

邓世广、王月、马亚伟, 2019, 基于 Python 爬虫技术的高精度地震目录更新程序的设计与应用, 中国地震, 35(3), 541~549.

基于 Python 爬虫技术的高精度地震目录 更新程序的设计与应用

邓世广 王月 马亚伟

中国地震台网中心, 北京 100045

摘要 首先, 基于 Python 爬虫技术, 实现了高精度定位结果的地震目录的自动下载; 然后, 针对省局台网快报目录汇总后存在重复定位地震的问题, 提出了一种能够有效剔除重复定位地震的方法; 最后, 分析 EQ3 格式地震目录的二进制结构, 实现了高精度定位结果的地震目录向 EQ3 格式的转换。实际应用表明, 该程序能够显著提高 EQ3 格式地震目录更新的工作效率, 更高定位精度的地震目录也有助于对地震活动性进行深入分析。

关键词: 网络爬虫 高精度地震目录 格式转换

[文章编号] 1001-4683(2019)03-0541-09 [中图分类号] P315 [文献标识码] A

0 引言

地震目录是地震危险性分析中的关键资料, 也是研究岩石圈动力学过程的重要数据, 在国民经济建设、地震分析预报、工程抗震、地震减灾等领域发挥着重要的作用(李圣强等, 2004; 王海涛等, 2006; 张卫华等, 2007; 何琳等, 2017)。地震目录又是一种时效性要求较高的数据, 特别是在强震发生后的应急工作中, 需要及时更新地震目录, 跟踪地震序列的发展演化。在日常的地震活动性分析工作中, 为了跟踪分析最新的地震活动情况, 也需要频繁地更新地震目录。准确、实时和完整地获取地震目录对地震科研人员以及防震减灾、城建规划等工作都至关重要(张卫华等, 2007)。

由于历史原因, 地震目录格式多种多样, 如 CapSeis 系统使用的 EQT 或 EQ2 格式、SSEPC 所用的 WKF 或 SHK 格式、地震速报目录、地震月报目录、Apnet 网络报送目录及其他特征格式等(李圣强等, 2004)。为了支持更高精度的地震定位结果, 同时包含地震参数数据和震源机制解数据, 李圣强等(2004)设计了 EQ3 格式的地震目录, 这种格式由 2 个相互关联的二进制结构来表示记录, 与 ASCII 码格式相比, 这种二进制结构的地震目录格式具有数据存储量少、存取访问速度快的优势。由于地震目录格式类型多样, 不同机构提供的地震目录格式不尽相同(马宝君等, 2005), 而不同地震分析系统软件支持的地震目录格式也各有不同, 因此, 不同地震目录格式之间的转换是地震活动性分析中经常遇到的问题。

[收稿日期] 2019-06-10; [修定日期] 2019-07-22

[项目类别] 国家重点研发计划课题“震例回溯研究”(2018YFC1503305)、国家重点研发计划课题“年尺度和短临强震发震紧迫程度判定技术研究”(2017YFC1500502)、中国地震局震情跟踪定向工作任务(2019010118)共同资助

[作者简介] 邓世广, 男, 1988 年生, 助理研究员。主要从事地震活动性研究。E-mail: dengsg@seis.ac.cn

为提高地震目录格式转换及更新的工作效率,前人做了较为丰富的工作。考虑到数据交换或不同地震系统软件分析处理的需要,李圣强等(2004)分析了应用较多的目录格式,并提供了格式转换工具,可以将多种格式的地震目录转换为 EQ3 格式,同时也支持 EQ3 格式地震目录转换为其他格式地震目录。马宝君等(2005)使用 VB 编制了地震目录快速转换软件,将网上检索的地震目录转换成 TXT、EQT、WKF 等格式目录。为了进一步提高地震目录更新的时效性,张卫华等(2007)开发了一种能从 FTP、HTTP 服务器上直接下载地震目录资料的软件,实现了地震目录的自动下载并转换为 EQT、TXT、WKF、CSV 等地震目录格式。由于 SuperSeis 提供的地震目录格式转换工具支持的地震目录格式有限,于洪波(2009)采用 Delphi 语言实现了 Excel 和 TXT 格式地震目录向 EQ3 格式的转换,有效提高了震后趋势会商的速度和准确性,但该方法没有考虑 EQ3 格式地震目录的辅助记录部分,因此未将发震地点写入 EQ3 地震目录内。梁阿全等(2016)应用 Excel VBA 实现了从 Excel、Q01、EQT 等格式至 WKF 格式之间的转换。何琳等(2017)通过对安徽地震编目数据库的访问,实现了地震目录查询以及数据提取、转换和导出等功能,减轻了数据维护人员的工作量,提高了数据的即时性和准确性,能够有效满足地震应急的迫切要求。由于 EQT、TXT、WKF、CSV 等地震目录格式属于 ASCII 码格式,读写相对简单,因此针对这些地震目录的格式转换方法研究相对较多。EQ3 地震目录格式属于二进制结构,具有数据存储量少、存取访问速度快的优势,但这种结构读写方式相对复杂,目前还没有针对 EQ3 格式地震目录的自动更新方法。

随着地震监测定位水平的不断提高,中国地震编目网站提供的地震目录中经纬度的精度可达到小数点后 3 位,更高精度的地震经纬度能够更加清晰地刻画地震分布、揭示断裂构造形态,有助于地震活动性的深入分析。目前常用的 EQT、WKF、CSV 等 ASCII 码格式的地震目录,一般仅支持经纬度精确到小数点后 2 位,难以满足更加精细的地震活动性分析工作。而 EQ3 格式的地震目录属于二进制结构,震中经纬度参数采用 Float 类型,其参数值精度保持在 7 位有效数值的精度范围内,经纬度的有效精度可以表示到小数点后 4 位,即相当于 10m 级(李闽峰等,2006)。因此, EQ3 格式地震目录是一种能够适应更加精细地震活动性分析工作的地震目录格式,具有更好的应用前景。然而, EQ3 格式地震目录虽然能够支持较高的地震定位精度,但目前尚没有针对高精度定位结果的地震目录的格式转换方法,高精度的地震目录还未能在地震活动性分析工作中充分发挥作用。

为实现高定位精度 EQ3 格式地震目录的自动更新,设计并实现了地震目录自动更新程序。该程序首先基于 Python 爬虫技术实现地震目录的自动下载;然后针对地震编目系统中多个省局的快报地震目录中存在重复定位地震的情况,提出了一种筛除重复定位地震的方法;最后研究分析 EQ3 格式地震目录的二进制结构,在此基础上实现了高精度定位结果的地震目录的格式转换、目录拼接等功能。实际应用表明,该程序能够完成高精度定位结果的地震目录的自动更新,提高了地震目录更新的工作效率,能够满足地震应急对地震目录的高实效性要求。

1 技术方法

为实现高精度定位结果的地震目录的自动下载,首先调用预设的地震目录下载相关参数,利用 Python 爬虫技术从地震编目网站自动下载指定的地震目录,针对邻省快报目录存在

重复定位地震的问题,采用本文提出的重复定位地震筛除方法对地震目录内的重复定位地震进行筛除,保证地震目录的唯一性,最后分析 EQ3 格式地震目录的数据结构,实现地震目录向 EQ3、EQT 等格式的转换,整体技术流程如图 1 所示。需要说明的是,地震编目网站提供的正式地震目录的更新时间是滞后的,而快报目录的实时性较好。为了提高地震目录的完整性,部分地震目录使用最近一段时间的快报目录来补充正式目录的不足。为了在目录更新时将快报目录替换为正式目录,下载地震目录的起始时间应早于待更新目录内省快报地震目录内最早的发震时刻。另外,为了保证地震目录更新时不遗漏地震事件,一般也将下载地震目录的起始时间设置早于地震目录的最近 1 次地震发震时刻。在地震目录更新时,程序将自动删除待更新地震目录内下载起始时间之后的地震目录,再将其与下载的最新地震目录进行拼接。

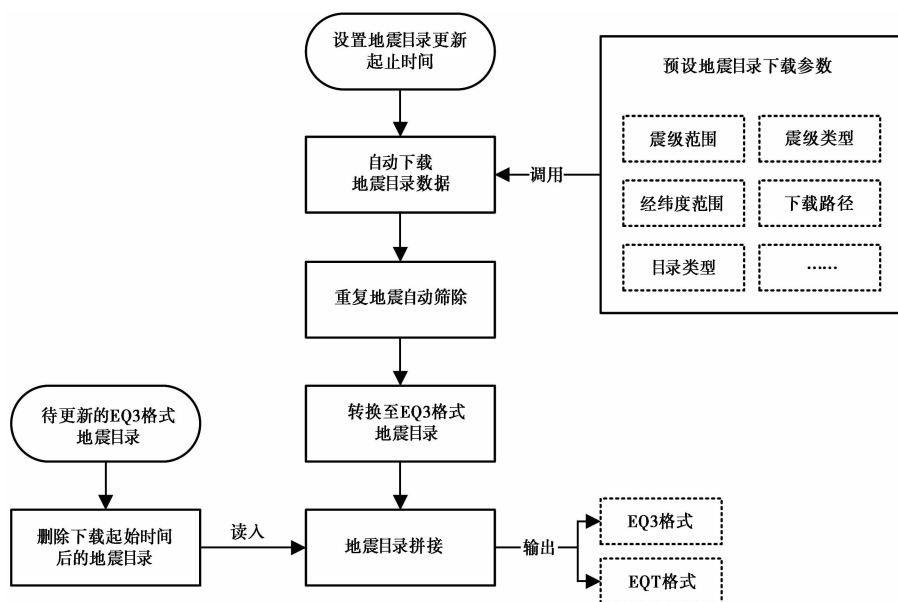


图 1 地震目录自动更新技术流程

1.1 基于 Python 爬虫技术的高精度定位结果的地震目录自动下载

随着互联网的快速发展,爬虫技术已经成为高效准确地获取数据的重要方式之一,在各个领域都有着广泛的应用(周中华等,2014;熊畅,2017;余豪士等,2018)。Python 中的第三方 HTTP 库——Requests 库被爬虫开发者广泛应用(余豪士等,2018)。Requests 集成了定制请求头部、发送请求、传递 URL 参数、获取相应内容等多种函数(谢克武,2017)。通过 Requests 库,能够方便实现网络数据的爬取。

目前,我国地震活动性分析使用的地震目录主要通过中国地震台网中心的地震编目系统网站下载获取。地震编目系统网站主要通过 GET 方式发送请求、传递参数。Requests 集成了多种 HTTP 协议中的请求方式,包括最普遍的 GET 和 POST 请求。因此,利用 Requests 库可以方便地访问地震编目网站并获取指定的地震目录数据。

地震编目系统网站的用户登陆通过 GET 方式完成,实现过程如下:

```
s=requests.Session()  
login_params={'tb000.t1':username,'tb000.t3':password}  
s.get("http://10.5.160.18/console/login.action",params=login_params)
```

登陆之后即可继续通过 GET 方式传递下载地震目录的参数,从而获取地震目录。下载地震目录的参数涉及地震目录类型、地震目录格式、震级类型、起止时间、震级范围、经纬度范围等,可以通过字典的形式构建 1 个参数字典,随后通过 GET 方式获取下载参数规定范围内的地震目录,实现过程如下:

```
DownloadParams={'key1':'value1','key2':'value2','key3':'value3',...}  
response=s.get("http://10.5.160.18/export/download.action",params=DownloadParams)
```

值得注意的是,在地震编目系统网站提供的地震目录格式中,通用格式文本记录的地震经纬度精度能够达到小数点后 3 位,定位精度高于其他格式。如需下载高精度定位结果的地震目录,应将地震目录格式参数设置为通用格式文本。

1.2 重复定位地震筛除

根据《测震台网运行管理细则》^①,地震编目工作产出的数据包括快报和正式报目录。省级台网中心应在每日 12 时前报送前 1 日本省及周边地区的省局快报目录,在每周日上午 12 时以前报送上周的省局正式目录。国家测震台网中心收集有人值守的国家测震台站和省级台网中心处理的地震目录和观测报告,并进行全国统一编目,保证目录和观测报告的唯一性。每日 24 时前产出前 1 日的全国快报目录,每周编辑国内及全球地震目录和观测报告,于第 3 周的周日 24 时前发布全国正式目录和观测报告。由此可见,省局快报地震目录的实时性最高,且在实际工作中,省级台网中心一般会及时发布最近的地震事件至省局快报目录内,因此省局快报的滞后时间一般仅在数小时内。而其他类型的地震目录则至少会滞后 1 天的时间,难以满足地震活动性分析的实时性要求。

为了跟踪最新的地震活动,在地震目录更新时,需要将省局快报目录进行汇总来补充正式目录尚未更新的部分。然而,由于各省级台网中心会记录本省周边地区的地震事件,因此,邻省之间的地震目录会存在重复定位地震。由于直接获得各台站的实时波形数据存在困难,我们仅能根据地震目录内的信息来识别重复定位地震并予以剔除,从而保证地震目录的唯一性。传统的工作方式需要通过人工判断,如果 2 个地震的发震时刻、震中位置以及震级均相近,且属于不同省局上传的地震事件,则认为 2 个地震为重复定位地震并删除。这种方式需要人工对每个地震事件逐一判断,操作繁琐且易出错。本文基于传统人工识别重复定位地震的经验,建立重复定位地震的识别准则,通过计算机高效地自动识别并筛除重复定位地震,技术流程如图 2 所示。基本的重复定位地震识别准则为:

- (1) 发震时刻小于给定阈值;
- (2) 震中距小于给定阈值;
- (3) 震级差小于给定阈值。

如果 2 次地震同时满足以上 3 点标准,即认为是重复定位地震。阈值需要根据实际地震监测水平设定。例如,当中国台湾发生较强地震时,多个省局会监测并记录该地震,但目

^① 中国地震局监测预报司,2015,中震测函〔2015〕115 号测震台网运行管理细则(2015 修订版)。

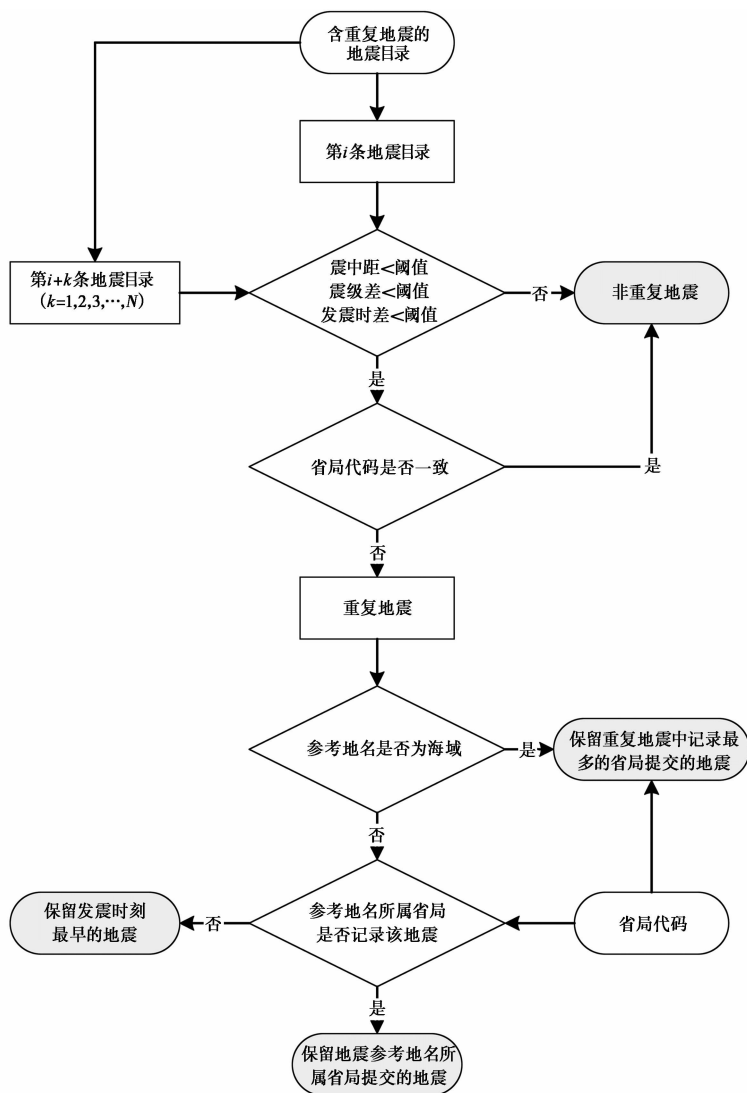


图 2 重复定位地震识别方法技术流程

前我国大陆对中国台湾地区的监测能力相对较弱,不同省局确定的地震震级、发震时刻以及发震位置偏差较大,因此,对于这种情况应给予较宽松的阈值。需注意的是,同一省局的快报目录内不存在重复定位地震,因此,对于同一省局的快报目录内的 2 次地震,即使满足以上标准,也不认为是重复定位地震。另外,针对以下特殊情况给出了相应的解决方法:

(1) 重复出现的地震未被本省记录而被其他多个省局记录,则保留发震时刻最早的记录。

(2) 重复出现的地震发生区域(如海域)没有明确省局归属,则保留记录到该区域地震数最多的省快报目录,如果各省快报记录的地震数相同,则保留发震时刻最早的记录。

1.3 EQ3 地震目录格式转换

为将下载得到的最新地震目录转换至 EQ3 格式,我们对 EQ3 格式地震目录的二进制结

构进行了分析研究。李闽峰等(2006)指出,EQ3 格式地震目录由 2 个相互关联的二进制结构来表示记录,这能够有效地控制数据冗余,并为应用程序节省空间资源提供了有效的机制。因此,完整的 EQ3 格式地震目录包含 EQ3 和 EQB 2 个文件,分别记录了主要信息和辅助参数(参考地名和地震震源机制解参数)。

EQ3 主记录结构定义如下:

```
typedef struct {
    long Date;           //地震日期(年 * 10000+月 * 100+日)
    long Time;           //地震时间(时 * 1000000+分 * 10000+秒 * 100)
    float Lat;           //震中纬度(单位:°)
    float Lon;           //震中经度(单位:°)
    char  $M_s$ ;          //震级( $M_s * 10$ )
    char  $M_L$ ;           //震级( $M_L * 10$ )
    char  $M_b$ ;           //震级( $M_b * 10$ )
    char  $M_w$ ;           //震级( $M_w * 10$ )
    float Depth;         //震源深度(单位:km)
    long Sequence;       //地震序列编号
    long Index;          //对应辅助记录位置的偏离记录数
} EQ3AUX; //EQB 格式
```

根据 EQ3 格式的结构定义可知,震中经纬度采用 Float 类型,其参数值精度保持在 7 位有效数值的精度范围内,这样经纬度的有效精度就可以表示到小数点后 4 位,即相当于 10m 级(李闽峰等,2006)。因此,EQ3 格式地震目录是一种能够有效支持高精度定位结果的地震目录格式。EQ3 格式中每个地震记录主记录结构的最后部分记录的信息为辅助记录索引 Index,它表示对应于该地震记录的辅助参数在辅助记录文件中的顺序号(李闽峰等,2006)。目前 EQ3 格式地震目录采用的辅助记录定义为 EQB 格式,参考李闽峰等(2006)给出的 EQA 格式的二进制结构,EQB 格式的辅助记录结构定义如下:

```
typedef struct {
    char Location[32];   //地震参考地名(16 个汉字)
    float StrikeI;       //节面 I 走向
    float DipI;          //节面 I 倾角
    float StrikeII;      //节面 II 走向
    float DipII;         //节面 II 倾角
    float Pazim;         //P 轴方位角
    float Pdip;          //P 轴倾角
    float Tazim;         //T 轴方位角
    float Tdip;          //T 轴倾角
    float RakeI;         //节面 I 滑动角
    float RakeII;        //节面 II 滑动角
} EQ3AUX; //EQA 格式
```



这种相互关联的结构,使得 EQ3 格式既有效地描述了较多的地震参数,同时也将数据冗余降到最低。

下载使用的地震目录资料主要包括地震的发震时刻、震中经纬度、震源深度、震级及参考地名信息。地震编目系统提供的高精度定位结果的地震目录的经纬度信息精确到了小数点后 3 位,EQ3 格式地震目录能够有效支持高精度的定位结果。在格式转换过程中,应将参考地名信息存储于 EQB 格式的辅助记录内,其他信息写入 EQ3 格式的主记录内,同时根据辅助信息所在文件的位置,将相应的 Index 顺序号写入主记录。如果没有对应的辅助参数,应将 Index 赋值为-1。

2 应用效果

使用 Python 的图形工具包 PyQt 编写了程序的可视化界面(图 3),并且预设了常用地震目录的下载参数。设置好待更新的地震目录路径后,更新地震目录时只需选择目录类型以及起止时间即可自动完成地震目录的更新工作,无需其他人工操作。为验证所提方法的有效性,以全国快报地震目录更新一周时间内的地震为例,分别测试人工手动和计算机自动 2 种更新方式的用时。表 1 对比了人工手动和计算机自动更新全国快报地震目录所涉及的操作,可以看出人工手动更新方式的操作步骤较为繁琐。表 2 列出了同一测试人员采用不同更新方式在多次测试中使用的时间。从对比结果可以看出,本文所提的地震目录自动更新方法能够有效提高地震目录更新的工作效率,且耗时稳定,受人为因素影响较小,而人工手动更新方式的耗时受人为因素影响较大。另外,由于操作人员的熟练程度不同,更新地震目录的耗时也会有较大差异。

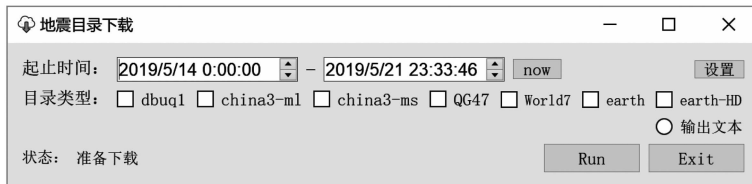


图 3 地震目录更新程序界面

表 1 人工手动更新和计算机自动更新全国快报目录的操作对比

序号	人工手动更新	计算机自动更新
1	登录地震编目系统	设置下载起止时间
2	设置参数下载正式快报目录	选择待更新的地震目录类型
3	设置参数下载省快报目录	运行
4	手动筛选省快报目录内重复定位的地震	
5	正式目录与快报目录拼接	
6	文本格式转换为 EQ3 格式	
7	删除待更新地震目录下载起始时间后的地震目录	
8	导入下载的地震目录	
9	输出 EQ3 格式目录	

本文所提方法支持高精度 EQ3 格式地震目录的自动更新,以全国统一正式报地震目录为例,对比传统地震目录与高精度定位结果的地震目录的定位结果。分别利用 2 种地震目录选取了九寨沟余震区 2018 年 1~12 月的 $M_L>0.0$ 的地震,绘制了相应的震中分布图(图 4)。从对比结果可以看出,通过高精度定位结果的地震目录绘制的震中分布(图 4(b))更符合地质构造背景。特别是在研究小区域尺度地震活动性时,如果地震定位精度过低,震中分布图上会呈现较明显的网格状分布,更高精度的地震目录能够刻画更加细致的地震活动分布,揭示断裂构造形态,有利于分析认识区域地震活动特征。

表 2 全国快报地震目录更新对比测试结果

序号	耗时/s		耗时比 (手动/自动)
	手动更新	自动更新	
1	339	14	24.21
2	327	11	29.73
3	297	11	27.00
4	372	13	28.62
5	321	9	35.67
6	357	12	29.75
7	341	10	34.10
平均	336.29	11.43	29.43

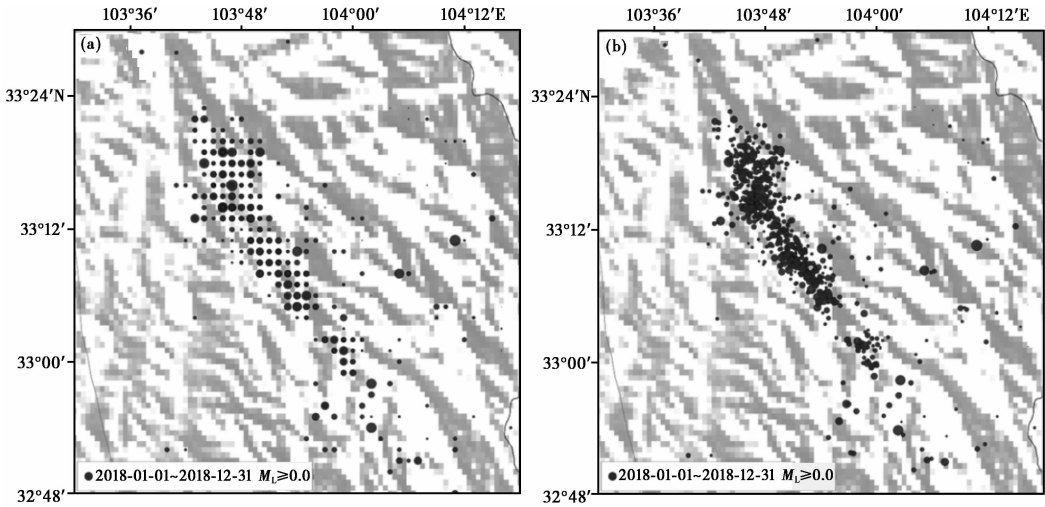


图 4 九寨沟 7.0 级地震余震区 2018 年 $M_L>0.0$ 地震分布对比
(a)使用传统地震目录;(b)使用高精度定位结果的地震目录

3 结语

本文设计了一种针对高定位精度 EQ3 格式地震目录自动更新的程序,该程序利用 Python 爬虫技术,从地震编目网站自动下载高精度定位结果的地震目录,并转换至 EQ3 格式。针对相邻省局快报地震目录内存在重复定位地震的问题,提出了一种自动筛选重复定位地震的方法,并应用于地震目录的自动更新中。采用 PyQt 编写了程序的可视化界面,同时预设地震目录下载参数,使该程序具有良好的实用性。实际应用表明,该程序能够有效提高地震目录的更新效率,相比传统的人工手动更新方式,能够提高近 30 倍的工作效率。

利用高精度的地震目录能够更细致地刻画地震分布、揭示断裂构造形态,有助于对地震活动性进行深入分析。目前,高精度的地震目录尚未在地震活动性分析工作中充分发挥作

用,本文提出的高精度定位结果的地震目录自动更新程序将有助于将高精度定位结果的地震目录进一步应用于地震活动性分析中。

致谢: 审稿专家对本文提出了宝贵的建议,中国地震台网中心孟令媛副研究员为本项工作及论文撰写给予了指导并进行了有益的讨论,薛艳副研究员对本项工作提供了有益的建议,本文部分图件采用 GMT 绘制,在此一并表示感谢。

参考文献

- 何琳、马奔、姜琪等,2017,地震目录更新软件设计,中国科技信息,(24),61~62.
- 李闽峰、李圣强、王斌,2006,EQ3 地震目录格式及其在 SQL Server 数据库中的实现方法,地震,26(2),133~138.
- 李圣强、李闽峰、王斌,2004,通用地震目录转换及打印程序的设计与实现,华北地震科学,22(2),50~53.
- 梁阿全、梁跃、马宝君等,2016,应用 EXCEL VBA 实现地震目录格式的转换,高原地震,28(4),59~63.
- 马宝君、马莉、崔东源等,2005,地震目录快速转换软件,地震地磁观测与研究,26(3),87~90.
- 王海涛、李莹甄、屠泓为,2006,新疆历史地震目录完整性分析,内陆地震,20(1),10~17.
- 谢克武,2017,大数据环境下基于 python 的网络爬虫技术,电子制作,(9),44~45.
- 熊畅,2017,基于 Python 爬虫技术的网页数据抓取与分析研究,数字技术与应用,(9),35~36.
- 于洪波,2009,厦门市地震局目录格式及转换,地震地磁观测与研究,30(6),114~118.
- 余豪士、匡芳君,2018,基于 Python 的反爬虫技术分析与应用,智能计算机与应用,8(4),112~115.
- 张卫华、李胜乐、廉超等,2007,地震目录资料自动搜索、下载软件的实现,地震,27(1),131~141.
- 周中华、张惠然、谢江,2014,基于 Python 的新浪微博数据爬虫,计算机应用,34(11),3131~3134.

Design and Application of High Resolution Earthquake Catalogue Auto Update Program Based on Python Crawler Technology

Deng Shiguang Wang Yue Ma Yawei

China Earthquake Networks Center, Beijing 100045, China

Abstract We first use the Python crawler technic to download the high resolution earthquake catalogue automatically. We proposed an effective repeat earthquake elimination method to remove the repeat earthquakes among different catalogues uploaded from different provincial earthquake bureaus. In the end, we transform the high resolution earthquake catalogue to EQ3 format based on the research of its binary structure. Practical application shows that the proposed program can improve the efficiency of the update of earthquake catalogue. In addition, high resolution earthquake catalogue is beneficial for us to analyze seismicity more deeply.

Key words: Web crawler; High resolution earthquake catalogue; Format transformation