引文格式：何春秀, 万远. 多源在线地理编码融合优化与应用[J]. 地理空间信息, 2023, 21(5): 45-47. doi:10. 3969/j.issn. 1672-4623.2023.05.011

多源在线地理编码融合优化与应用

何春秀[1]，万　远[1*]
（1. 湖北师范大学 城市与环境学院，湖北 黄石 435000）

摘　要：不同地图服务商提供的地理编码规则和数据源不同，导致地理编码结果也有所差异。目前所有在线地理编码服务均提供JSON文本数据作为返回数据格式，但JSON文件并不是矢量文件格式，无法直接转换成矢量数据文件，给实际工作带来了不便。对比分析了高德地图、百度地图、天地图等地理编码服务接口的结果差异，针对该差异设计了一种地理编码结果优化算法，并利用ArcPy将最终优化的编码结果转换为矢量文件，实现了多种地理编码服务的整合优化和编码结果的直接矢量化，减小了地理编码差异带来的结果偏差，提高了地理编码工作效率。

关键词：地理编码；ArcPy；编码优化；编码融合
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2023）05-0045-03

Optimization and Application of Online Multi-source Geocoding Fusion

HE Chunxiu[1],WAN Yuan[1]
(1. School of City and Environment, Hubei Normal University, Huangshi 435000, China)

Abstract: However, due to the different geocoding rules and data sources provided by different map service providers, the results of geocoding will be different. At present, all online geocoding services provide JSON text data as the return data format, but JSON file is not a vector file format and cannot be directly converted into vector data file, which brings inconvenience to the actual work. In this paper,we compared the results of geocoding service interfaces such as Gaode map, Baidu map and MapWorld,designed an optimization algorithm for geocoding results according to the differences of geocoding results, and used ArcPy to convert the final optimized coding results into vector files, which could realize the integration and optimization of a variety of geocoding services, reduce the result deviation caused by the difference of geocoding, realize the direct vectorization of coding results, and improve the efficiency of geocoding.

Key words: geocoding, ArcPy, coding optimization, coding fusion

研究表明，日常生活中有80%的信息都与地理空间位置密切相关，但这些信息基本没有空间坐标属性[1]。同时，许多行业都需将地名地址的自然语言描述信息转换为相应的空间位置信息[2]。地理编码技术能将非空间化的地名地址描述信息转换为具体的空间属性信息，从而实现行业专题信息与空间数据的融合与共享[3]，是一种基于空间定位技术的编码方法[4]，主要通过地址标准化、地址匹配和地址量化3个步骤实现，但需要依赖相应的地址模型、地址词典、地名地址数据库和量化规则[5]，实现过程较繁杂。因此，在实际生产中通常采用在线地理编码服务，可省去地名地址数据库等步骤实现地理编码。目前百度地图、高德地图、腾讯地图和天地图等互联网地图服务商均提供地理编码服务接口，但各接口采用的数据源和坐标系统不一致，导致用户使用过程的不便，且接口返回结果还需进一步处理才能转换为通用的矢量文件格式。鉴于此，本文利用Python语言封装了针对不同地图的地理编码请求模块，并对不同接口返回的结果坐标进行转换，统一到WGS84坐标系下进行结果对比和优化，最终利用ArcPy二次开发的Python库对优化坐标结果进行坐标数据矢量化，即将JSON点数据转换为Shapefile矢量文件。

1　地理编码方法

1.1　国内外研究进展

地理编码是将自然语言描述的地址位置信息通过既定的地址模型或编码规则与空间位置相关联，从而确定其代表的地理实体的位置[6]。许多学者对地理编码技术进行了研究，如廖薇薇[7]等利用系统聚类和随机森林法对多源在线地理编码服务结果进行优化，改

收稿日期：2022-01-06；修回日期：2022-05-12。
第一作者简介：何春秀（1994—），硕士研究生，研究方向为地图制图及遥感影像分类研究，E-mail：2412308653@qq.com。
通信作者：万远（1984—），博士，副教授，E-mail：42414120@qq.com。

善了编码精度，提高了海量社会经济信息数据转换为空间信息数据的效率；程钢[8]等通过对比国内主流在线地图API，提出了优化对策；张剑[9]等利用精细化地名地址库、地名地址特征分词库以及基于用户搜索行为的大数据分析来提高地址匹配度；陈志云[10]等利用大数据融合技术，将POI点与旅游用户行为大数据进行融合，进一步分析了梅州市旅游文化特点，优化了旅游资源发展格局；田沁[11]等对比分析了主流地理编码服务的质量情况。

1.2 地理编码接口服务

百度地图、高德地图、天地图的在线地理编码服务是通过HTTP/HTTPS协议访问远程服务的接口，提供结构化地址与经纬度之间的相互转化能力，使用时只需按照对应的请求规则发送需要编码的地址描述信息，请求结束后再解析编码结果获取对应的坐标等信息。由于各地理编码服务接口返回的坐标系不一致，因此需将坐标统一转换到WGS84坐标，本文利用高斯反算公式进行坐标转换，得到的坐标精度不低于0.000 1″[12]。实验过程中为了更好地对比编码结果，需对数据进行投影处理，将数据坐标单位由经纬度换算为m。其中，高德地图地理编码返回的坐标数据采用GCJ-02坐标系（火星坐标系），是由中国国家测绘局制订的地理信息系统坐标系统，是WGS84坐标系加密后的坐标系；百度地图地理编码返回的坐标数据采用BD09坐标系（百度坐标系），是在GCJ-02坐标系基础上再次加密得到；天地图地理编码返回的坐标数据采用WGS84坐标系。

2 模型设计

2.1 处理流程

结合ArcPy进行二次开发，整合了3种地理编码服务接口，其中在线地理编码服务接口是一种简单的HTTP接口，其编码效率可达0.2 s/条[7]。根据不同在线编码服务接口规范，对需要编码的数据进行地理编码转换，对多个地理编码结果进行优化对比分析，并获得最优的地理编码结果，此时该数据仍为JSON格式，需将数据由JSON格式转为Shapefile矢量数据格式。

2.2 优化算法

相对算数平均数而言，几何平均数受奇异值的影响较小[13]，因此优化算法优先选择几何平均数进行计算。几何平均数是对各变量值的连乘积开项数次方根，然后在计算的基础上增加阈值判断。阈值判断是在编码结果出现异常值时做出阈值筛选。

同一个地址描述信息经过在线地理编码服务接口处理后，将返回具体的编码结果坐标值，这里将3个编码结果分别记为 A、B、C 三个空间坐标点，若某点与其他两点的长度大于阈值，则该点为异常点，需剔除，不参与几何平均数的计算；若3个点都为异常点，则全部保留参与计算。如图1所示，当 LA 和 LB 的长度大于阈值 LP 时，则 C 点不参与计算，最终结果则是 A、B 两点的几何平均数；否则是 A、B、C 三点的几何平均数，计算结果更接近真实结果，阈值 LP 则需根据大量数据验证确定。

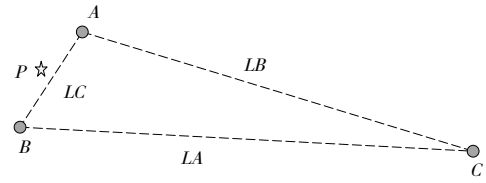


图1 编码结果选取示意图

几何平均数的计算公式为：

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n} \tag{1}$$

最终优化结果 P 点的坐标为：

$$\begin{cases} P_X = \sqrt[3]{A_X B_X C_X} \\ P_Y = \sqrt[3]{A_Y B_Y C_Y} \end{cases} \tag{2}$$

式中，A_X 为 A 点的 X 坐标；A_Y 为 A 点的 Y 坐标；B、C 点类似。

2.3 阈值计算

部分POI点通过不同编码后的最大距离值见表1，可以看出，相同数据采用不同在线编码服务处理后的结果存在较大差异，需通过阈值筛选才能最大程度地消除和避免存在的差异性。

表1 不同地理编码结果间的最大距离

名称	地址	市级名称	县区名称	距离/m
爱佳雅苑	迎宾大道46号	黄石市	黄石港区	248.10
奥山磁湖天下	湖滨大道76号	黄石市	黄石港区	74.16
奥体铭苑	北煞湖巷20号	黄石市	阳新县	32.46
白马山小区	白马路附近	黄石市	下陆区	75.05

利用不同的在线编码服务接口对黄石市多个POI数据进行编码处理，并计算编码后结果的距离，以最大距离作为该点的最大误差距离，所有点最大误差距离见图2，可以看出，最大误差距离值分布差异较大，但大多数在200 m以下，因此在判断异常值时，只需将阈值设为200 m即可完成对异常值的筛选过滤。

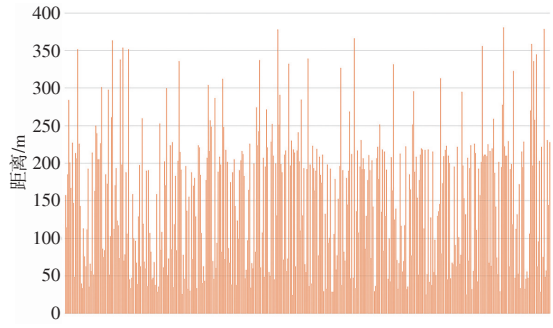


图2　不同编码结果点位距离差异分布图

2.4　程序模型设计

根据在线地理编码服务接口，结合实际工作生产需要，本文开发了基于ArcPy的全流程处理工具软件。ArcPy是ArcMap软件的二次开发程序框架，提供了地理数据分析、数据转换、数据管理等[14]实用高效的开发接口。程序模块设计的底层基础框架是基于ArcMap的ArcPy库，应用模块包括Shapefile格式转换模块、编码结果优化模块、HTTP网络请求模块和Excel数据读取模块，编码服务模块包括百度地图服务、国家天地图服务和高德地图服务。

3　实验与结果分析

3.1　实验数据与方法

本文以黄石市部分小区、社区的地名地址点为实验数据（表2），数据字段属性包括名称、地址、所在市、所在县，数据总量为998条。数据的地址描述内容并非标准的地址描述，而是包含“附近”、“东北侧”等自然语言描述信息，但由于每条数据都包含所在的市县，因此在进行地理编码时能快速缩小匹配范围。地理编码工具对融合字段后的数据进行多源在线编码处理，得到百度地图、高德地图、天地图3种编码结果，再利用优化算法对3种编码结果进行融合优化处理，从而得到最优编码结果。

表2　实验示例数据表

名称	地址	市级	县级
江北管理区灵官村社区	灵官路22号附近	黄石市	黄石港区
电线厂小区	龙港路10号	黄石市	阳新县
东屏巷和谐文明小区	黄厂街642号	黄石市	西塞山区
幸福家园临江街小区	黄石大道中窑小学东北侧180米	黄石市	西塞山区

3.2　结果分析

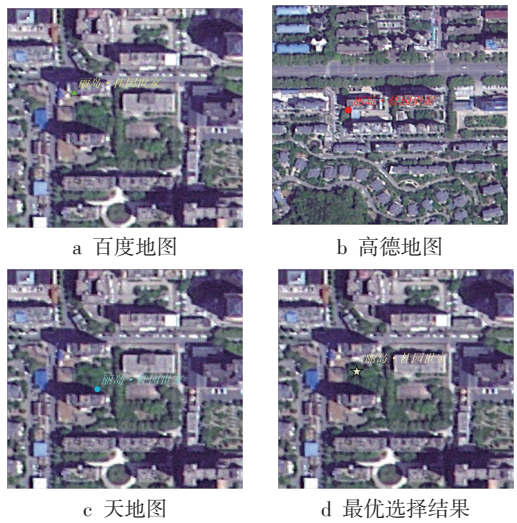
对百度地图、高德地图、天地图3种编码结果进行阈值对比分析，剔除异常值；然后求取剩下结果的几何平均数，作为优化结果；最后将3个地理编码结果与优化结果展示到卫星遥感影像地图上进行分析。

以丽岛·桂园世家小区的POI点为例，其地址描述为：湖北省黄石市下陆区杭州西路76号附12东南方向60 m，编码结果和算法择优选取的结果见表3，距离1、距离2为当前编码结果与另外两种编码结果的距离，在该点结果中，高德地图的编码结果与另外两种编码结果的差异较大，距离大于阈值200 m，因此高德地图的地理编码结果作为异常值不参与优化算法的计算。

表3　丽岛·桂园世家小区地理编码结果数据表

地理编码服务	编码经纬度	距离1/m	距离2/m
百度地图	115.0292, 30.2004	336.448	43.378
高德地图	115.0258, 30.2009	336.448	366.114
国家天地图	115.0295, 30.200	366.114	43.378

为了更好、更直观地对比编码结果，将具体编码结果叠加到高分辨率卫星遥感影像中，如图3所示，可以看出，百度地图的编码结果为该小区的旁边，经分析由于POI数据是沿路边采集，因此采集点落在小区的旁边，但不存在位置偏移；该小区高德地图的编码结果是错误的，点位并不在真实的小区内部；天地图的地理编码结果坐标点落在小区内部，编码结果正确；最优算法选取的结果位于小区内部，是通过百度地图和天地图的编码结果计算而得。



a 百度地图　b 高德地图　c 天地图　d 最优选择结果

图3　各种编码结果对比

4　结　语

本文研究了多种地理编码结果 （下转第116页）