测绘观察
2021 年 2 月第 3 卷第 1 期

一种 K-Means 算法的坐标转换公共点自动选取方法

刘国栋　秦　浩　刘　佳　刘　浪

重庆交通大学土木工程学院，重庆

摘　要 | 背景：坐标转换中，在具有多个已知数据的重合点时，重合点的选取是关键问题之一。目标：一个基于 K-Means 算法在众多重合点中自动选出合适重合点的方法。方法：通过分析 K-Means 聚类算法的基本原理和最佳 k 值的确定，针对空间直角坐标数据提出了一个初始聚类中心的选取方法，并结合 Python 实现了从多个重合点中快速、自动地选取出重合点，同时使用 Bursa 模型解算出七参数和内符合精度达到要求。结果：利用 K-Means 聚类算法能够快速、自动地选出较为合适的重合点，内符合精度在最佳 k 值处最优，由此确定的重合点求解出的七参数能够用于坐标转换中。结论：研究表明基于 K-Means 算法自动选出重合点的方法不但可行而且能够满足坐标转换精度要求，对坐标转换工作具有非常重要的意义。

关键词 | 坐标转换；重合点自动选取；K-Means 算法；Bursa 模型；Python

1　引言

坐标系统是测绘工作中的基石，所有测绘成果均基于坐标而成。2018 年 7

基金项目：重庆市研究生教育优质课程建设项目（编号：JDYZ2019009）。
通讯作者：秦浩，重庆交通大学土木工程学院在读硕士生，E-mail：1752014095@qq.com。
文章引用：刘国栋，秦浩，刘佳，等．一种 K-Means 算法的坐标转换公共点自动选取方法［J］．测绘观察，2021，3（1）：9-21.
https://doi.org/10.35534/go.0301002

月 1 日起，自然资源部将全面启用 2000 国家大地坐标系，在以往的测绘工作中使用的坐标系往往不是该坐标系，为了便于管理需要进行坐标转换至国家 2000 大地坐标系上。坐标转换的精度一方面在于坐标转换模型的选取，另一方面则是在于重合点的自身精度和选取方式[1, 2]。重合点的选取需要按照均匀分布、具有一定的密度的原则[3-5]。传统的方法是通过人工进行选取，重合点数量较多时，往往难以决策出选多少个合适和具体选出哪几个点。针对以上问题，本文通过 Python 结合 K-Means 算法[6]实现了从多个重合点中快速、自动地选择出具有一定的密度和均匀分布原则的重合点，并将选出的重合点使用 Bursa 模型[7]进行了七参数的求解和对其内符合精度的评定。

2 K-Means 聚类算法

2.1 K-Means 算法原理

K-Means 是一个将没有分类标签的数据集，根据样本之间的相似性分成若干簇的聚类算法，属于无监督分类算法。其算法流程如下：

Step1：选取 k 个初始聚类中心点后，计算每个样本到各个初始聚类中心的相似程度，与 k 个初始聚类中心最相似的原则来划分数据集，将样本集合划分为 k 个子集构成 k 个类或簇；

Step2：在 Step1 中所得的 k 个类中，分别计算出每一类的聚类中心，作为新的聚类中心点，然后计算每个样本点到新的聚类中心点的相似程度，重新对数据集进行划分，从而得到不同于 Step1 得到的 k 个类；

Step3：重复 Step2 直到出现的 k 个类中的结果不再发生变化为止。

2.2 相似性度量方法的选取

聚类的核心是相似程度的度量，可分为距离或相似度。距离度量方式有闵可夫斯基距离、马哈拉诺比斯距离，其中闵可夫斯基距离又可分为欧氏距离、曼哈顿距离和切比雪夫距离；相似度度量的方法有相关系数和夹角余弦。采用距离去辨别相似程度时，距离越小样本越相似；使用相似度度量时，相似度值

越大样本越相似。不同相似性度量方法得到的结果并不一定完全相同，可能存在着差异。针对空间直角坐标并结合程序设计的复杂性，经分析决定以欧式距离作为相似性度量的方法。欧式距离的定义如下：若 X 是 m 维实数向量空间 R^m，其中 x_i，$x_j \in X$，$x_i=(x_{1i}, x_{2i}, \cdots, x_{mi})$，$x_j=(x_{1j}, x_{2j}, \cdots, x_{mj})$，样本间闵可夫斯基距离可表示为

$$d_{ij}=\left(\sum_{k=1}^{m}|x_{ki}-x_{kj}|^P\right)^{\frac{1}{P}} \tag{1}$$

式（1）中 $P \geqslant 1$，当 P=2 时为欧氏距离，其表达式如下：

$$d_{ij}=\left(\sum_{k=1}^{m}|x_{ki}-x_{kj}|^2\right)^{\frac{1}{2}} \tag{2}$$

2.3 最佳 k 值的选取

K-Means 无法直接根据数据集自动选择出最佳的 k 个类别，这也是 K-Means 中的最大弊端。k 值的选择往往会对聚类的结果造成影响，因此需要根据一定方法来确定出最佳的 k 值，从而得到最优的聚类结果。最为常用的方法有手肘法[8]和轮廓系数法[9]，其中手肘法的原理和实现较为简单，但是并不能够很直观的说明出某一个值就是最佳的 k 值，文中通过与轮廓系数法结合起来使用，以便于直观的看出最佳 k 值和聚类效果，用时也便于检验聚类结果的可靠性。

2.4 初始聚类中心的选取

K-Means 算法中还有一大缺陷就是如何选取初始聚类中心，不同的初始聚类中心会导致不同的聚类结果[10]。因此需要根据数据自身的情况根据一定的方法确定出初始聚类中心，已有的方法有基于样本的密度和样本的距离来确定初始聚类中心[11]。文中针对空间直角坐标数据特点，考虑通过分析重合点的范围来确定出初始聚类中心的方法。首先通过统计分析出所有的坐标数据点，确定出各方向的最大、最小值。由于 Z 方向的值变化一般相对较小，取 Z 值的中心范围处作为 Z 方向的值记为 z。然后通过比较 X 与 Y 的范围来确定以哪一方向来进行划分数据，若 X 的范围大，则划分 X 方向的坐标数据来确定 X 方向的值，Y 方向则取中心范围处作为 Y 方向的值记为 y；若 Y 的范围大，则划分 Y 方向的坐标数据以确定 Y 方向的值，X 方向则取中心范围处作为 X 方向的值记为 x。再

https://doi.org/10.35534/go.0301002

根据确定出的最佳 k 值，结合是以 X 方向还是以 Y 方向来对所有坐标点的范围来进行划分，以此来确定出 k 个初始聚类中心。

假设需要从一系列坐标点中筛选出三个初始聚类中心点，若统计分析出 X 方向的范围比较大时，可见筛选出的三个初始聚类中心如图 1 所示，此时的各个初始聚类中心点坐标计算公式如下：

$$\begin{aligned} x_n &= X_{\min} + (X_{\max} - X_{\min}) \div 2k \times (2n-1) \\ y &= Y_{\min} + (Y_{\max} - Y_{\min}) \div 2 \\ z &= Z_{\min} + (Z_{\max} - Z_{\min}) \div 2 \end{aligned} \tag{3}$$

若统计分析出 Y 方向的范围比较大时，可见筛选出的三个初始聚类中心如图 2 所示，此时的初始聚类中心点坐标计算公式如下：

$$\begin{aligned} x &= X_{\min} + (X_{\max} - X_{\min}) \div 2 \\ y_n &= Y_{\min} + (Y_{\max} - Y_{\min}) \div 2k \times (2n-1) \\ z &= Z_{\min} + (Z_{\max} - Z_{\min}) \div 2 \end{aligned} \tag{4}$$

式（3）、（4）中，k 为初始聚类中心点的个数，n=1，2，3，…，k 表示需要计算的第几个聚类中心点的坐标。

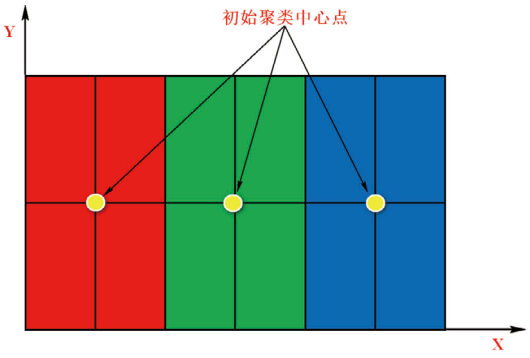


图 1　X 大于 Y 的范围

Figure 1　Range of X greater than Y

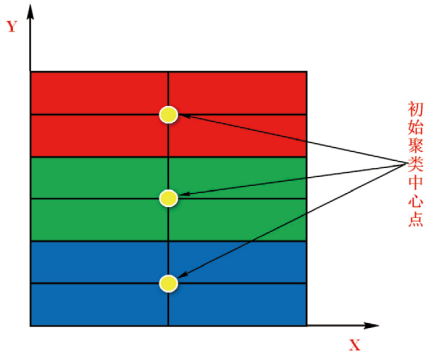


图 2　X 小于 Y 的范围

Figure 2　The range where X is less than Y

3　Bursa 七参数模型

在空间直角坐标点之间的坐标转换，Bursa 模型是最为常用，可适用于全国范围[12]。其模型如下：

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_N = \begin{bmatrix} 1 & 0 & 0 & 0 & -Z_O & Y_O & X_O \\ 0 & 1 & 0 & Z_O & 0 & -X_O & Y_O \\ 0 & 0 & 1 & -Y_O & X_O & 0 & Z_O \end{bmatrix} \begin{bmatrix} T_X \\ T_Y \\ T_Z \\ \varepsilon_X \\ \varepsilon_Y \\ \varepsilon_Z \\ m \end{bmatrix} + \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_O \tag{5}$$

设有 N 个重合点，利用间接平差模型将七个参数作为必要观测数 t=7，则总观测数为 n=3N，多余观测数 r=n−t。由此关系知至少需要 3 个坐标点才能对七参数进行解算，可建立误差方程为

$$\begin{bmatrix} V_{X_1} \\ V_{Y_1} \\ V_{Z_1} \\ \vdots \\ V_{X_N} \\ V_{Y_N} \\ V_{Z_N} \end{bmatrix}_N = \begin{bmatrix} 1 & 0 & 0 & 0 & -Z_{O_1} & Y_{O_1} & Y_{O_1} \\ 0 & 1 & 0 & Z_{O_1} & 0 & -X_{O_1} & Y_{O_1} \\ 0 & 0 & 1 & -Y_{O_1} & X_{O_1} & 0 & Z_{O_1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & -Z_{O_N} & Y_{O_N} & X_{O_N} \\ 0 & 1 & 0 & Z_{O_N} & 0 & -X_{O_N} & Y_{O_N} \\ 0 & 0 & 1 & -Y_{O_N} & X_{O_N} & 0 & Z_{O_N} \end{bmatrix} \begin{bmatrix} T_X \\ T_Y \\ T_Z \\ \varepsilon_X \\ \varepsilon_Y \\ \varepsilon_Z \\ m \end{bmatrix} - \left(\begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \\ \vdots \\ X_N \\ Y_N \\ Z_N \end{bmatrix}_N - \begin{bmatrix} X_1 \\ Y_1 \\ Z_1 \\ \vdots \\ X_N \\ Y_N \\ Z_N \end{bmatrix}_O \right) \tag{6}$$

将式（6）简化为矩阵的形式为

$$V = B\hat{X} - L \tag{7}$$

接着使用最小二乘法来求解七参数，此方法利用了所有选取的公共点，所得结果更为可靠[13]。七参数 $\hat{X}$ 计算式为

$$\hat{X} = (B^T B)^{-1} (B^T L) \tag{8}$$

将重合点视为等精度的独立观测值，因此各点之间的权阵 P 即为单位矩阵。由式（7）所求的改正数 V 可得出单位权中误差 σ_0 计算式为

$$\sigma_0 = \sqrt{\frac{V^T PV}{n-t}} \tag{9}$$

4 实验设计与分析

4.1 实验设计思路与流程

由于坐标转换参数保密的原因，本文采用模拟数据。在坐标点（60796.654，64094.649，253.465）的基础上，在 X 与 Y 方向上通过 Python 生成（1，5000）和（0，1）的随机数再加入到初始的 X 与 Y 的值作为转换前的 X 与 Y 方向的值，由于高程上 Z 值变动范围一般较小，因此在原基础上加上（1，80）和（0，1）

https://doi.org/10.35534/go.0301002

的随机数作为转换前的 Z 方向的值，以上就生成原坐标系的数据。接着通过指定三个平移参数、三个旋转参数和一个尺度参数其数值见表 2，通过 Bursa 七参数坐标转换模型得到转换后的新坐标系数据，由此方法得到 30 组重合点数据如图 3 所示，生成的新坐标系的坐标位置和点号在 ArcMap 中显示见图 4，将这些重合点作为本实验所用的数据。

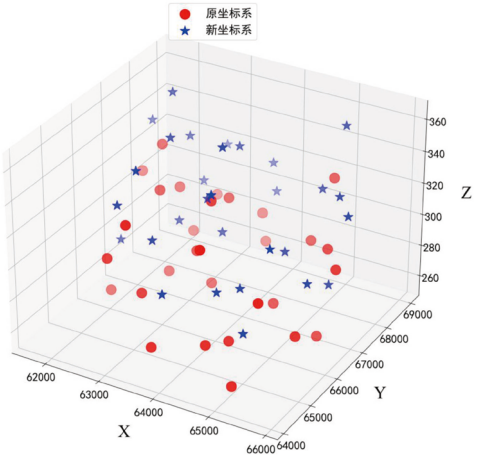


图 3　重合点坐标空间位置

Figure 3　Coordinate space position of coincidence point

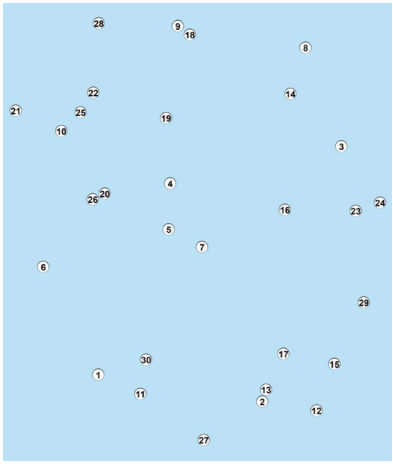


图 4　新坐标系数据点位点号

Figure 4　New coordinate system point number

接着通过手肘法与轮廓系数法对转换后的坐标数据进行聚类分析得出最佳 k 值，然后根据文中初始聚类中心的选取方法确定出 k 个初始聚类中心，接着通过 K-Means 算法对转换后的坐标数据进行聚类得出最终的 k 个聚类中心，最后筛选出各个类与其该类中心点最近的点作为选出重合点。将选出的重合点通过 Bursa 七参数模型计算出的七参数以及计算内符合精度，经分析即可得出结论，其流程图详见图 5。

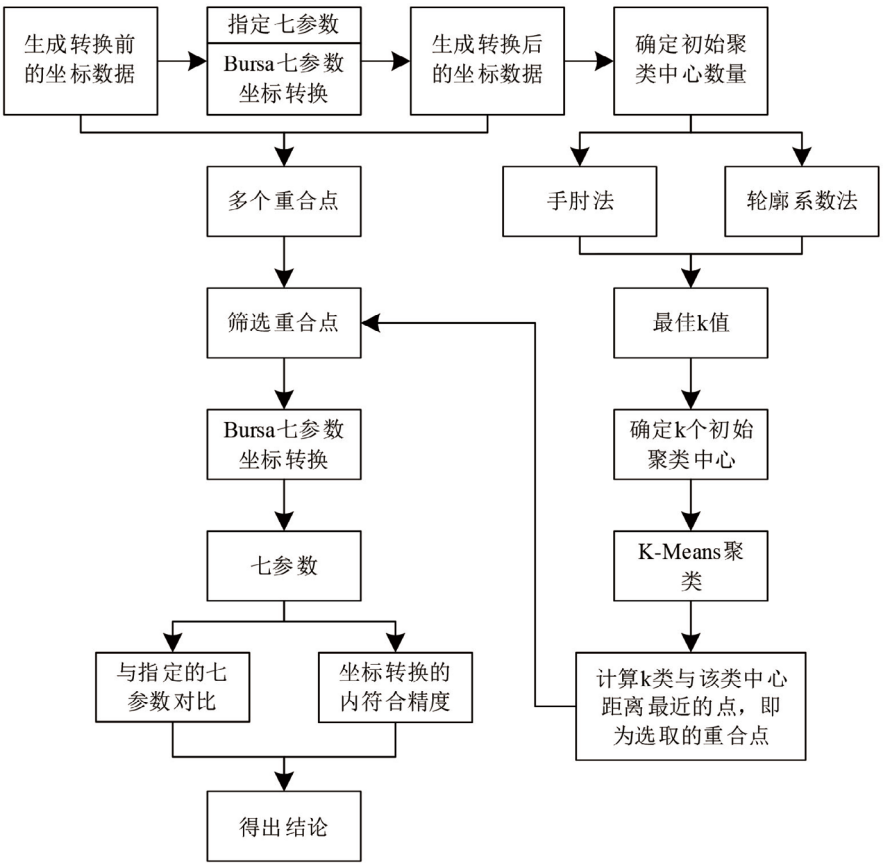


图 5　实验流程图

Figure 5　Experimental flowchart

4.2　实验结果分析

首先将生成的新坐标数据通过手肘法和轮廓系数法通过 K-Means 聚

https://doi.org/10.35534/go.0301002

类分析，绘制出的簇内离差平方和之和与轮廓系数对簇的个数的关系图如图 6 所示。根据手肘法可以看出当聚类为 7 个簇的时候为肘点，得到的聚类效果最佳，因此可以考虑选取 k=7 作为最佳 k 值从而进行接下一步的验证。由于从图 6 中手肘法所确定最佳 k 值不太直观，因此结合轮廓系数法也对每次聚类的结果进行一个评价，从图 6 中可以更直观的看出当簇的个数为 7 时的轮廓系数是最大的，以上两个方法都可以同时确定出最佳聚类个数为 7 个簇。

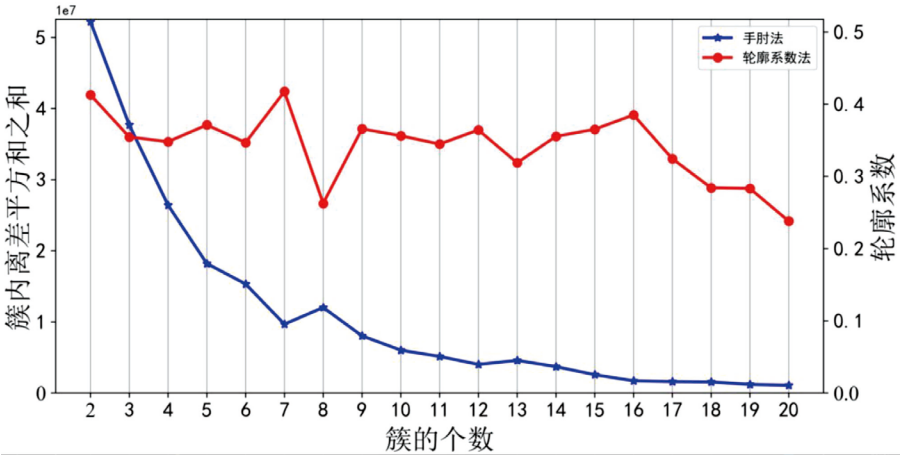
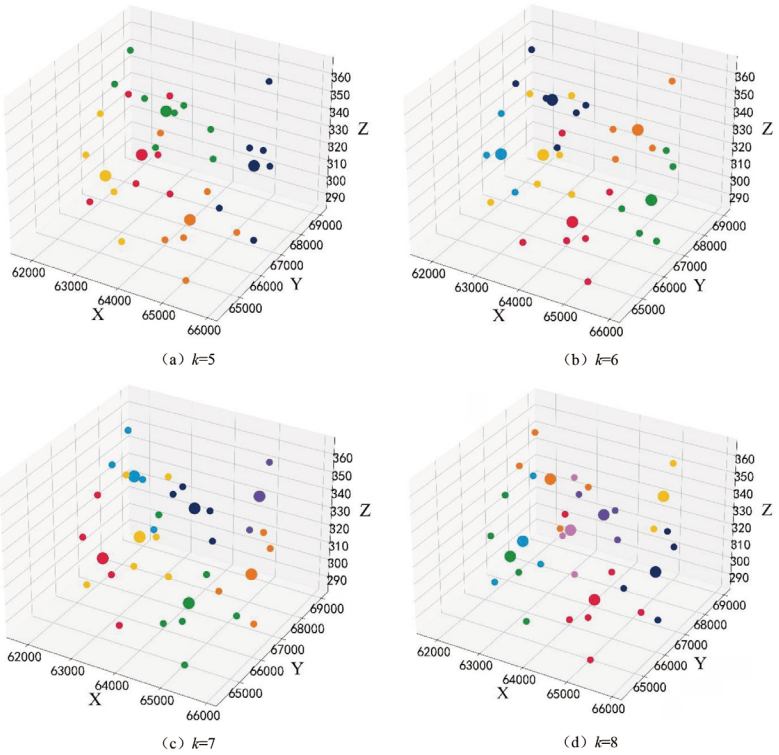


图 6　手肘法与轮廓系数法

Figure 6　Elbow method and contour coefficient method

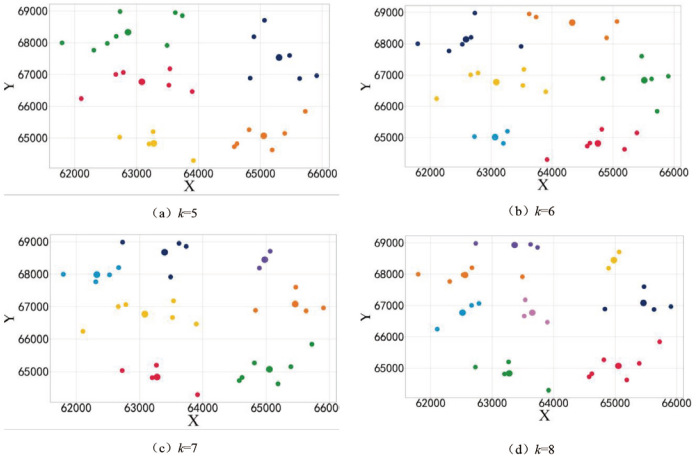
根据不同的 k 值处确定出不同的 k 个初始聚类中心后，然后将生成新坐标系数据分为 k 个簇，从而得到它们的聚类结果。为了验证在最佳 k 值处所筛选出的重合点所求得的七参数是最优的，结合图可以看出当 k=5 和 6 时的轮廓系数也较高，因此对它们进行聚类，另外选取最佳 k 值处较近的 k=8 也进行聚类。k=5、6、7 和 8 的聚类结果见图 7，因三维效果有时候看起来会误认为聚类结果是错误的，故同时也绘制了二维效果图见图 8，从图中不难看出聚类的结果都还比较好，聚类结果都是正确的。其中每种颜色表示同一簇，较大的圆为各簇的聚类中心。



（a）k=5　（b）k=6　（c）k=7　（d）k=8

图 7　K-Means 聚类结果

Figure 7　K-Means clustering results



（a）k=5　（b）k=6　（c）k=7　（d）k=8

图 8　二维效果

Figure 8　Two-dimensional effect

https://doi.org/10.35534/go.0301002

接下来计算出每一簇的各点与该簇聚类中心的欧式距离，选取以欧式距离最小的点作为每一簇所选出来的一个重合点，k 个簇即可得到 k 个重合点。通过计算，当 k=5 时所筛选出来的重合点的点号为 1、2、6、16、21；k=6 时所筛选出来的重合点的点号为 1、6、9、21、27、29；k=7 时所选出的重合点的点号为 1、2、14、16、19、21；k=8 时所选出的重合点的点号为 1、2、5、6、14、16、21、28，各点的分布位置见图 4。将选出的重合点分别通过 Bursa 七参数坐标转换模型解算出七参数和内符合精度。表 1 中是将比较不同的 4 个 k 值所解算出的七参数结果都是与其对应指定参数的差值，可以看出通过 K-Means 聚类算法筛选出的重合计算出的七参数与原先指定的七参数都比较接近，内符合精度上都能够满足坐标转换的规范要求，综合比较下这四组数据可以看出在最佳 k 值处得到结果最优。因此通过 K-Means 算法在最佳 k 值处所选取出的重合点，能够应用七参数的求解中。

表 1　七参数与精度评定表

Table 1　Seven parameters and accuracy evaluation table

	指定七参数	k=5	k=6	k=7	k=8
X 方向平移参数（m）	121.624	−0.00569	0.00182	−0.00377	−0.00149
Y 方向平移参数（m）	155.887	−0.00845	−0.00512	−0.00559	0.00000
Z 方向平移参数（m）	31.898	−0.02250	0.00215	−0.01615	−0.01694
X 方向旋转参数（"）	0.186	−0.02386	−0.00080	−0.01406	−0.01827
Y 方向旋转参数（"）	−0.067	0.04743	−0.00981	0.03693	0.03529
Z 方向旋转参数（"）	0.171	−0.01094	−0.01801	−0.00950	−0.00427
尺度参数（ppm）	17.579	0.14708	0.06274	0.10914	0.04841
内符合精度（m）	——	0.0002547	0.0002650	0.0002263	0.0003078

5　结束语

重合点的选取首先要自身精度达到一定要求，还需要根据重合点的均匀分布和具有一定的密度的原则来选择出最终进行计算七参数的重合点。在重合点自身精度达到一定要求下，通过 K-Means 聚类算法对重合点进行聚类，根据所得到 k 个簇即可选择出 k 个重合点，符合了重合点应具有一定的密度这一原则；根据聚类的最终结果，计算出各簇点与该簇聚类中心点的欧式距离最近的点作

为最终选取出的重合点，这样选择的重合点满足了均匀分布的原则。因所得到的重合点是与各簇的聚类中心最近的一个点，相当于是每一簇的伪聚类中心点。

文中之所以要先通过手肘法和轮廓系数法事先对坐标点进行聚类分析出最佳 k 值，是为了验证在最佳 k 值处得到的结果是最优的。不同的初始聚类中心会促使聚类的结果不一致，一方面是为了让程序能够多次运行且聚类结果不发生变化，另一方面则是为了加快 K-Means 收敛的速度，即加快得到最终的聚类结果。通过 Python 与 K-Means 算法实现快速、自动地选取出重合点，并以 Bursa 解算出七参数和验证了内符合精度，结果表明利用 K-Means 能够快速、自动地选取出较合适的重合点，求解出的七参数和内符合精度都比较好，并且得到在最佳 k 值处所得到的结果是最优的。因此在最佳 k 值处选取出的重合点，能够应用在重合点的选取中，但进行坐标转换的结果不一定是最好的。

K-Means 算法在重合点的选取上会因为一些离群点而导致聚类的效果较差，从而在自动选取重合点中造成一定的影响，因此接下来可以在此基础上研究如何过滤掉一些不必要的离群点后，再使用此方法进行重合点的选取，但需要注意有些坐标点是不一定能够去掉。

参考文献

[1] 王玉成，胡伍生．坐标转换中公共点选取对于转换精度的影响［J］．现代测绘，2008，31（5）：13-15.

[2] 王晓南，王庆宾，苏玉瑞．局域地区坐标转换七参数的求解及精度分析［J］．测绘与空间地理信息，2017，40（12）：158-160.

[3] 焦立芬．基于坐标转换重合点的分布、密度、精度与转换精度分析［J］．测绘技术装备，2013，15（4）：25-28.

[4] 周跃寅，潘国荣．公共点分布对坐标转换精度的影响［J］．大地测量与地球动力学，2013，33（2）：105-109.

[5] 赵宝锋，张雪，蒋廷臣．坐标转换模型及公共点选取对转换成果精度的影响［J］．淮海工学院学报（自然科学版），2009，18（4）：54-56.

[6] 杨俊闯，赵超．K-Means 聚类算法研究综述［J］．计算机工程与应用，

https://doi.org/10.35534/go.0301002

2019，55（23）：7-14+63.

［7］张敬伟．布尔莎模型坐标转换适用范围及精度分析［J］．测绘与空间地理信息，2013，36（1）：175-176+182.

［8］Rezaee M R，Lelieveldt B P F，Reiber J H C．A new cluster validity index for the fuzzy c-mean［J］．Pattern Recognition Letters，1998，19（3/4）：237-246.

［9］朱连江，马炳先，赵学泉．基于轮廓系数的聚类有效性分析［J］．计算机应用，2010，30（S2）：139-141+198.

［10］Celebi M E，Kingravi H A，Vela P A．A comparative study of efficient initialization methods for the k-means cl- ustering algorithm［J］．Expert Systems with Applications，2013，40（1）：200-210.

［11］Tanir D，Nuriyeva F．On selecting the initial cluster centers in the K-means algorithm［C］//2017 IEEE 11th International Conference on Application of Information and Communication Technologies，Moscow，2017：1-5.

［12］程鹏飞，成英燕，秘金钟，等．大地测量控制点坐标转换技术规范：CH/T 2014-2016．国家测绘地理信息局，2017．3：4-6［2020-11-20］．http://www.nrsis.org.cn/portal/stdDetail/211741.

［13］李潇，尹晖．基于最小二乘配置的三维空间坐标转换［J］．测绘工程，2008，17（2）：16-18，29.

A K-Means Algorithm for Automatic Selection of Public Points for Coordinate Conversion

Liu Guodong　Qin Hao　Liu Jia　Liu Lang

Chongqing Jiaotong University, Chongqing

Abstract: Background: In coordinate conversion, when there are multiple coincident points with known data, the selection of coincident points is one of the key issues. Objective: A method based on the K-Means algorithm to automatically select a suitable coincidence point among many coincidence points. Method: By analyzing the basic principles of the K-Means clustering algorithm and determining the optimal k value, an initial clustering center selection method is proposed for spatial rectangular coordinate data, and combined with Python to achieve rapid selection from multiple coincident points, the coincidence point is automatically selected, and the Bursa model is used to calculate the seven parameters and the internal compliance accuracy meets the requirements. Result: The K-Means clustering algorithm can quickly and automatically select a more suitable coincidence point, and the internal coincidence accuracy is the best at the best k value. The seven parameters solved by the determined coincidence point can be used in the coordinates Converting. Conclusion: The study shows that the method of automatically selecting coincident points based on the K-Means algorithm is not only feasible but also able to meet the accuracy requirements of coordinate conversion, which is of great significance to coordinate conversion work.

Key words: Coincidence point selection; Coordinate conversion; K-Means; Bursa; Python

https://doi.org/10.35534/go.0301002