

10. 致谢

时间飞逝，转眼就要毕业。首先，我要感谢我的所有专业课老师，数学科学学院的老师们真正践行了“学为人师，行为世范”的校训，不仅教会了我知识，还为我树立了榜样。我要特别感谢何辉老师，王颖喆老师和程志云老师，三位老师在我申请学校的时候给予了我很大的助力。

我还要感谢我的朋友们，你们让我的大学生活多姿多彩，希望大家都拥有光明的未来。

我也要感谢我的父亲李群英先生，母亲刘天明女士，感谢你们相信我的选择，无条件的支持我。

此外，我还要感谢王梓坤先生，虽然我和王先生素未蒙面，但我大一的时候在图书馆五楼拜读了《王梓坤文集》，本文的灵感就来源于那本书。

当然，我还要感谢每一位认真阅读本文的读者。本人才疏学浅，表述能力较差，为了让文章尽可能地通俗易懂，在本文的许多章节中配备了示例帮助读者理解。若读者在本文中发现谬误，还请及时与我联系。邮箱：1244497667@qq.com