1. **各指标的相关性分析**

相关系数是最早由统计学家卡尔·皮尔逊设计的统计指标，是研究变量之间线性相关程度的量，用字母r表示，如公式1。

其中，为X与Y的协方差，为X的方差，为Y的方差。

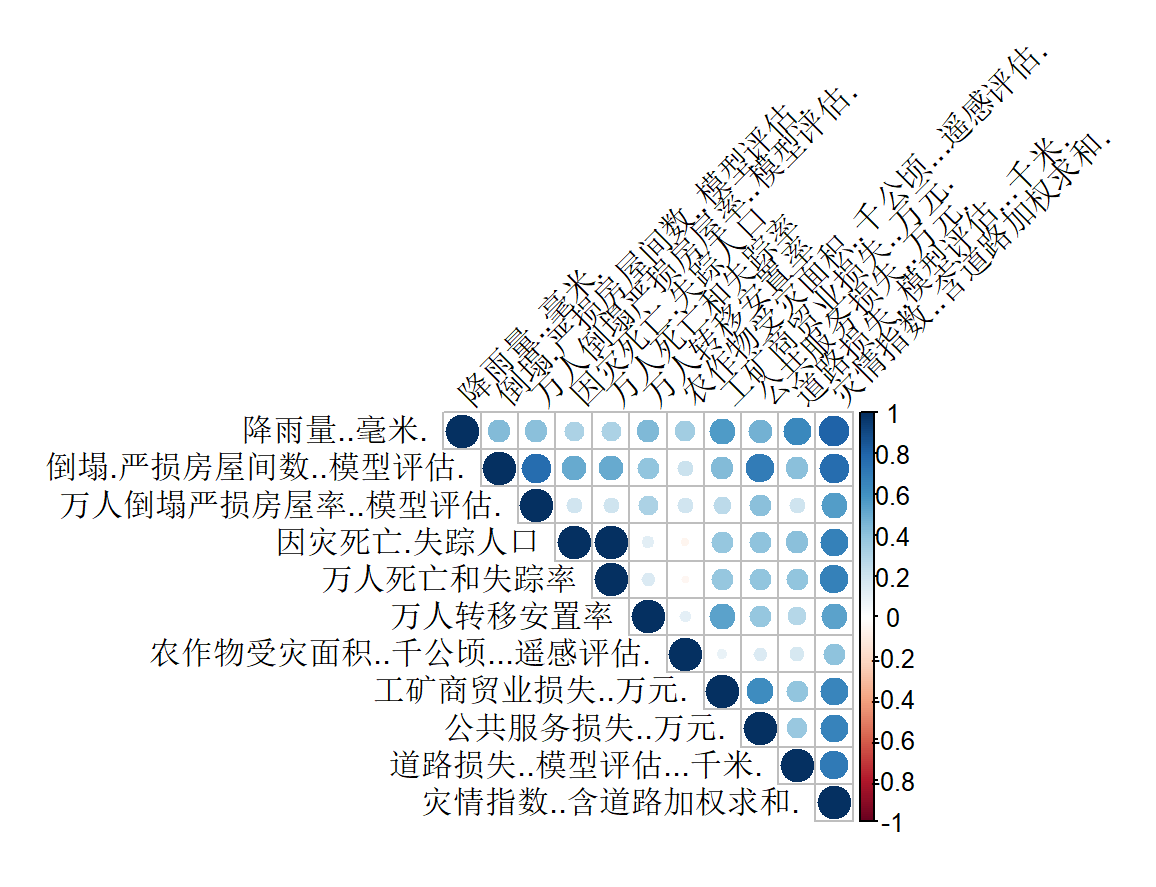


图1指标相关性表示图



表1指标相关系数值

图1中2个指标的相关性越大，对应的图形越大，颜色越深。例如，降雨量与工矿商贸损失和道路损失的相关性较大，对照表1，相关系数r为0.566884875和0.646525541；倒塌严损房屋间数与万人倒塌严损房屋率和公共服务损失的相关性较大，相关系数r为0.763454255和0.693976006；

工矿商贸业损失与公共服务损失的关性较大，相关系数r为0.621837436。

同时，还可以通过相关性反应各指标与灾情指数的关系，其中农作物受灾面积相关系数r为0.404887，相关系数较小，为低度相关。

|  |  |
| --- | --- |
| 相关系数r | 指标间关系 |
|  | 显著性相关 |
|  | 高度相关 |
|  | 中度相关 |
|  | 低度相关 |
|  | 相关性极弱 |

表2 相关系数与指标间的关系

1. **熵权法权重与直观感觉不一致的原因分析（如公共服务设施损坏的权重高于降水量）**

熵权法是根据各指标传输给决策者的信息量的大小来确定指标权数的方法。熵权法作为客观赋权法，能反应指标数据间的重要性和差异性，评价结果具有客观性。某项评价指标的差异越大，则信息熵越小，该指标包含和传输的信息越多，相应权重越大。根据熵的特性，可以用熵权来判断某个指标的离散程度，指标的离散程度越大，该指标对综合评价的权重影响越大。比如样本数据在某指标下取值都相等，则该指标对总体评价的影响为0，权值为0.

熵权法的计算公式如下：

1. 建立评价指标矩阵为
2. 归一化处理
3. 计算信息熵：

式中：*fij*为第*j*项指标下第*i*个样本数量的比重；*k*为待定常数；*Ej*为第*j*项指标的信息熵；*m*为样本总数；*n*为指标总数。

（4）确定第*j*项指标权重*wj*：

熵权法权重与直观感觉不一致，是因为熵权法计算出的权重仅依赖于数据本身的离散性，降雨量的离散程度并没有公共服务设施损坏的离散程度高，所以降雨量的权重低于公共服务设施损坏权重。

其次降雨量不是评价地方灾情损失严重程度的直接指标，例如相同降雨量，在不同的城市就会产生不同的效果，经济发达和不发达的城市损失会差别很大。所以降雨量的权重低于公共服务设施损坏权重是合理的。

1. **如何认为BP神经网络出来的结果是最优的，如误差（公式）。**

为了判断BP神经网络输出结果的优劣程度，采用了均方误差，公式如下。

式中，为第*i*个受灾城市样本的输出层的期望输出，为第*i*个受灾城市样本的输出层的输出，均方误差越小准确性越高，均方误差控制在了10-4之内。

1. **分等定级划分的原理和公式**

最优分割法是一种对有序样本数据进行分段的统计方法，在给定的分段数下，各分段的起讫点是客观确定的，其实质是一定目标函数下不破坏样本顺序的聚类法。有序样本最优分割法在水文、气象、地震、地质、环境等自然科学和经济、管理等社会科学领域均获得了广泛应用。最优分割法对序列数据进行分段，使所分的段内差异最小、段间差异最大由于段内离差平方和与段间离差平方和所构成的序列数据的总离差平方和为固定量，故所分的各段内离差平方和最小化的分割即为最优分割。

现设长度为N的序列数据为，被分成S段，第s段内的数据量为。显然。序列数据的均值为、总离差平方和、第s段内均值、段内离差平方和的公式分别为：

式中是指序列前s段数据个数。

段间离差平方和为：

5. 降维的作用

PCA（Principal Component Analysis，主成分分析），线性DR技术。降维致力于解决三类问题：

1. 降维可以缓解维度灾难问题；
2. 降维可以在压缩数据的同时让信息损失最小化；
3. 理解几百个维度的数据结构很困难，两三个维度的数据通过可视化更容易理解。

PCA作为一个非监督学习的降维方法，它只需要特征值分解，就可以对数据进行压缩，去噪。因此在实际场景应用很广泛。为了克服PCA的一些缺点，出现了很多PCA的变种，比如为解决非线性降维的KPCA，还有解决内存限制的增量PCA方法Incremental PCA，以及解决稀疏数据降维的PCA方法Sparse PCA等。

PCA算法的主要优点有：

1、仅仅需要以方差衡量信息量，不受数据集以外的因素影响。

2、各主成分之间正交，可消除原始数据成分间的相互影响的因素。

3、计算方法简单，主要运算是特征值分解，易于实现。

PCA算法的主要缺点有：

1、主成分各个特征维度的含义具有一定的模糊性，不如原始样本特征的解释性强。

2、方差小的非主成分也可能含有对样本差异的重要信息，因降维丢弃可能对后续数据处理有影响。

在将高维数据映射到适合执行K-均值的特定空间上花费了大量精力。为此，采用了多种技术，包括主成分分析（PCA）、典型相关分析（CCA）、非负矩阵分解（NMF）和稀疏编码（字典学习）。除了这些线性DR算子（例如，投影矩阵），还考虑了非线性DR技术，如谱聚类（Ng等人，2002）和稀疏子空间聚类（Elhamifar&Vidal，2013；Y ou等人，2016）中使用的非线性DR技术。

近年来，由于深度神经网络（DNN）在有监督学习中的成功，无监督深度学习方法在聚类前被广泛应用于DR。例如，堆叠式自动编码器（SAE）（Vincent等人，2010年）、深度CCA（DCCA）（Andrew等人，2013年）和稀疏式自动编码器（Ng，2011年）分别从PCA、CCA和稀疏编码中获得见解，并利用DNN学习从数据域到低维潜在空间的非线性映射。这些方法将其DNN视为预处理阶段，与后续聚类阶段分开设计。希望这些DNN学习到的数据的潜在表示自然适合于聚类。然而，由于在学习过程中没有明确纳入聚类促进目标，学习的DNN不一定输出适合的降维数据

在（De Soete&Carroll，1994；Patel et al.，2013；Y ang et al.，2017）中，考虑了联合DR和聚类。这项工作背后的基本原理是，如果存在一些潜在空间，实体可以很好地分成簇，那么自然会寻求揭示这种结构的DR变换，即产生较低的K-均值聚类成本。这促使在潜在空间中使用K-means成本作为先验，帮助选择正确的DR，并推动DR产生K-means友好表示。通过联合进行DR和K-均值聚类，在（Y ang等人，2017年）中观察到了令人印象深刻的聚类结果。这些工作的局限性在于，通过简单的线性变换，可以从潜在的聚类友好空间生成可观测数据。虽然简单的线性变换在许多情况下都能很好地工作，但在其他情况下，生成过程更为复杂，涉及非线性映射。

6、联合聚类公式