“华为杯”第十五届中国研究生

数学建模竞赛

学校	复旦大学
参赛号	**18102460009**
队员姓名	**1.李渝方**
	2.崔丹妮
	3.邓兴旺

“华为杯”第十五届中国研究生

数学建模竞赛

题 目　　恐怖袭击事件记录数据的量化分析

摘　　要：

本文主要根据全球恐怖主义数据库（GTD）中 1998-2017 年世界上发生的恐怖袭击事件的记录对恐怖袭击事件进行分析。考虑到恐怖袭击事件发生的主要原因、时空特性、蔓延特性、级别分布等规律对恐怖袭击事件的影响，为此我们建立模型来分析影响恐怖袭击事件的各种因素。在这些模型中我们利用 excel，R，Python 等一系列数学建模工具对模型进行求解，综合统计分析各种方法，从而得出最终的结论。

首先，对于任务 1，依据危害性对恐怖袭击事件分级。我们选取了危害性相关特征，在对数值变量进行归一化处理后，建立线性模型，人为定义权重，通过矩阵乘法得到此事件危害程度的分数。最后对危害分数设定阈值，将其划分为 1-5 级。找到近二十年来危害程度最高的十大恐怖袭击事件。

对于任务 2，我们采用 kmondes 聚类方法对 2015、2016 年度发生的、尚未有组织或个人宣称负责的恐怖袭击事件进行聚类。我们对给出的特征进行筛选，最终筛选出 5 个特征——地区、攻击类型、受害者类型、国籍、武器类型为最终聚类分析的特征。我们结合问题 1 的量化分级模型，将每个事件与量化等级评分相关联，计算每一个类中所有事件的量化等级评分的平均值作为该类的等级评分，我们将等级评分按降序排列取前五类。由于题目所给数据为离散数据，我们通过计算事件与每一类的 modes（类似于 kmeans 中的聚类中心点）之间汉明距离（hamming distance）以判断事件与类的相关性。汉明距离越短，说明事件属于该类的概率越大。而在本题中，汉明距离越短，事件属于该组织所为的概率就越大。最后，我们将汉明距离转化为恐怖分子关于典型事件的嫌疑度评分，得到最终结果。

对于任务 3，我们按不同的特征变量对案件群体进行划分，并计算频数和频率，然后通过图表的方式对主要原因、时空特性、蔓延特性、级别分布等规律进行可视化分析，主要用到的图表类型包括饼图、折线图、柱状图等。然后通过近三年的信息对未来反恐态势进行分析。

对于任务 4，我们重点分析了一下金砖四国往年发生的恐怖袭击数量以及伤亡人数。

最后，我们对模型优缺点进行评价，并提出了改进建议，以用于实际的推广和应用。

关键字

矩阵乘法，线性模型，Kmodes 聚类分析，汉明距离，R 软件，Python

1. 问题重述

恐怖主义是人类的共同威胁，打击恐怖主义是每个国家应该承担的责任。对恐怖袭击事件相关数据的深入分析有助于加深人们对恐怖主义的认识，为反恐防恐提供有价值的信息支持。附件 1 选取了某组织搜集整理的全球恐怖主义数据库（GTD）中 1998-2017 年世界上发生的恐怖袭击事件的记录，需要完成的任务如下：

任务 1 依据危害性对恐怖袭击事件分级

任务 2 依据事件特征发现恐怖袭击事件制造者

任务 3 对未来反恐态势的分析

任务 4 数据的进一步利用

2. 模型假设

1. 假设持续时间，入选标准，袭击是否成功，伤亡人数，财产损失，人质情况，以及是否为连续事件等因素可以反映恐怖事件的危害程度，即可作为恐怖事件分级的判定依据。

2. 假设可以通过地区、攻击类型、受害者类型、国籍、武器类型特征判断恐怖袭击事件制造者类型。

3. 名词说明

名词	对照
汉明距离	Hamming distance
全球恐怖主义指数	Global Terrorism Index
危害等级	riskscore
类平均危害等级	mean_riskscore

4. 模型建立与求解

4.1 问题一模型建立与求解

4.1.1 问题分析

任务一的问题是依据危害性对恐怖袭击事件分级，由于恐怖袭击事件的复杂性，除了人员伤亡和经济损失外，还需要其他指标共同判定事件的级别。在查阅了恐怖袭击衡量的相关文献和 GTD 数据库给出的变量的相关说明后，我们人为筛选了一些特征，如持续时间，袭击是否成功，伤亡人数，人质情况，以及是否为连续事件等因素。

由于数据库中并没有提供先验的分级信息，即没有训练样本，故我们不能采取有监督学习的分类算法。受到 GTI（Global Terrorism Index）[1]的启发，GTI 使用了 GTD 的四个变量，建立一个线性模型，通过人为定义权重，计算出了 GTI，即各国家的恐怖活动严重程度分数，我们决定使用类似的线性模型进行分级预测。

4.1.2 模型建立与求解

4.1.2.1 数据预处理

GTD 提供的变量包括数值变量，text 变量，分类变量 3 种类型。一些变量含有大量的 unknown 信息，对模型的计算和评估效果不利，建议弃用。文本信息的内容不全且过于冗杂，因此我们选取了数值变量和分类变量作为模型的输入向量。通过对数值变量的分布进行分析，我们发现一些变量如死亡人数，人质数等，区间较大，不利于代入模型与分类变量一起计算，因此我们对于所有的数值变量进行了归一化处理，归一化公式如下：

$$Xi^{*} = \frac{Xi - Xmin}{Xmax - Xmin}$$

其中Xi^{*}和 Xi 分别代表正则化后数值和原始数值，Xmax 和 Xmin 分别代表第 i 类数值变量的最大值以及最小值，经过归一化处理后，数值变量的取值将在 0，1 之间，便于后续计算。

4.1.2.2 模型建立

模型主要运用到了矩阵乘法，它的公式表示如下：

$$\text{score} = \sum_{i=1}^{13} \text{向量取值(i)} * \textit{对应}\text{权重(i)}.$$

即对于 GTD 事件，给其每一个向量赋予一个权重，再求和得到此事件危害程度的分数。

公式包括的特征和对应的权重如下表所示:

表 4.1.1 特征和权重对应表

特征	权重	变量类型	说明
extended	3	0，1 分类变量	是否为持续事件
crit1	1	0，1 分类变量	入选标准 1
crit2	1	0，1 分类变量	入选标准 2
crit3	2	0，1 分类变量	入选标准 3
doubtterr	-1	0，1 分类变量	疑似恐怖主义

multiple	3	0，1 分类变量	事件组的一部分
success	5	0，1 分类变量	成功的攻击
nkill	20	数值变量 标准化 [0,1]	死亡总数
nwound	2	数值变量 标准化 [0,1]	受伤总数
property	1	0，1 分类变量	有无财产损失
propextent	20	分类变量 标准化 [0,1]	损失程度 1 级=1，2 级=0.2
nhostkid	2	数值变量 标准化 [0,1]	人质/绑架受害者总数
ndays	1	数值变量 标准化 [0,1]	人质/绑架事件的天数

4.1.3 实验结果与分析

4.1.3.1 实验结果

线性模型的危害打分结果的分布如下表所示：

表 4.1.2 线性模型的危害打分结果分布

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	9	9.513	9.138	10.03	52.66

对危害分数设定阈值，将其划分为 1-5 级，划分的依据如下所示：

表 4.1.3 危害分数设定阈值表

级别	对应分数	案件数目
1 级	>20	15
2 级	15-20	1004
3 级	12-15	19079
4 级	10-12	29129
5 级	0-10	64956

近二十年来危害程度最高的十大恐怖袭击事件：

表 4.1.4 十大恐怖袭击事件

事件编号	危害级别
200109110004	1
200109110005	1
200109110006	1
200109110007	1
201408090071	1
201406150063	1
201406100042	1
199811010001	1
199808070002	1
201710140002	1
200708150005	1

表 4.1.5 典型事件危害级别

事件编号	危害级别
200108110012	2
200511180002	2
200901170021	2
201402110015	5
201405010071	4
201411070002	3
201412160041	4
201508010015	5
201705080012	3

4.1.3.2 结果分析

通过查阅互联网信息，表明我们的分级结果是符合现实的，我们判定为十大恐怖袭击事件的前四名均发生在 2001 年 9 月 11 日，也就是举世哗然的 911 事件，其他事件也造成了大量人员伤亡和财产损失，在此不一一赘述。

4.2 问题二模型建立与求解

4.2.1 问题分析

任务二的目的是为了找出新生的或者隐匿的恐怖组织或个人，这些数据是无标签的数据，所用到的算法应为无监督的算法。因为多起恐怖袭击事件未确定作案者，所以题目要求根据已有的信息找出同一个恐怖组织或个人在不同时间、不同地点多次作案的若干案件，这属于聚类问题。同时，题目要求按该组织或个人的危害性从大到小选出其中的前 5 个，这属于量化分级问题。由于第一问中已经求解出量化分级模型，所以问题 2 中只需要将组织或个人与量化分级模型一一对应，就能求该组织或个人的危害性。为了确定每个事件与这 5 个组织之间的关系，我们通过计算每个事件与 5 个聚类中心点之间的距离，来判定事件与组织或个人之间的相关性。

4.2.2 模型建立与求解

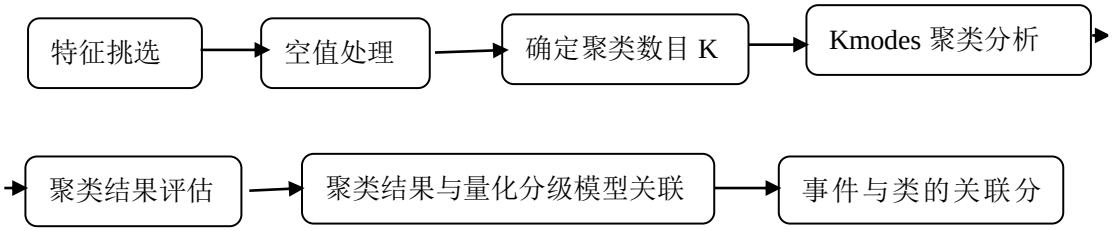

图 4.2.1 模型建立流程图

为了确定同一个恐怖组织或个人在不同时间、不同地点多次作案的若干案件，我们对附件 1 给出的数据进行预处理，筛选出 2015、2016 年度发生的、尚未有

组织或个人宣称负责的恐怖袭击事件，一共 22740 条记录。同时，我们对给出的特征进行筛选，最终筛选出 5 个特征——地区、攻击类型、受害者类型、国籍、武器类型为最终聚类分析的特征。为了保持数据的完整性，我们补全了数据缺失值。由于所选特征均为分类型变量，所以对于每个特征中的缺失值，我们分别自定义一个数值代表该特征中的空值（表 4.2.1），以得到最终用于聚类的数据（附件表 4.2.1）。我们通过 R 语言中的“mclust”包确定最终聚类的数目为 15 类，利用“klaR”包中的 kmodes 算法对 5 特征进行聚类，该算法适用于离散型数据。然后，将聚类结果与问题一中的风险模型进行关联，计算出每个类中所有事件的危害等级（riskscore）的平均值代表该类的危害等级（mean_riskscore），取前五类。最后，我们通过汉明距离(Hamming distance)计算每个事件与前五类的距离，以此来判断事件与类的相关性。

表 4.2.1 每个特征中 NA 的自定义取值

特征	region	attacktype1	targtype1	natlty1	weaptype1
NA 填充值	无缺失值	无缺失值	无缺失值	888	无缺失值

4.2.2.1　聚类分析模型

1.恐怖袭击事件聚类

由于附件 1 所给的数据为离散型数据，我们通过 R 语言中 mclust 包计算最佳的聚类数 K[2]。对于离散型数据的聚类我们采用 Kmodes 聚类方法。Kmodes 是 K-means 的扩展，该方法是一种无监督的聚类方法，需要事先确定聚类的数目 K。

Kmodes 算法适用于离散型数据，我们假设有 N 个样本，M 个属性且全是离散的，簇的个数为 k，其实现步骤如下：

步骤一：随机确定 k 个聚类中心 $C_1, C_2 \ldots C_k$，C_i 是长度为 M 的向量，$C_i=[C_i^1, C_i^2, \ldots, C_i^M]$

步骤二：对于样本 $x_j(j=1,2,\ldots,N)$，分别比较其与 k 个中心之间的汉明距离

步骤三：将 x_j 划分到距离最小的簇，在全部的样本都被划分完毕之后，重新确定簇中心，向量 C_i 中的每一个分量都更新为簇 i 中的众数

步骤四：重复步骤二和三，直到总距离（各个簇中样本与各自簇中心距离之和）不再降低，返回最后的聚类结果[3]。

2.恐怖袭击事件与类的相关性判断

通过 kmodes 算法对所有未认领的事件进行聚类，将这些事件聚成 15 类。我们将事件与问题一种的危害等级相关联，以每类的平均等级作为该类的危害等级。按照危害等级降序进行排序，取前五。在 kmodes 聚类结果中我们取出前五类的 modes 值，分别计算每一个事件与每一个类的汉明距离。在信息论中，两个等长字符串之间的汉明距离是两个字符串对应位置的不同字符的个数，即两个字符串中不同的字符个数：1011101 与 1001001 之间的汉明距离是 2[4]。汉明距离越短，说明该事件属于该类的概率越大。

4.2.3 实验结果与分析

4.2.3.1 聚类数 K

通过 R 包中的 mclust 函数对聚类数 K 进行确定，我们自定义的聚类数目为 20 类。mclust 包中内置 14 种聚类模型，通过计算每一种模型的 1 到 20 类的 BIC 值来判断聚类数 K 的取值。通过 R 语言代码，计算得出最佳聚类数为 17 类，我们将结果进行可视化（图 4.2.2）。观察图片我们发现 14 中模型的 BIC 值随着聚类数 K 的增加儿缓慢上升，在聚类数 K 大于等于 15 时增加比较缓慢，故我们选取最佳聚类数为 15 类。

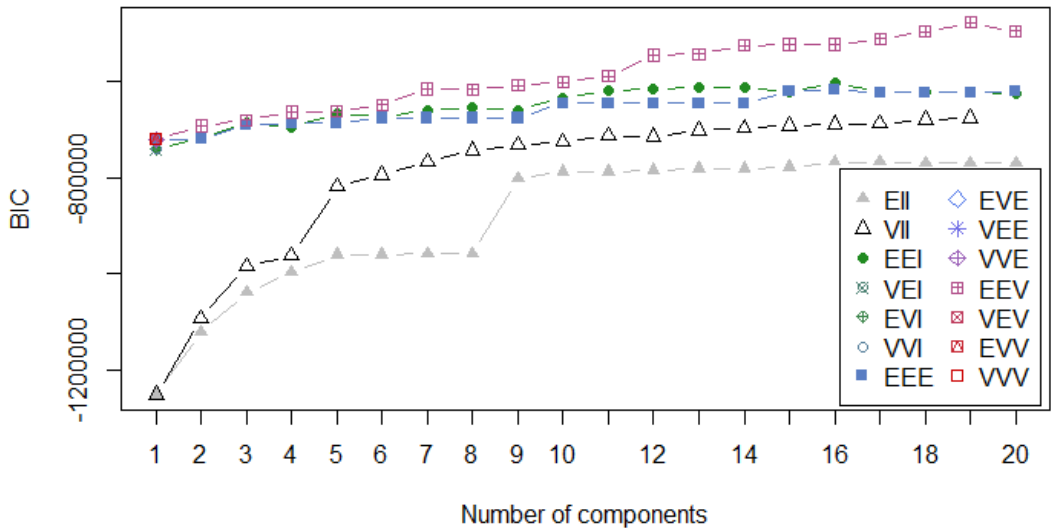

图 4.2.2 聚类数目与 BIC 之值的关系

4.2.3.2 聚类结果分析

1. 聚类结果可视化

通过 klaR 包中的 kmodes 函数对数据进行聚类以及选取前 5000 行数据进行可视化，如图 4.2.3 所示，聚出的 15 类在 5 个判定特征的维度的分布。

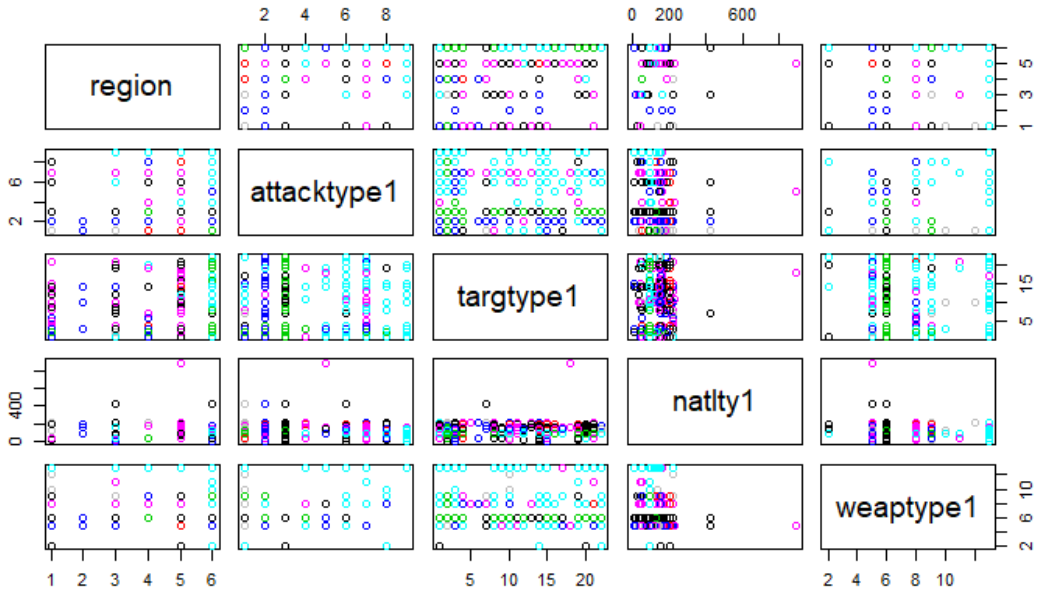

图 4.2.3 15 各类类在 5 个判定特征的维度的分布

2.筛选打分 top5 的类

在完成聚类之后，我们将聚类结果与问题 1 的量化分级模型进行关联，确定每一个类中的每一个样本的危害等级。通过计算每一个类中所有样本危害等级的平均值作为该类的危害等级，排名前五的类及其打分详见表 4.2.2。

表 4.2.2 量化分级分数排名前五的类

cluster	9	10	4	5	12
mean_score	11.04853	10.22524	10.2004	9.965624	9.388669

3. 筛选 top5 的聚类 modes

事件与类的相关性一般通过距离进行判断，在 Kmodes 算法中我们需要计算每一个样本到每一类 modes 的距离。Kmodes 是 Kmeans 算法的衍生，这里的 modes 相当于 Kmeans 中的聚类中心点。而每一类的 modes 在聚类结果中已经给出，我们将给出的每一类 modes 值与危害等级分数进行关联并按照降序排序（表 4.2.3），得到前五类的 modes。

表 4.2.3 15 类的聚类 modes 以及量化等级分数

cluster	region	attacktype1	targtype1	natlty1	weaptype1	meanscore
9	10	6	14	113	5	11.04853
10	5	2	14	205	5	10.22524
4	11	2	14	147	5	10.2004
5	6	9	14	92	13	9.965624
12	6	2	3	153	5	9.388669
7	11	3	14	147	6	9.20522
1	6	3	14	4	6	8.79665
13	10	2	3	95	5	8.570659
2	10	3	4	95	6	7.699888
6	5	2	4	160	5	7.644722

11	6	3	2	92	6	7.550265
15	11	6	14	195	5	7.391615
8	10	1	2	160	5	7.333351
14	8	7	14	233	8	7.175714
3	10	3	14	95	6	6.39988

4.计算事件与类的汉明距离

判断事件与类的相似性常用的方法是计算距离，由于附件 1 所给数据为离散型数据，计算离散型数据的距离可以用汉明距离。我们通过 R 语言代码实现每个事件与每个类的汉明距离计算，详细结果见附件（附件表 4.2.2），由于篇幅有限表 4 展示了 20 条数据。

表 4.2.4 事件与每一类的汉明距离

eventid	cluster1	cluster2	cluster3	cluster4	cluster5
201501030050	3	2	2	4	3
201501030052	4	4	4	5	4
201501030054	4	4	4	5	4
201501030055	4	4	4	5	4
201501030056	4	4	4	5	4
201501030059	5	5	5	5	5
201501030060	5	5	5	5	5
201501030061	5	5	5	5	5
201501030073	4	4	4	4	5
201501030074	4	4	4	4	5
201501030075	4	3	3	5	3
201501030082	5	5	5	5	5
201501030083	4	3	3	5	3
201501030084	5	4	4	5	3
201501030085	3	3	3	4	4
201501030090	4	3	3	5	2
201501030091	5	5	5	4	5
201501030092	3	3	3	4	4
201501030093	5	5	5	5	5
201501030096	4	4	4	5	4
201501030097	5	4	4	5	3

5．汉明距离转化为嫌疑分数

我们通过汉明距离表示事件与类的相似性，汉明距离越短，说明事件属于该类的可能性越大。通过观察事件与每一类的汉明距离的结果，我们发现汉明距离的范围为 0 到 5 的整数，根据题目要求我们需要对距离进行评分，将其分为五个等级，由 1 到 5 表示嫌疑度逐渐降低，打分表详见表 4.2.5。

表 4.2.5　汉明距离与嫌疑度的转换评分表

汉明距离	嫌疑度评分
0	1
1	2
2	3
3	4
4	5
5	

根据题目意思以及上述分析结果可知，我们的每一个类代表的是每一个组织或个人，评分则代表恐怖分子关于典型事件的嫌疑度。我们通过汉明距离与嫌疑度评分的转换为排序结果（附件表 4.2.3），表 4.2.6 仅展示 20 条数据。

表 4.2.6 恐怖分子关于典型事件的嫌疑度

eventid	嫌疑人 1	嫌疑人 2	嫌疑人 3	嫌疑人 4	嫌疑人 5
201603040003	4	1	3	5	4
201603070032	4	1	3		2
201603080042	2	1	2		2
201603080043	3	1	2	4	3
201603080054	3	1	2		2
201604050006	3	1	2		2
201604050007	2	1	2	2	
201604050008	3	1	2	4	3
201605310031	2	1	2		2
201606020041	3	1	2		2
201606050039	3	1	2	4	3
201606050040	2	1	2	2	
201606060001	2	1	2	2	
201607260022	2	1	2	3	3
201607260023	3	1	2	4	3
201607290046	2	1	2	2	
201607310055	2	1	2	2	
201506190017	3	1	2	4	3
201506190018	2	2	2		2
201506190033	3	1	2	4	3
201506200017	3	1	2		2

4.3 问题三模型建立与求解

4.3.1 问题分析

对未来反恐态势的分析和研判需要现有数据的提供依据，解答此问可主要按不同的特征变量对案件群体进行划分，并计算频数和频率，然后通过图表的方式对主要原因、时空特性、蔓延特性、级别分布等规律进行可视化分析，主要用到的图表类型包括饼图、折线图、柱状图等。

4.3.2 恐怖袭击的主要原因

恐怖袭击的主要原因可以从受害者类型中进行挖掘，下图是近三年不同受害者类型在总事件中所占比例：

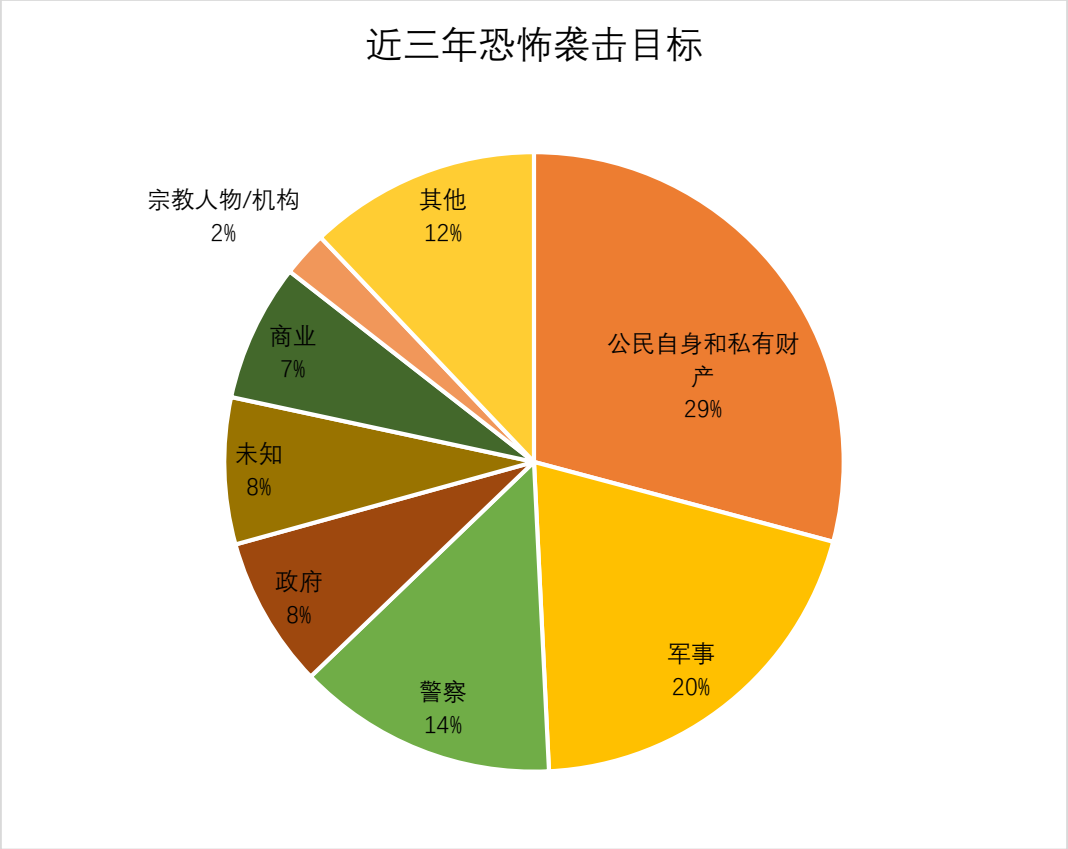

图 4.3.1 近三年不同受害者类型在总事件中所占比例

饼图说明，恐怖袭击的最主要原因是对公民生命和财产的掠夺，以及对军队和政府的反抗，如在战乱地区的反政府武装冲突等。此外，我们也考察了判定恐怖袭击的 3 条入选标准，结果如下图，可以看到，绝大部分的恐怖袭击事件出于以实现政治、经济、宗教或社会目标为目的以及意图胁迫、恐吓、或将其他信息传达给比直接的受害者更多的观众（或听众）；有少部分性质更加恶劣，超出合法战争活动的背景范围，以非战斗人员为目标（即行为必须超出 1949 年 8 月 12 日，日内瓦公约的附加议定书中国际人道主义法所允许的范围）。

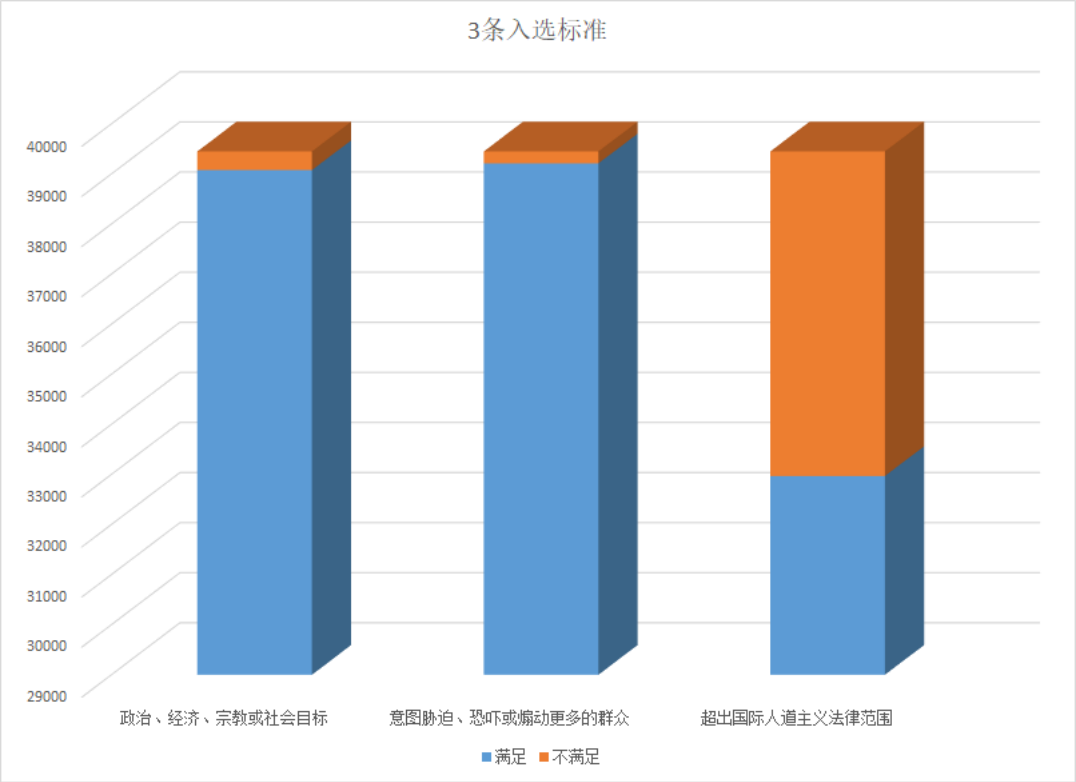

图 4.3.2 考察判定恐怖袭击的 3 条入选标准

4.3.3 恐怖袭击的时空特性

我们可以从时间和空间两个层面来研究近三年的恐怖袭击。下图是按照国家类别统计的恐怖袭击发生件数的地图表示，颜色越深表示恐怖袭击发生越频繁，影响越严重。

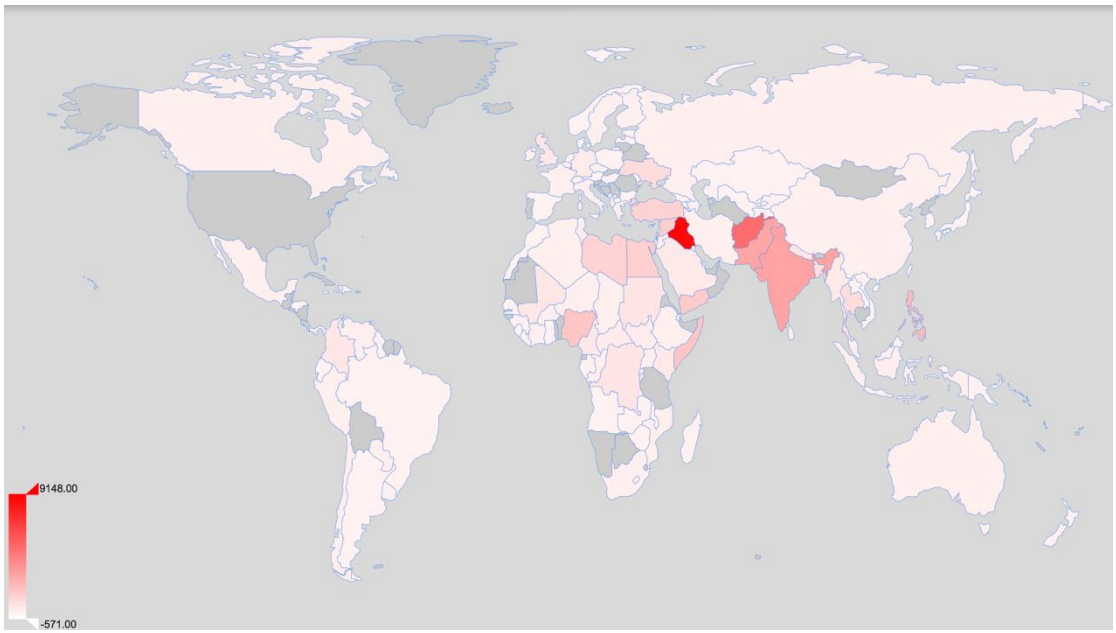

图 4.3.3 近三年的恐怖袭击的时空特性

我们也制作了按国家和地区划分的恐怖袭击事件发生比例的饼图，如下图所示，可以看到伊拉克、阿富汗、印度以及巴基斯坦是恐怖袭击最激烈的国家，与

上述地图相符。从地区而言，中东非洲以及南亚是冲突最频发的地区。

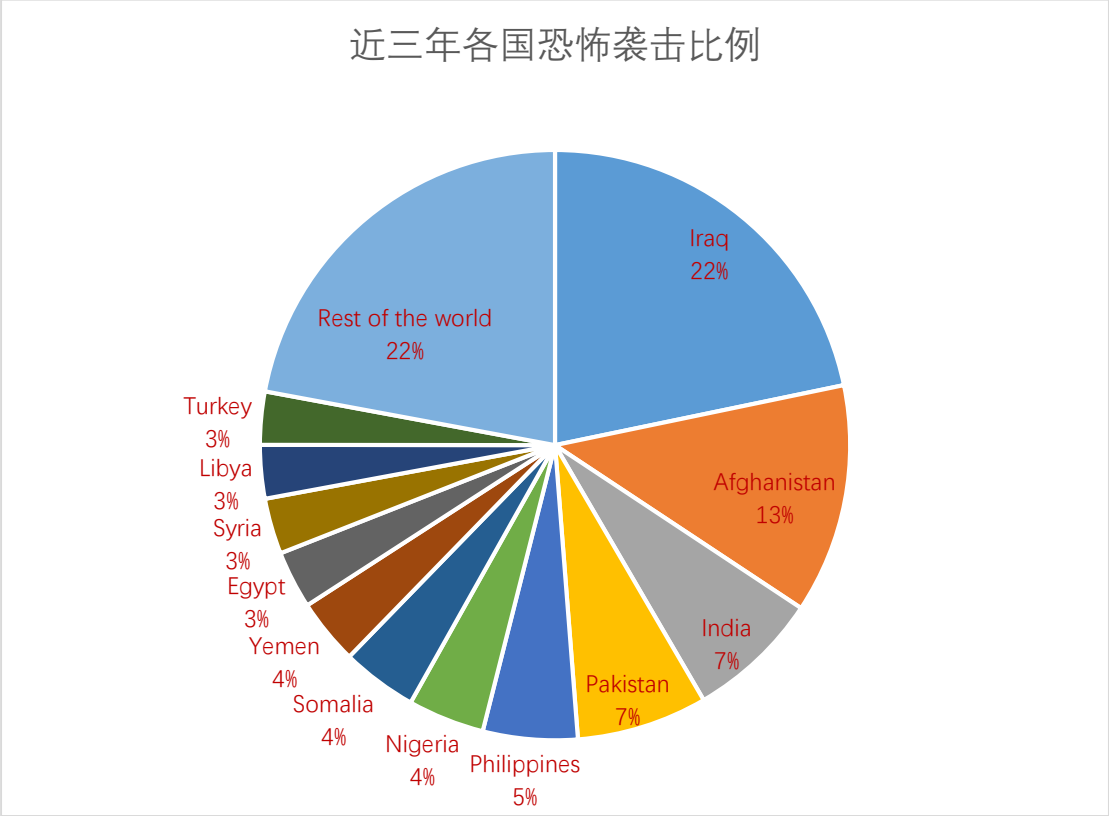

图 4.3.4 近三年各国恐怖袭击比例

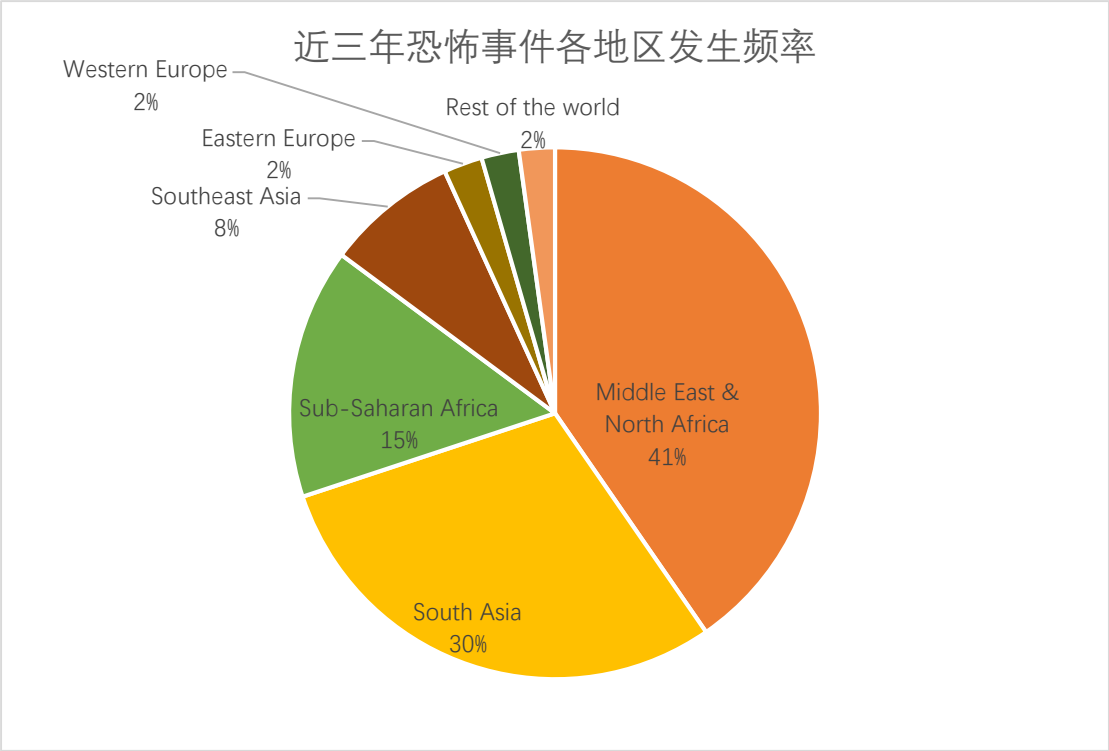

图 4.3.5 近三年恐怖事件各地区发生频率

从时间上来看，恐怖袭击年总数呈逐年下降的趋势，说明全球各国的反恐努力发挥了一定成效。

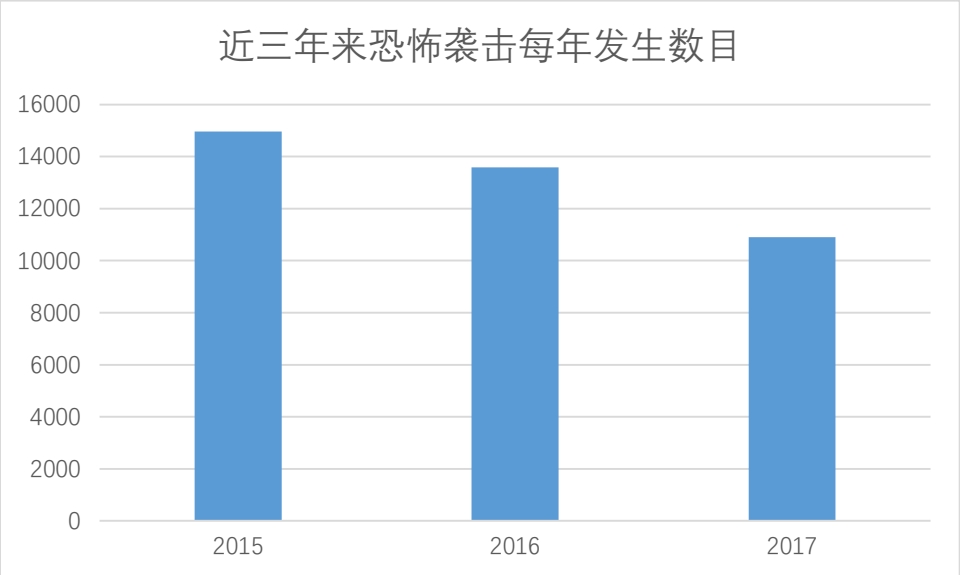

图 4.3.6 近三年来恐怖袭击每年发生数目

4.3.4 恐怖袭击的蔓延趋势

占据恐怖袭击事件总数排行前四的国家名单在近三年没有变化，可以看到，尽管恐怖袭击事件总数在近三年有递减的倾向，但是重点国家的恐怖袭击事件数仍居高不下，事件总数的减少可以归因于世界其他国家在反恐中做出的努力，我们仍要对排在前几的重点国家提供援助和严密看守。

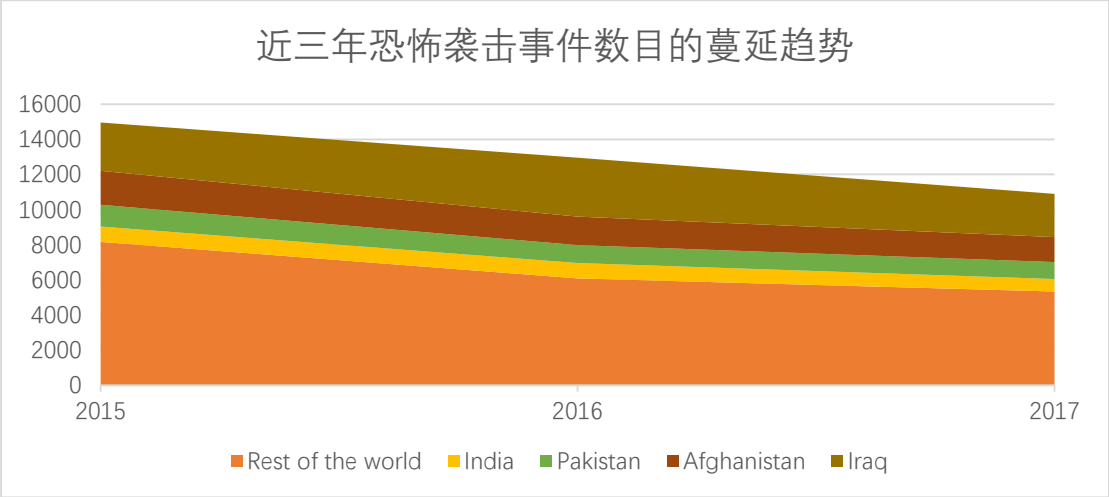

图 4.3.7 近三年恐怖袭击事件数目的蔓延

4.3.5 恐怖袭击的级别分布

除了数量外，恐怖袭击的级别也值得关注。利用在第一题中我们建立的恐怖事件分级模型，对近三年的事件级别分布进行计算，结果如下图所示，虽然恐怖袭击事件数目逐年下降，但是危害程度最大的一级事件仅在 2017 年发生一起，在 2015 和 2016 年未有一级事件发生，这说明除了减少事件的发生以外，还要对性质极其恶劣的一级二级事件进行额外监测。

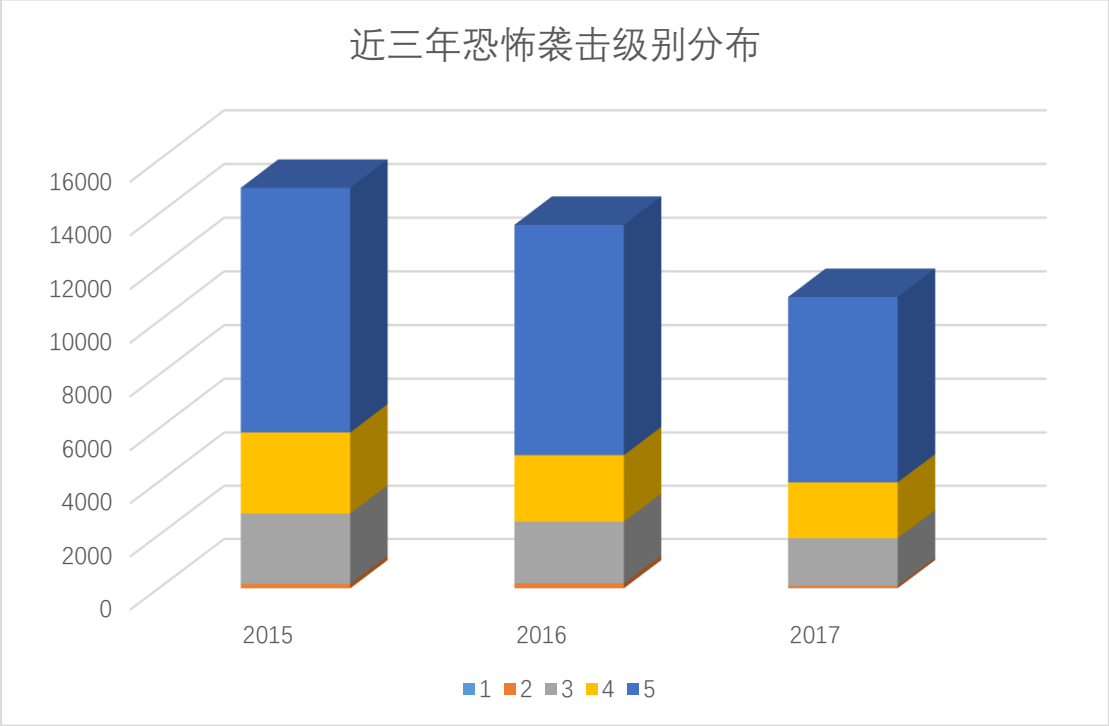

图 4.3.8 近三年恐怖袭击级别分布

4.3.6 对未来反恐态势的分析和建议

预计下一年依旧会延续近三年的恐怖袭击事件数目下降的趋势。恐怖分子活跃国家如伊拉克，阿富汗、印度以及巴基斯坦仍应受到密切关注，各国应联合起来打击惯常在这些区域进行恐怖犯罪的组织，并且对于高伤害级别的恐怖事件做好战略应对准备。随着先进的大规模杀伤性武器的投入使用，未来在恐怖分子活跃区域生活的平民的生命和财产将继续受到较大威胁，需要发展更先进的监测技术和武器来预警和打击极端恐怖犯罪。

4.4 问题三模型建立与求解

4.4.1 问题分析

数据的进一步利用，统计分析了往年金砖四国发生的恐怖袭击数量以及伤亡人数。

4.4.2 模型建立与分析

我们统计分析了往年金砖四国发生的恐怖袭击数量以及伤亡人数，结果如下图所示。

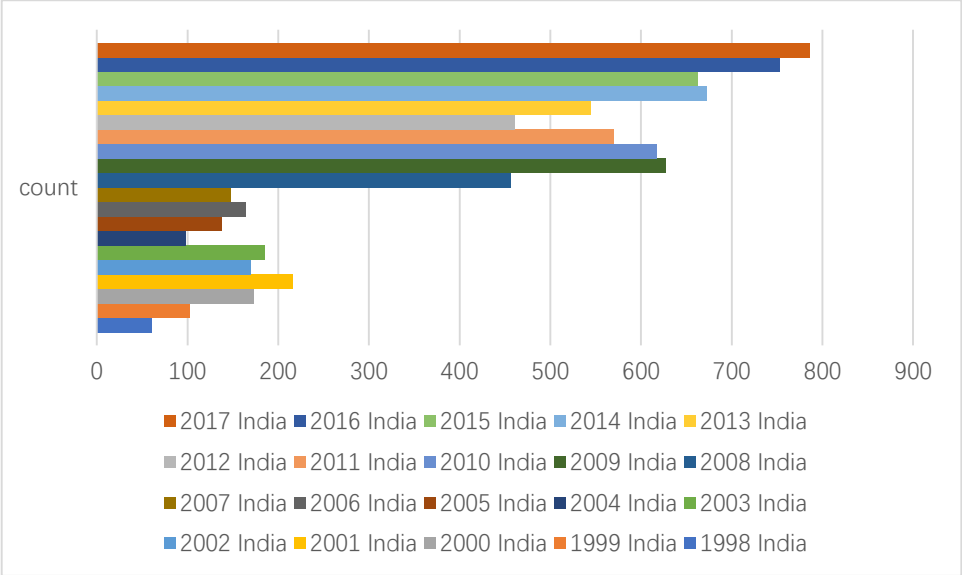

图 4.4.1 印度恐袭成功数量

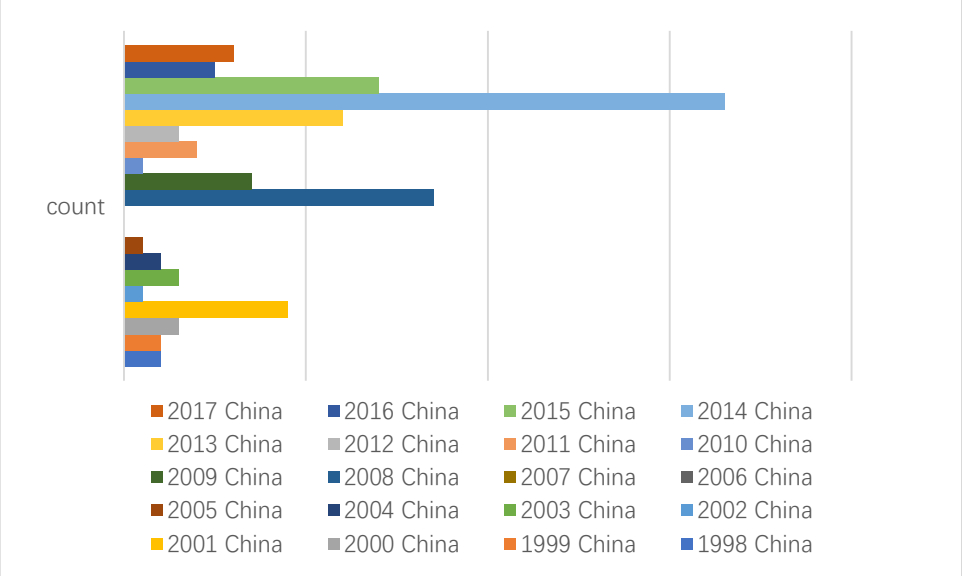

图 4.4.2 中国恐袭成功数量

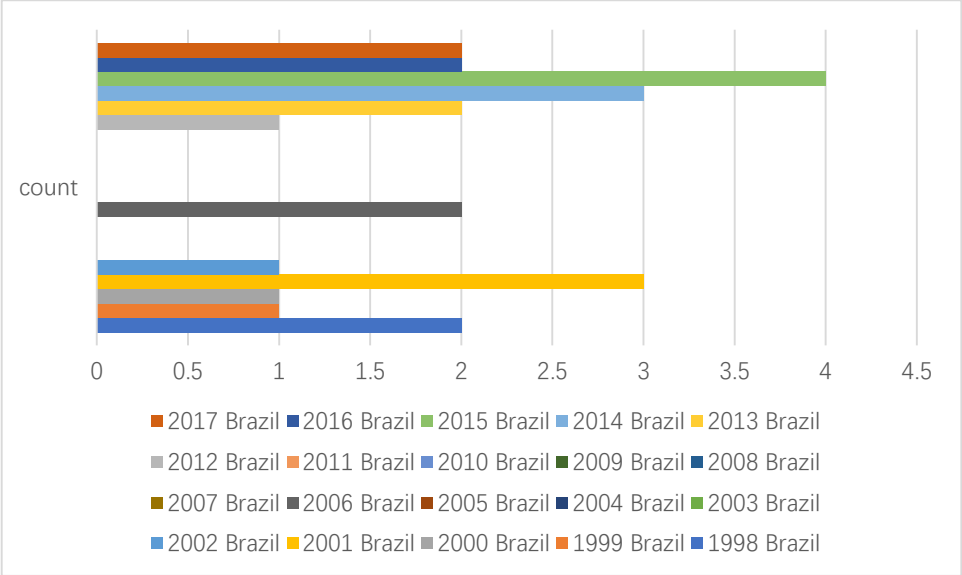

图 4.4.3 巴西恐袭成功数量

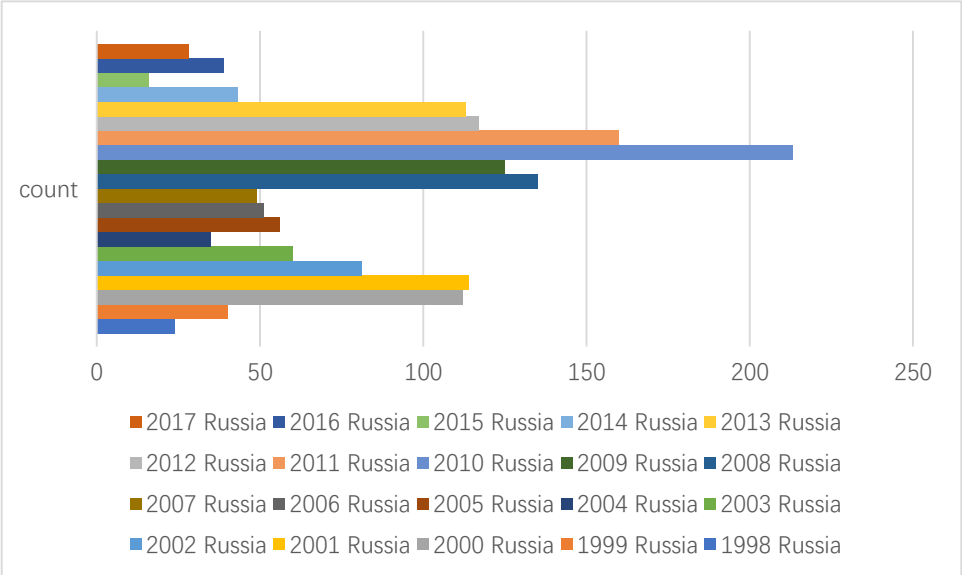

图 4.4.4 俄罗斯恐袭成功数量

综合来看，我们能够发现饱受恐怖主义威胁的大部分国家，经济发展也不好，扼制好恐怖主义，维护安全稳定，显然是国家经济发展的重要保障。一些发达国家受恐怖主义事件威胁，主要与内部的分离主义和国际地位有一定关系，但这些国家都能有效地空置恐怖主义，让恐怖袭击造成的死亡人数降下来。但是防范恐怖主义要常抓不懈，否则一旦懈怠，单词突如其来的事件造成的伤亡就非常大。

从金砖四国的恐怖袭击死亡人数来看，印度和俄罗斯的恐怖主义负担很大，中国和巴西在未来发展中受这个因素的影响比较小。从恐怖袭击成功次数的比较来看，在中国和巴西真正成功的恐怖袭击次数并不多。

结合上面几张图的分析来看，中国和巴西比另外两块“金砖”受恐怖主义的威胁更小，更具有发展优势/中国民众应该对政府在遏制恐怖主义的能力有信心。最后想说的是，恐怖主义最大的威胁并不是杀伤生命，而是借机制恐慌和创造更多的矛盾，将局面搅得错综复杂，从而造成社会动荡。所以恐怖主义事件发生后，最重要的固然是首先提高预防水平和级别，然后则要致力重塑民众对于社会生活安全的信心。

5.模型评估与反思

5.1 模型评价

5.1.1 任务一模型评价

任务一建立的是线性模型，用简单的方法去解决复杂的问题，简明易懂。本模型受到GTI（Global Terrorism Index）[1]的启发，GTI使用了GTD的四个变量，建立一个线性模型，通过人为定义权重，计算出了GTI，即各国家的恐怖活动严重程度分数，该模型参考对象专业，求解过程严谨，结果可信度高，说服力强。

5.1.2 任务二模型评价

任务二的解决是基于任务一的结果的，巧妙使用 python、R 语言编写程序，通过feature筛选，我们选出的变量都是分类型变量，聚类方法中k-modes模型专门用来解决分类型变量聚类问题。模型理论严谨，假设大胆合理。

5.1.3 任务三模型评价

任务三通过计算频数和频率，然后通过图表的方式对主要原因、时空特性、蔓延特性、级别分布等规律进行可视化分析。

5.2 反思

题目提供了数量型、分类型和文本型数据，我们在模型中没有使用文本型数据，只考虑了一部分常见的变量，可能会忽略一些影响因素。对于任务三，我们可以应用RNN、LSTM等自然语言处理算法进行时间序列预测。对于任务四，可以考虑关联规则，利用Apriori算法挖掘数据集，找出存在于项目集合或对象集合之间的频繁模式、关联、相关性或因果结构。可能会从大量数据中发现项集之间有趣的关联和相关联系，得出更多有价值的结论。

参考文献

[1] Countries L O. Global Terrorism Index[J]. 2016.
[2] 雷锋网，机器学习之确定最佳聚类数目的 10 种方法，http://tech.sina.com.cn/roll/2017-10-07/doc-ifymrqmp9824367.shtml, 2018/9/19
[3] tyh70537, k-modes 聚类算法介绍, https://blog.csdn.net/tyh70537/article/details/78158674, 2018/9/19
[4] zdy0_2004, 汉明距离, https://blog.csdn.net/zdy0_2004/article/details/49054113, 2018/9/19