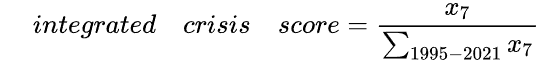
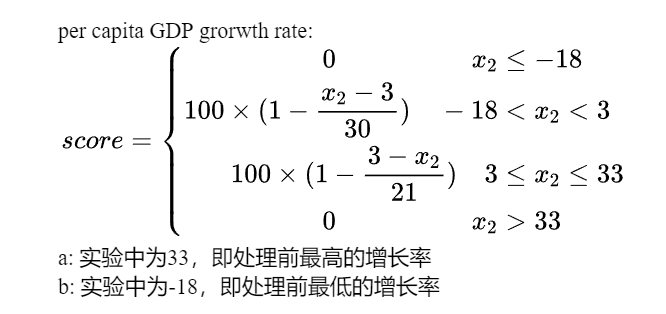
### 4.2.1 数据处理

我们的数据集中包含21个国家的7个指标在1995年至2021年间（共27年）的变化，为了方便建模，我们建立一个矩阵（567\*7）来储存所有的数据。其中，每一行代表同一个国家在同一年的七个不同指标数据，每列代表同一个指标在不同国家的27年间，每年的取值。

考虑到收集到的数据中，综合性社会危机的取值与国家体量有非常大的关联性，一些大国的综合性社会危机的指标在数值上甚至能超过小国两个数量级。因此为了排除国家体量对模型带来的干扰和误差，我们对综合性社会危机这一指标作以下处理：将其转换为该国该年的综合性社会危机占该国这27年总社会危机的比值：



在实验开始前，我们设想人均GDP增速与社会稳定性呈正相关，即人均GDP增长越快，社会越稳定。但初步实验后我们发现：在11个发生了颜色革命的国家，颜色革命发生时，该国的人均GDP增速也往往比较快，这与我们最初的设想不符。在查阅相关资料和文献后，我们发现实际上除了经济倒退以及发展过慢，经济发展得过快同样不利于社会稳定，只有社会经济以适中的速度发展时，社会才是最稳定的状态，而这个适中的发展速度通常指年人均GDP增幅在百分之3左右。因此我们将该指标设为区间型指标，处理原则如下：



同时，为了避免量纲不同带来的系统误差，我们对上述所有数据按列进行归一化，将各个指标的量纲降至同一水平。

### 4.2.2权重确定

### 4.2.2.1 熵权法

AHP法通过主观赋予7个指标各自的权重，经我们尝试后得到的实验结果非常不理想，不能明显区分出发生过颜色革命的国家，因此我们选择尝试了非主观的赋权重方法——熵权法。

首先计算每个指标的熵：

然后计算每个指标的变异系数：

得到的变异系数向量代表了每个指标所包含的信息量大小，也就是说指标所包含的信息量越大，其在变异系数向量中所对应的分量就越大，以变异系数向量作为权重向量是非常客观的一个选择。

### 4.2.2.2 随机森林算法

经实验，熵权法得到的权重向量仍然无法有效使Topsis模型有着良好的评价效果，因此我们采用预警模型常用的随机森林算法。

我们团队认为：社会稳定预警模型需要较为激进，即宁可让模型在社会稳定时错误地发出不稳定预警，也要尽可能减少社会发生动荡，而模型没有出现预警的情况。因为两种情况之下，后者的代价明显更大。同时我们认为发生颜色革命，只是社会动荡的诸多表现之一，而并非社会动荡的决定性因素（其他因素诸如游行示威、恐怖袭击、大罢工等），因此经过讨论，我们将模型需要达到的效果简化为：

1. 对社会的稳定程度进行区分，尽可能得将爆发了颜色革命的社会归为“不稳定”
2. 模型不能过多地预测社会是“不稳定”的。数据的正样本为“发生了颜色革命的国家和年份”，负样本为“没有发生颜色革命的国家和年份”。所以我们选择准确率和召回率作为随机森林模型的评价标准，希望我们的模型尽可能拥有较高的召回率，即成功的把颜色革命的发生归为“不稳定”，也要尽可能维持做出社会“不稳定”判断的比例，避免模型盲目将所有社会认定为“不稳定”，以最大化召回率。

因为随机森林预测得到的是样本为正的概率，我们认为只要概率不为0，即可以认为社会“不稳定”。根据这个法则，经过多次实验，我们最后选定的随机森林模型的准确率和召回率如下：

此时，模型对应的权重向量如下：

经过讨论，我们团队计划使用随机森林算法得到的权重向量作为Topsis模型中计算所需的权重向量，而不是直接使用随机森林模型作为社会稳定预警模型。原因主要是随机森林模型预测的可读性差，模型内部结构非常复杂，同时随机森林模型的可解释性很差，很难被精确的解释。

### 4.2.3 Topsis确定阈值

将上述随机森林算法得到的权重向量用于Topsis模型，即可计算出各个国家在各个年份的得分，得分越高代表社会越稳定。经过在召回率和准确率之间的权衡，我们团队选择的不稳定阈值为0.814，稳定阈值为0.904。意为若综合评价分数低于0.814（数据中比例约为百分之25），就认为社会稳定程度较差，容易发生包括颜色革命在内的各种社会运动；若综合评价分数高于0.904（数据中的比例约为百分之25），则认为社会相对稳定，基本不会发生较大的动荡；若综合评价得分介于两个阈值之间，则认为社会稳定程度一般，有较小可能出现动荡。对于发生了颜色革命的11个国家，在它们发生颜色革命的年份中，有9个被认定为社会不稳定，另外两个国家的颜色革命发生时，社会稳定性被认定为一般。经过综合考量，我们认为0.904和0.814最合适的临界点，可以成功预测绝大多数的颜色革命的发生，将其认定为“不稳定”，同时又没有过多得对社会做出不稳定的判断。该模型的结果具有较强的参考和实践意义。在该模型的调节下，若某国在某年被认定为“不稳定”，那么该国家的政府则更需要采取举措来稳定社会正常运转，防止社会运动的发生。

同时，该模型综合考量了7项评价指标，不同指标之间相互补充，相互制衡，共同参与构成最后的预测分数，使得评价结果也更加科学。

### 4.2.4 灵敏度分析

最后，我们从敏感性的角度，对模型的每个指标进行了单独的分析，横轴代表各项指标，纵轴代表分数。分析结果如下：

