**共享单车实时定位与移动趋势的大数据分析预测**

王宇洋

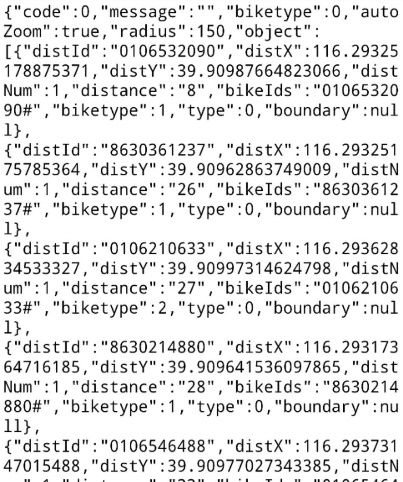
（北京育英学校）

**摘要：**近年来，很多城市的共享单车数量快速达到饱和状态，存量非常大且处于流动状态，在为市民提供骑行便利的同时，也对城市管理提出了更高的要求和挑战。如何及时掌握和分析城市中海量共享单车的数量、实时位置和移动趋势规律，智能预测未来的用车高峰地点与时段，对于做好共享单车的高效管理与便捷服务，具有很好的现实意义与应用价值，而且共享单车的骑行数据也是分析城市人群出行爱好与规律的最佳数据。在大数据环境下，采用Python语言编写网络爬虫程序，实现了对共享单车位置信息进行实时抓取，在GIS系统中对位置数据进行可视化展示和直观分析，从而了得出了一天之中不同地段、不同时段的共享单车移动规律和市民出行偏好趋势的研究结论。

**关键词：**网络爬虫；共享单车；大数据；GIS；Python

近年来，很多城市的共享单车数量快速达到饱和状态，存量非常大且处于流动状态，在为市民提供骑行便利的同时，也对城市管理提出了更高的要求和挑战。如何及时掌握和分析城市中海量共享单车的数量、实时位置和移动趋势规律，对于做好共享单车的高效管理与便捷服务，具有很好的现实意义与应用价值，而且共享单车的骑行数据也是分析城市人群出行爱好与规律的最佳数据。在大数据环境下，采用Python语言编写网络爬虫程序，实现了对共享单车位置信息进行实时抓取，通过GIS系统对骑行数据进行可视化的直观分析，从而了得出了一天之中不同地段、不同时段的共享单车移动规律和市民出行偏好趋势的研究结论。

**1共享单车位置数据的获取**

在手机上对微信摩拜小程序进行抓包，得到以下JSON数据。

于是，通过对本地POST发送数据分析，获得摩拜单单车GPS信息获取API接口。

此后利用Python，将API中发送数据中的latitude变量，和longitude分别写入两个嵌套循环，在北京（经度：116~116.8000 ; 纬度：39.6000~40.3000）范围内，以0.008作为步长对北京市六环以内进行快速扫描，从而得到共享单车GPS当前数据。

经过与手机APP（共享单车）的对比，可以确定抓取数据为有效数据，非反爬虫机制的模拟数据。

对于车辆类型，JSON(JavaScript Object Notation, JS 对象标记)数据中的biketype有1和2两种可能，经过与手机客户端的对比，1代表lite型车，2为老款车。

至此，数据获取部分结束。

**2数据处理**

得到抓取的原始数据后，还不能将其输入GIS系统进行分析。由于其中某些数据由于GPS波动造成车辆频繁移动或大幅度一定，因此需要对数据进行进一步筛选。

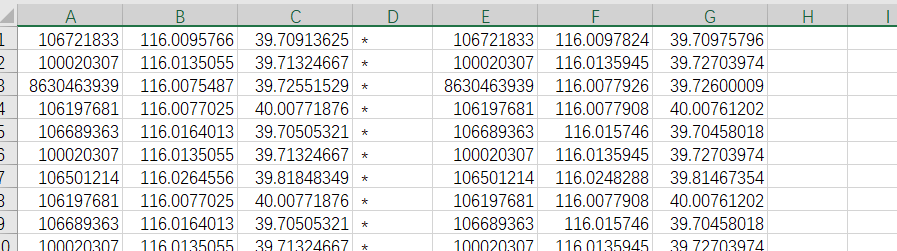
根据研究目的，选择以下筛选方法：

保留：车辆 ID相同且位置变化适中 的车辆

去除：车辆 ID不同或位置变化极大（小） 的车辆

根据所选研究的时间区段，移动经纬度超过0.4可能性不大，于是选择0.4作为数据的上限，由于车辆的小规模移动（GPS波动）同样可能，于是选择经纬度变化0.0002作为下限。

处理后，分别将同一ID自行车的前后位置存在excel表格中的同一列，如下图：



自左向右，分别为ID号，起点经度，起点纬度，ID号（同上），终点经度，终点纬度。

编程实现的核心代码如下:

for temline1 in range(1,line1):

for temline2 in range(1,line2):

if file1[temline1][id] == file2[temline2][id]:

if (file1[temline1][x] != file2[temline2][x] ) and (file1[temline1][y] != file2[temline2][y] ):

#print(file1[temline1][x])

#print(file2[temline2][x])

subx = abs(str2float(str(file1[temline1][x])) - str2float(str(file2[temline2][x])))

suby = abs(str2float(str(file1[temline1][y])) - str2float(str(file2[temline2][y])))

#print(str(subx) + str(suby))

#print(temline2)

if (subx > 0.0002 ) and (suby > 0.0002):

if (subx < 0.4) and (suby < 0.4 ):

print('subX : ' + str(subx))

print('subY : ' + str(suby))

print(file1[temline1])

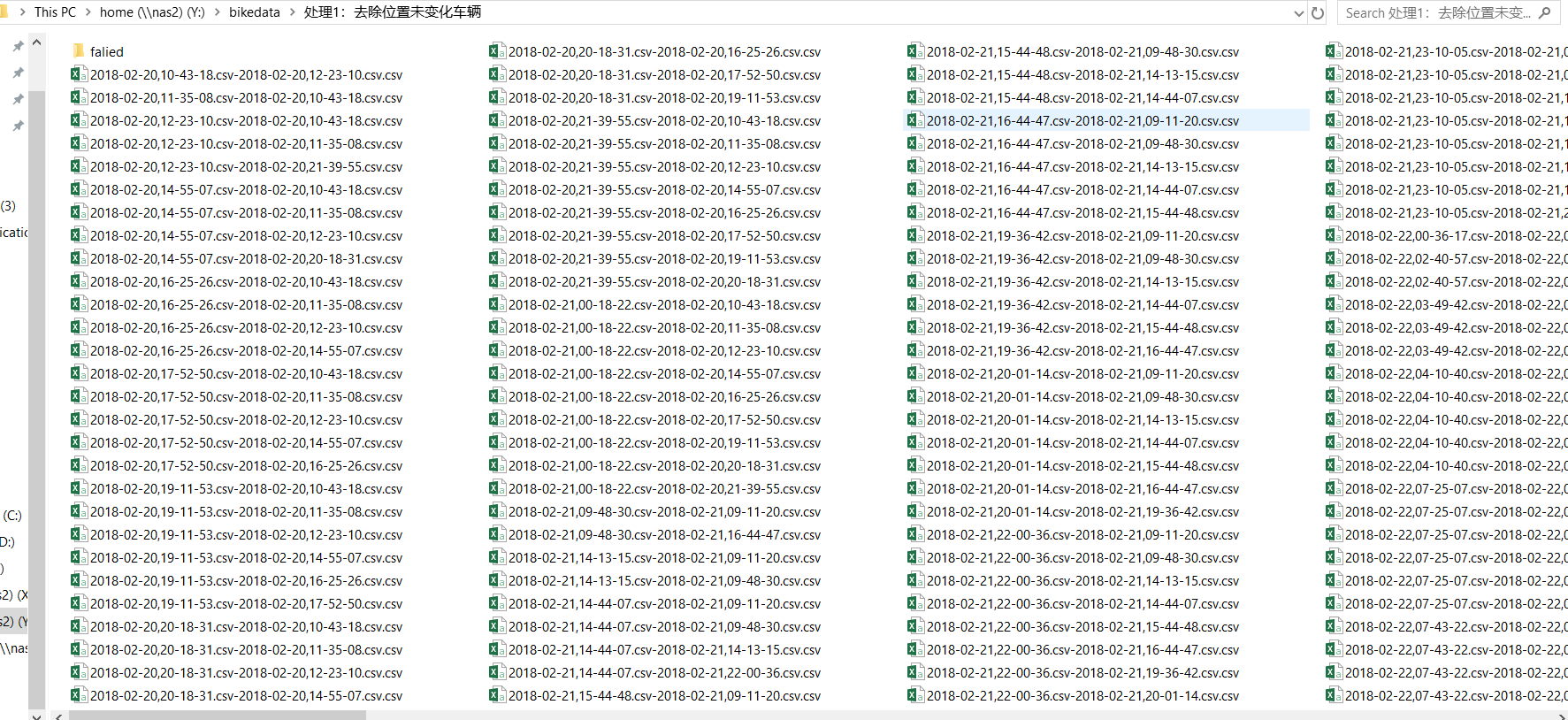
print(file2[temline2])

csv\_writer = csv.writer(out)

csv\_writer.writerow([file1[temline1][0],file1[temline1][1],file1[temline1][2] , " \* " , file2[temline2][0],file2[temline2][1],file2[temline2][2]] )

break

以下为处理后文件：

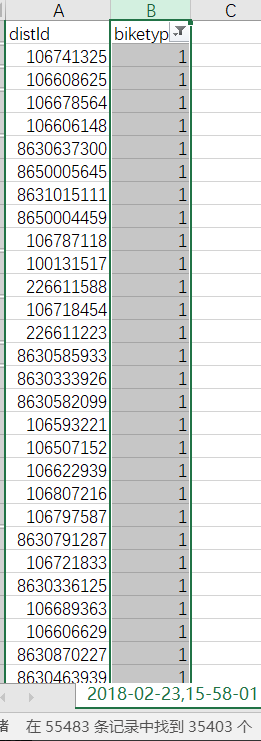
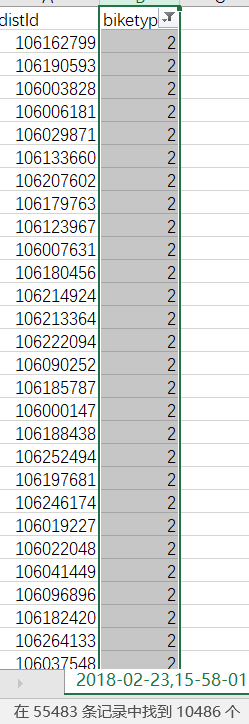
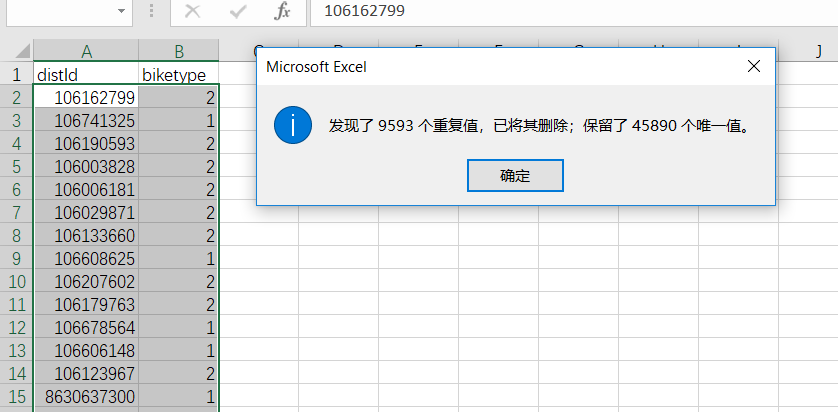


至此，数据处理部分结束。

**3数据分析**

**3.1车辆类型分析**

首先对ID进行去重处理：



除去重复项后，分别统计1和2类型的数量：

2类型包含10486个

1类型包含35403个

摩拜北京官方总投放量：37万辆

对北京市的总车辆类型估计如下：  
2类型约为: (69985辆

1类型约为: (236284辆

可以看出摩拜公司在车辆改进后，大量投放新车的现象，原旧版本车辆单车成本高达5000元而新车仅3000元，数据符合客观规律。

与时隔一年前的网上非官方数据分析产生了较大变化：

原文地址：https://www.jianshu.com/p/2a20d2a97ac0

**3.2车辆移动分析**

根据数据统计，有近4成车辆在研究时间内从未发生移动。对于此现象分析有两种可能：

* 1. 由于时间是冬天，骑行人较少
  2. 车辆处在小区内、偏僻位置或无法开始骑行（损坏，上锁）

**3.3移动趋势分析：**

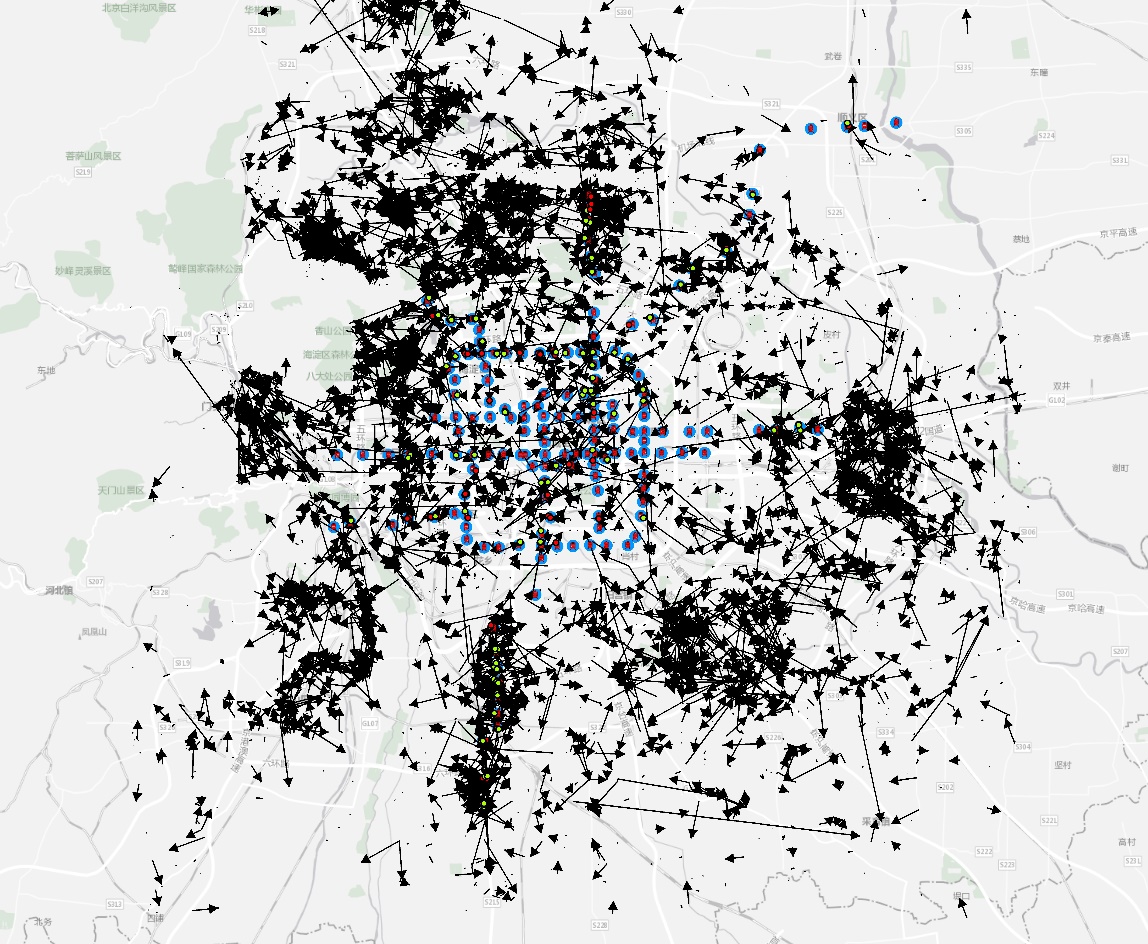
首先，通过百度地图开放平台获取北京市内主要地铁站的坐标，并将其输入GIS系统，并在其周围200米建立缓冲区，便于统计车辆移动信息。

之后，将处理后文件输入GIS系统，由起点向终点连线。同时将起点和终点坐标输入GIS系统，与缓冲区做相交处理，将起点设为红色，终点绿色。

以下为数据分析：

**3.4宏观分析**

根据总体数据，用车集中区不在北京市市中心附近，而是呈现围绕地铁的条带状，或是在郊区的聚集形态



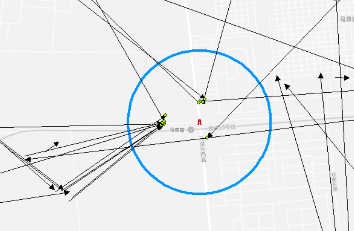
例如在地图的左下方的聚集区均围绕地铁，而位于地图左上的聚集区经百度地图查询，为三个产业园聚集处，青年人居多，故出现了聚集现象。

同时由图可见，中心城区骑行次数并不多

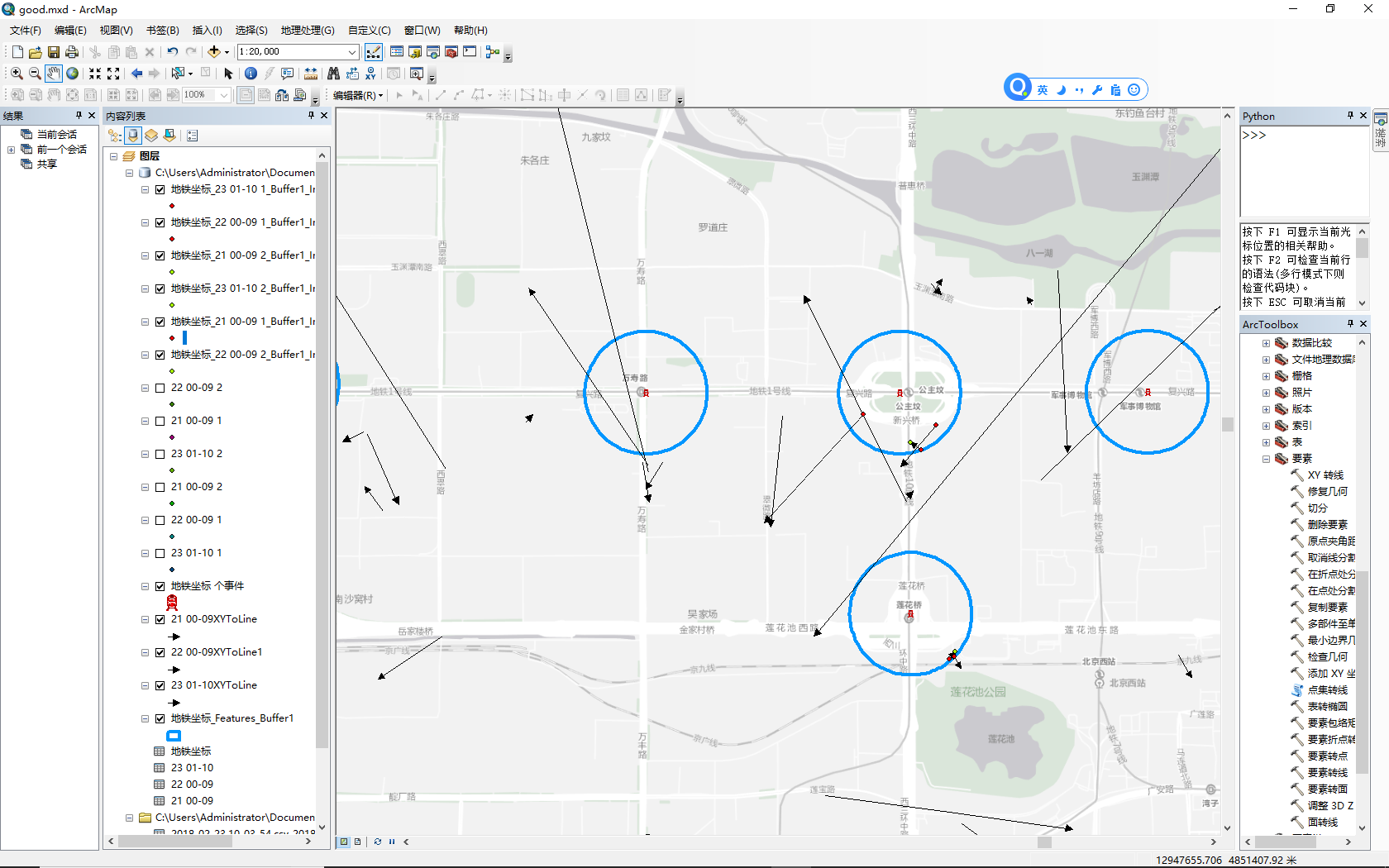
可能原因：

1. 城市生活水平较高，大部分居民在冬天选择乘私家车出行。
2. 城市公共交通便利，路网发达，冬天骑行相比较于公交车舒适性较差。

### 0时-9时：

 此时间段内，在郊区的地铁多呈现为绿色点聚集，也就是大量居民选择地铁站作为终点。

此为北京市东北五环外的马泉营地铁站流向图，在上班高峰期，居民多乘坐地铁向城市中心移动，而几乎没有人从地铁站出发向外骑行。



在市中心的情况就与之相反，多呈现红色聚集，多数地铁乘客下车后，乘摩拜自行车前往上班的地点。例如：处于西三环的公主坟地铁站，大多数骑行都呈现向外的发散状。

### C:\Users\ADMINI~1\AppData\Local\Temp\WeChat Files\271521428608022795.png9时-16时：

此区间内城市和郊区地铁区域无较大差别，而相比于早上，产业园附近车辆移动更为密集，且出发点与终点大多重合，经手机百度地图确认，这些地点附近大多有麦当劳，肯德基之类的餐饮企业。

**参考文献**

要重视参考文献的著录，参考文献按在正文中出现的先后顺序编码并标注，未公开发表的文章、研究报告、内部资料等不列入参考文献，可用页下注的方式注明。参考文献严格按如下格式著录：

期刊论文[序号]作者.题名[J].刊名，出版年份，卷次（期号）：起-止页码.

专著[序号]作者.书名[M].版本（初版不注）.译者（为译著时）.出版地：出版者，出版年：起-止页码.

论文集[序号]作者.篇名[C]//编者.论文集名.出版地：出版者，出版年:起-止页码.

**参考文献示例如下：**

期刊论文：

[1]范文科,张福东,王宗礼,等.中国石油“十一五”天然气勘探新进展与未来大气田勘探新领域分析[J].中国石油勘探,2012,17(1):8-13.

专著：

[5]关增淼，李剑．非洲油气资源与勘探[M]．北京：石油工业出版社，2005．

论文集：[12]张光亚，李宏伟，温志新，等．近年来全球油气勘探发展趋势与启示[C]//中国石油学会石油地质专业委员会．第四届中国石油地质年会论文集．北京：石油工业出版社，2012,567-573．

**2．**图

2.1 图的内容、图题和图注要正确，且要与正文（含表格）相关并一致，应在正文相应位置注明“如图1所示”或括注“（图1）”。文中所引用的图均应线条清晰、文字清楚，图的分辨率应该达到300dpi以上。

2.2 地图是一类特殊的图件，它更要注重科学性、政策性、规范性。（涉及国界地图须报国家测绘局送审，尽量避免使用）

2.2 图题

2.2.1 图题表示插图画面内容的主题，每一幅插图要有图题。

2.2.2 图题要同插图内容相符，确切地表明画面内容的主题。

2.2.3 图题要同正文相呼应，既不能游离于正文之外，更不能与正文相矛盾。

2.2.4 图题要确切、简明，不用修饰性词语。字数较多的图题，可以用简称。

2.2.5 图题后必要时可括注有关内容。

2.3 只有正文中的图编号。需编号的图按顺序编号，例如“图1、图2、图3......”，分图序号以(a)、(b)……表示。分图应有分图名或说明。例如：

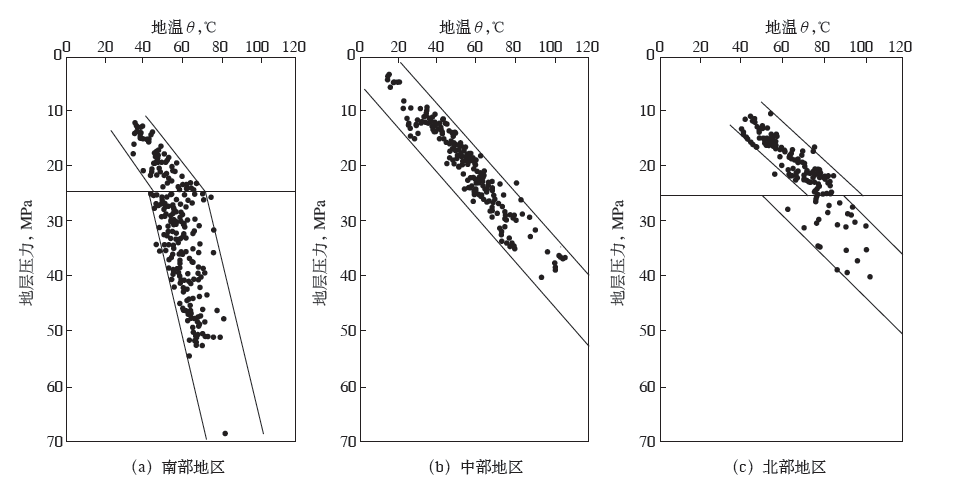


图1 准东地区温压系统分类

2.4 图注

2.4.1 图注是对插图画面某些部位的文字说明。插图中的符号、代号要在图注中或文中加以注释。

2.4.2 图注应内容正确、简洁明了，应与正文及插图一致。

2.4.3 显微放大的图片，要注明放大倍数。

2.4.4 图注置于图题之下。

2.5 图的具体排法

**特别注意：**坐标图中横、纵坐标物理量与单位之间用逗号隔开。横、纵坐标标注有数字的，则除给出横、纵坐标的代表事项外，还应给出相应的单位。例如：

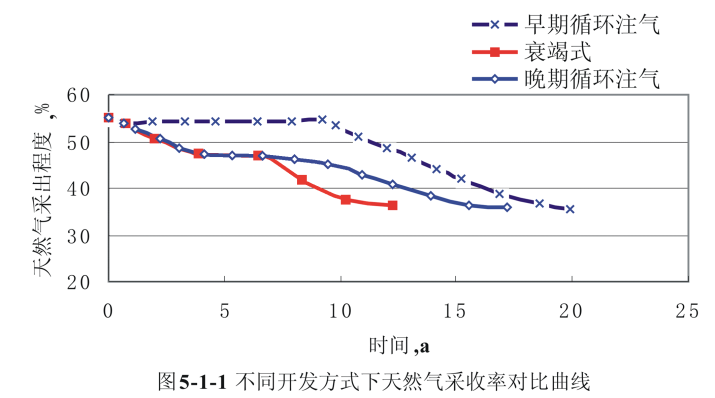


图5 不同开发方式下天然气采收率对比曲线

**3.**表格

3.1 表的内容应与正文一致，不能出现与正文矛盾的现象。同时在正文相应位置注明“见表1”或相应位置括注“（表1）”。

3.1.1 表式，应简单明了，层次清楚，说明什么问题应十分明确。如需要说明问题过多，应分别列表，避免臃肿庞杂、综合内容太多的表格。

3.1.2 标题，应简明扼要，高度概括表的主要内容。若表中数据只用一个计量单位，标题的右上角应注明表的计量单位。

3.2 表注（包括呼应注和全表注）置于表的下方。呼应注在表中加注之处的右上角编码，如①、②；全表注前空两格写“注：”后接序号“（1）、（2）”等，再接写注义。若该表既有全表注，又有呼应注，则先列全表注，后列呼应注。

3.3 表内文字最后一律不用句号。

3.4 表的格式如下：

**表1 \*\*\*\*\*\*\*\*\*\*\*\***

|  |  |  |  |
| --- | --- | --- | --- |
| 压力*p*，MPa | 渗透率*K* , mD | 井数①，口 | 冲数②，次/min |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

注：（1）表中数据为3月份的平均值；

（2）表中数据未考虑\*\*变化的影响。

①…… ；

②…… 。

**4．**公式

4.1 公式的内容应与正文一致，不能出现矛盾的现象，公式按顺序编号

4.2 公式中所有符号均应给出解释，注明单位，公式格式如下：

*K*d=0.00133*H*+0.870 （1）

式中 *K*d──地层可钻性级值；

*H*──井深，m。

特别注意：公式中物理量的符号为斜体（如上式中的*K*d，*H*），单位符号为正体（如上式中的井深，m）。一般下角标若为物理量符号、坐标轴、连续数或表示数的字母符号者应为斜体，其他为正体。