梳理并填写全国全部县区、城区（当前只覆盖项目的8个试点省份）的区域多维特征向量。采用加权聚类算法，计算形成聚类集合结果。在聚类算法中，采用了K-Prototypes聚类方法，对于类别指标采用K-mode方法进行距离计算，对于数值指标采用K-means方法进行距离计算，最终对两种距离进行加权求和得到数据的混合距离。

首先，对县区、城区各个维度的特征指标进行预处理。各个维度的特征指标主要分为数值类型和类别类型两类。对于数值类型的指标，保留其数值并做归一化处理。数值类型的指标可以细分为经度、纬度、上一年度人均地区生产总值、第一产业比重、第二产业比重、第三产业比重、高等教育毛入学率、经常参加体育锻炼人口比例、上一年度人均文化活动场次、千人拥有医生数、千人拥有床位数、私家车人均拥有量、互联网用户比例、移动互联网用户比例、人均接待游客数量、上一年度人均用电量、上一年度人均工业用电量、上一年度人均居民用电量、近三年平均PM2.5浓度、近三年平均人均水面面积、近三年平均气温、近三年平均降水量等共22个维度。

对于类别类型的指标，采用独热编码（One-hot）进行处理。各指标的具体处理方式如下：

* 民族：定义一个56维的编码向量，每个维度表示一个民族，将人数占比前五的民族置1，其余置0
* 方言：定义一个方言one-hot编码向量，向量的维度是统计的方言种类数。
* 戏曲：定义一个戏曲编码向量，向量的维度是统计的戏曲种类数
* 交通：有无高铁、机场指标作为二维向量，有则置1，无则置0
* 地貌：定义一个5维的地貌one-hot编码向量，将该地区地貌置1，其余置0

综上，在数据预处理之后，民族、方言、戏曲、交通的有无高铁/机场、地貌是类别指标，其余部分为数值指标。

加权K-Prototypes聚类算法流程：

1. 首先筛选出满足双向用户占比达到xx%的县区和城区，在这些县区和城区中间选取聚类簇数K与初始聚类中心。

（1）采用手肘法选取聚类簇数K。

首先进行经验估计，K值一般在4-8左右最好，然后利用手肘法绘制出随着K值变化损失函数变化的曲线，找到损失函数的拐点，该拐点数值就是最佳聚类簇数K。手肘法的损失函数是SSE（sum of the squared errors）：

其中代表第i个簇，是簇里的样本点，是簇的质心。随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。并且，当k小于最佳聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达最佳聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的最佳聚类数。

（2）计算所有数据点的高密度最小距离，按照从大到小进行排序，将值最大的k个密度极值点作为初始聚类中心。

1. 权重初始化，对于5种类别指标，每种指标赋值权重1/5，对于数值指标，共有22个维度，因此每个维度的数值指标赋值权重1/22。
2. 计算县区城区数据集中每个县区城区到这K个聚类中心的加权部分欧氏距离，并将每个县区城区分配给离它最近的中心点所代表的簇。计算每个簇中的所有数据点的平均值作为每个簇的新的中心。

（1）民族、方言、戏曲、交通、地貌这5个指标进行了one-hot编码，按照K-mode方法计算距离。

例如，对于数据点(j=1,2,… …，N)，分别计算其与k个中心之间的距离，这里的距离为不同编码值的个数，例如在地貌属性中，=[0，0，0，1，0],=[1，0，0，0，0]，那么与之间的距离为2。

（2）其余22个维度由数值来表示，采用K-means方法计算距离。加权欧式距离的计算方法如下：设县区城区数据集为C，其中有n个数据{Xi∈C, i=1,2,…,n},每个数据的维度是m=22，则对任意Xi，Xi=(xi1,xi2,…,xim)，对任意两个数据Xi和Xj，定义加权部分距离如下：



式中，wk是第k维的权重，Ik是第k维缺失的度量，如果xik与xjk有一方数据缺失，那么令Ik=0，否则Ik=1.令,那么I就表示未缺失指标的维数，使用来按比例扩展未包含缺失维度的欧式距离。

数据点的数值指标采用K-means方法得到，类别指标采用K-modes方法，最终混合距离*D*：

其中，和分别是数值指标和类别指标的权重，和为1.

计算数据对象与聚类中心的加权系数cor：

1．用pearson相关系数计算数据对象间的相关性系数

2．寻找数据集中，相关性系数绝对值最大值Max\_r和绝对值最小值Min\_r

3．通过离差标准化对数据对象间的相关性进行规范化，

4．距离加权系数 Q取1

5．利用前面计算出的欧式距离与cor相乘作为数据对象到聚类中心的真实距离

1. 根据当前划分与聚类中心动态调整数值指标各个维度的权重值。县区城区数据集中所有数据项各维度的权重是统一的，采用如下方法确定：



其中，

式中，表示聚类数，表示第j类聚类中心的第k维数值， 表示整个数据集第k维的均值，表示第j类的数据项数量，表示第j类在第k维数值上完整值的个数， 是第k维数值权重的度量值，表示该维度在聚类的类间影响因子与类内影响因子的比值。类间影响因子越大，类内影响因子越小，表示该维度对分类的影响越大，那么分配给该维度的权重也相对增大，反之亦然。

1. 重复C、D步骤，直到聚类代价函数收敛或者聚类中心点与上次完全相同。聚类代价函数为：

其中，代表数据点被分配的聚类的索引值，∈{1,2,...,K}，是类别为k的聚类中心。代表数据点被分配聚类索引值所对应的聚类中心。

聚类结束，输出结果。