**2017年第四届中国可视化与可视分析大会**

**数据可视分析挑战赛 挑战1**

**（ChinaVis Data Challenge 2017）**

## 背景介绍

“恭喜中大奖了！”、“您可以兑换银行卡积分了！”、“您需要无抵押银行贷款吗？”这类垃圾短信相信你一定收到过不少。用“伪基站”发送垃圾短信已成为不法分子诈骗、推销的新手段，特别是一些附带恶意链接的木马短信，用户点击后还会造成手机中毒和网银盗刷等现象。“伪基站”即假基站，能够搜集以其为中心、一定半径范围内的手机卡信息，利用GSM验证漏洞伪装成运营商的基站，冒用银行、运营商、国家机关或他人号码，强行向用户发送诈骗、色情、赌博、广告等短信息。不法人员利用伪基站发送短信，不仅干扰公共[频率资源](https://www.baidu.com/s?wd=%E9%A2%91%E7%8E%87%E8%B5%84%E6%BA%90&tn=44039180_cpr&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1YLPyPBrHm4uWmzn1TdmWD30ZwV5Hcvrjm3rH6sPfKWUMw85HfYnjn4nH6sgvPsT6KdThsqpZwYTjCEQLGCpyw9Uz4Bmy-bIi4WUvYETgN-TLwGUv3EPWcLnWTdn1m4nW0vPH6YrHmY)，影响人们的正常通讯，而且通过群发短信的方式发送内容不受控的短信息，严重侵害了社会秩序。

QHNet公司推出手机卫士应用软件，安装该软件的用户在收到垃圾短信时可以主动上报，这样手机卫士应用软件积累了大量被用户标注的垃圾短信样本数据，其中每一条短信都包含正文、接收时间、发送者号码以及接收该条垃圾短信前最后连接的正常合法基站信息。如果能够定位到伪基站发送垃圾短信时的位置就能帮助执法人员打击不法分子，但为了保护用户隐私，手机卫士应用软件并未获取用户的精确GPS位置信息。那么如何确定伪基站的位置信息呢？目前，我们将用户主动上报垃圾短信前最后连接的合法基站的位置，近似作为伪基站的位置。但伪基站有非常强的流动性，因此依据近似位置和传统数据分析方法，仍然很难准确把握伪基站的活动规律。QHNet公司这次将部分垃圾短信样本提供出来，希望您使用可视分析方法，揭示伪基站的行为模式，为有关部门建言打击伪基站的有效方案。

## 数据说明

本次比赛提供北京市被标记是伪基站发送的垃圾短信的样本数据，时间跨度为两个月，数据格式为csv，共有300多万行记录，数据大小为753M（压缩前）。数据中可能存在少量缺失和错位的情况，请参赛者自行甄别处理。

垃圾短信样本表包含多个字段，具体字段说明如下：

|  |  |
| --- | --- |
| **字段名称** | **字段含义** |
| phone | 伪基站伪装的发送方电话号码 |
| content | 短信具体正文 |
| md5 | 短信正文MD5 |
| recitime | 垃圾短信接收时间戳 |
| conntime | 与伪基站的连接时间戳- |
| lng | 伪基站发送此条短信时的近似位置经度 |
| lat | 伪基站发送此条短信时的近似位置纬度 |

注：时间戳转化为时间，JavaScript转化方法为new Date(timestamp)，Python转化方法为 time.localtime(timestamp)。

## 问题说明

1. 伪基站常流动于人口密集的区域，以各种名义向一定范围内的手机发送垃圾短信，因此，了解掌握伪基站出行的时空模式，能够帮助执法人员尽早阻止和抓获不法分子，从而更好地维护社会秩序。然而仅仅从垃圾短信中很难确定其对应的伪基站，即无法确定来自同一台伪基站设备的垃圾短信，相同的垃圾短信有可能来自不同的伪基站，同一个伪基站可能不送不同的短信。请从宏观时空分析的角度出发，对垃圾短信数据进行可视分析，揭示伪基站的总体时空活动规律。（建议参赛者回答此题文字不多于1500字，图片不多于8张）
2. 不法分子通过设置伪基站设备能够发送不同类型的垃圾短信，请尝试对垃圾短信的具体内容进行分类，分类标准不限，例如：按垃圾短信类型可以分为广告、诈骗等等，按垃圾短信对人们的人生经济危害程度可以分为一般、严重等等。请尝试在问题1的基础上进一步分析伪基站发送不同类型垃圾短信的时空分布规律。（建议参赛者回答此题文字不多于1500字，图片不多于8张）
3. 伪基站不仅破坏正常电信秩序，危害公共安全，扰乱市场秩序，而且严重损害群众财产权益，侵犯公民个人隐私，社会危害严重。据《人民网》统计，每年通过“伪基站”设备发送诈骗、赌博、推销、中奖等短信近千亿条，伪基站已成为社会一大公害。请结合以上两题中得到的伪基站行为模式，向执法人员提出打击整治伪基站的有效建议和方案，并结合数据分析结果进行说明。（建议参赛者回答此题文字不多于1000字，图片不多于5张）