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 - mini challenge 1）

答　卷

参赛队名称：　　天津大学-侯伟婷-挑战 1

团队成员：　侯伟婷，天津大学，hhhouwt@163.com，队长

林培文，天津大学，409548297@qq.com

于阜甲，天津大学，fujiazhiyu@sina.com

张加万，天津大学，jwzhang@tju.edu.cn，指导老师

是否学生队（是或否）：　是

使用的分析工具或开发工具（如果使用了自己研发的软件或工具请具体说明）：D3，MySQL，ECharts，Python

共计耗费时间（人天）：　60 人天

本次比赛结束后，我们是否可以在网络上公布该答卷与视频（是或否）：是

（灰色字为参赛信息填写模板，请参赛者在提交时参照模板填写）

挑战 1.1：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。（请将回答尽量控制在 1500 字和 8 张图片内）

一、伪基站时间活动规律：

1、首先，按接收时间统计每天的垃圾短信数量，如图 1-1 左图所示，对大部分日期来说，伪基站活动相对稳定，有个别日期出现了同邻近日期相比活动频率大幅增长或降低的现象，例如 02-24 的大幅度减少、03-04 和 03-05 的大幅度增加及 04-02 至 04-04 的减少等。我们推测 03-04 和 03-05 的大幅度增加和北京正值召开两会有关，而 04-02 至 04-04 的减少则和正值清明假期有关。

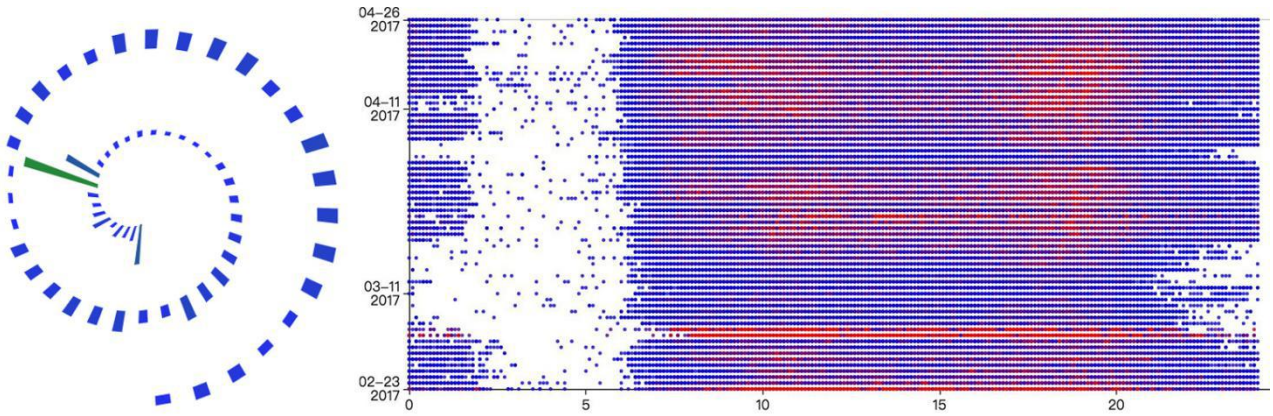

图 1-1　02-23 至 04-26 垃圾短信 24 时内数量变化图

其次，使用散点图显示伪基站每天各个时段的短信数，如图 1-1 右图所示，伪基站一般在凌晨 1~5 点之间休息，7 点后开始活动，一天之内伪基站活动有三个高峰期，第一个是 10:00~12:30 之间，此段时间正是人们上午工作的高峰期，包含了人流较多的午休时间。第二个是 18:00~20:00 之间，此段时间是人们下班时间，交通线上的人流非常密集。第三个是 00:00 左右，这个时刻相较于其邻近时间来说是一个小高峰。

二、伪基站空间活动规律：

1、去掉数据集中重复的经纬度位置后，把所有短信记录的位置投射在地图上，相同位置记录个数多的会更高亮地显示在地图上，得到如图 1-2 所示的分布图，投影在地图上的点勾画出北京各大主干道和高速公路轮廓，由此可知，伪基站多活动于北京的各大主干道和主要高速公路附近，并且可以观察出伪基站活动最密集的区域是朝阳区和东城区的交界地带，这里也是北京最繁华的中心商业地区。

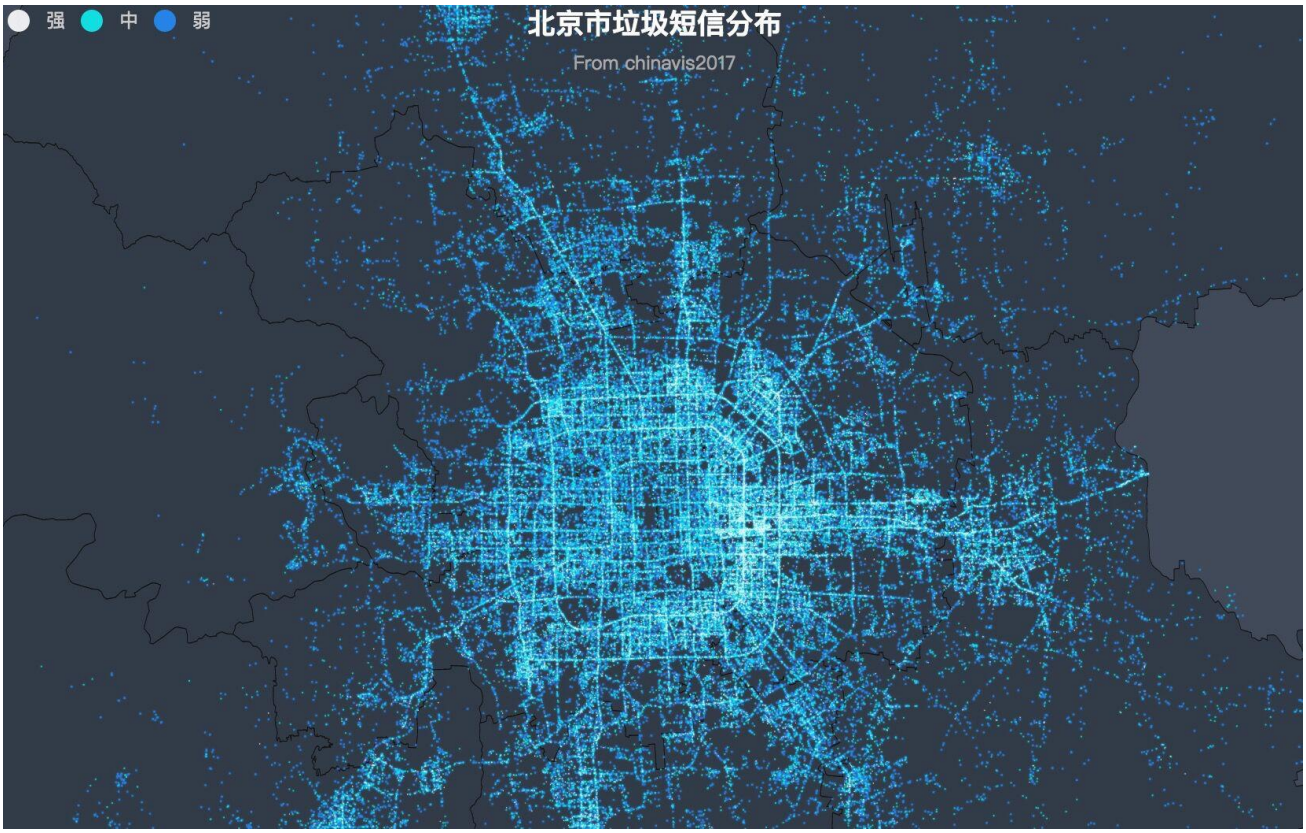

图 1-2　垃圾短信位置分布图

2、图 1-3 是根据垃圾短信经纬度所得的北京市各行政区垃圾短信数量分布图，由图可知，垃圾短信主要集中在朝阳区、丰台区和海淀区等中心城区，其中又以朝阳区的短信数量最多，高达 178 万，而平谷区、密云区、怀柔区及延庆区的分布数量较少。

3、根据短信所在行政区，得到各行政区的短信数据。此时，接收时间连续，经纬度在一定范围内且 MD5 相同的短信可以被认为是由同一个基站发出，据此算出各行政区每天的伪基站总数量，该数值

将一天中不同时间点的同一个伪基站视为不同的伪基站，但基于效率方面考虑，伪基站的工作时长一般为 9-12 小时，所以还需将各区每天的伪基站总数量除以各区伪基站工作时间，则可得到各行政区域内伪基站的平均数量，因为工作时间具有一定的波动，所以计算了各区伪基站数量的最大值和最小值，如图 1-4 所示。由图可知，伪基站主要分布在朝阳区，其次是丰台区和海淀区，而延庆、怀柔、平谷、和密云等区则分布较少。

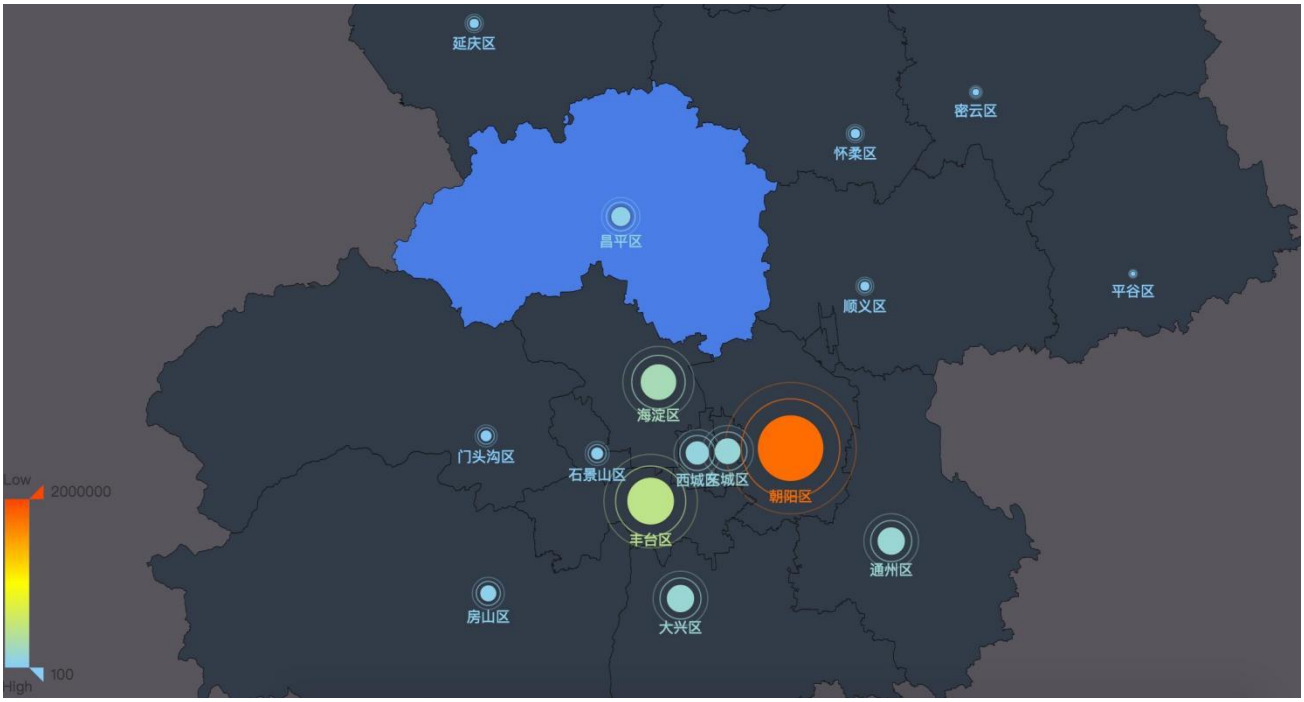

图 1-3　垃圾短信数量空间分布图

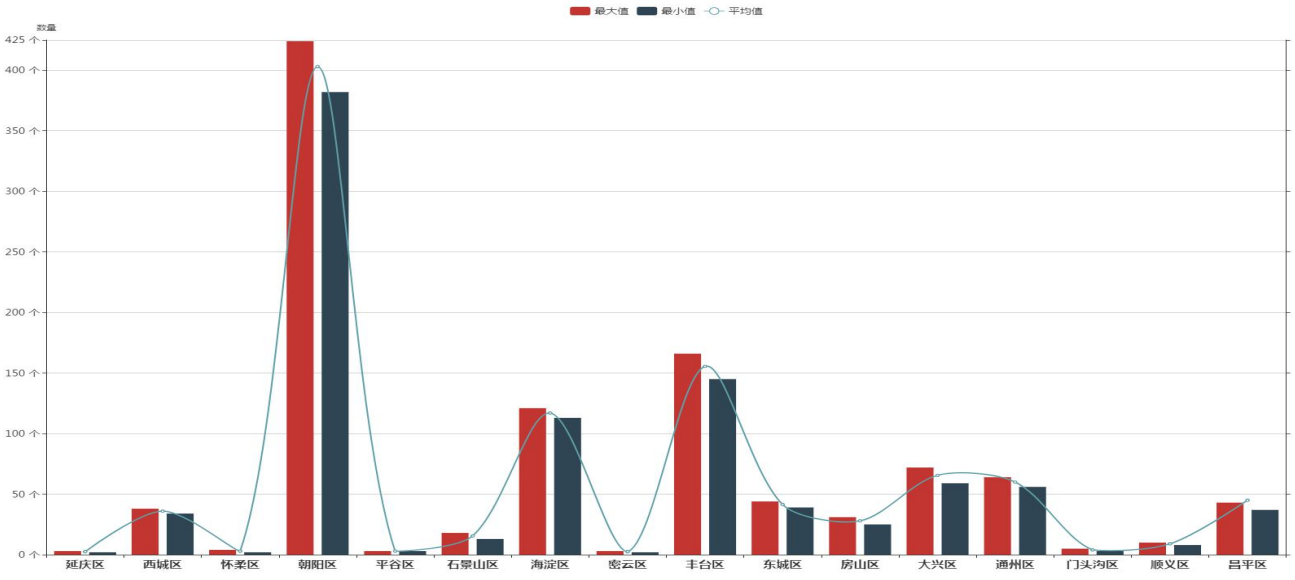

图 1-4　伪基站数量空间分布图

三、伪基站时空活动规律：

1、综合上述分析的伪基站时间和空间分布规律，我们再次对数据集进行处理分析，在同一时间段内，设位置相近的基站的中心点为伪基站经过的真实位置，从这些位置中找出一天中每个时段伪基站出没的高频点，并将其映射在地图上，如图 1-5 中左图所示。图 1-5 右侧折线图实时显示了各时间点内不同行政区域的伪基站数量，大体趋势就是在凌晨 2 点左右趋近于 0，在早晨 6 点左右数量逐步上升，在 10 点左右伪基站数量达到峰值，之后数量小幅度降低并趋于稳定，在晚上 18 点左右数量达到第二个峰值，并开始逐步下降。折线图的最右侧为伪基站在北京 16 个行政区内 24 个时段的活动指数，将其分为了活动非常猖獗、猖獗、一般、较少、轻微、和几乎没有六种情况，可以观察出，伪基站频繁出没于朝阳，东城，西城等地区，其中朝阳区伪基站的猖獗程度无出其右，由于东城、西城的占地面积小，所以虽然伪基站活动频数稍小，但是其活动猖獗等级很高。

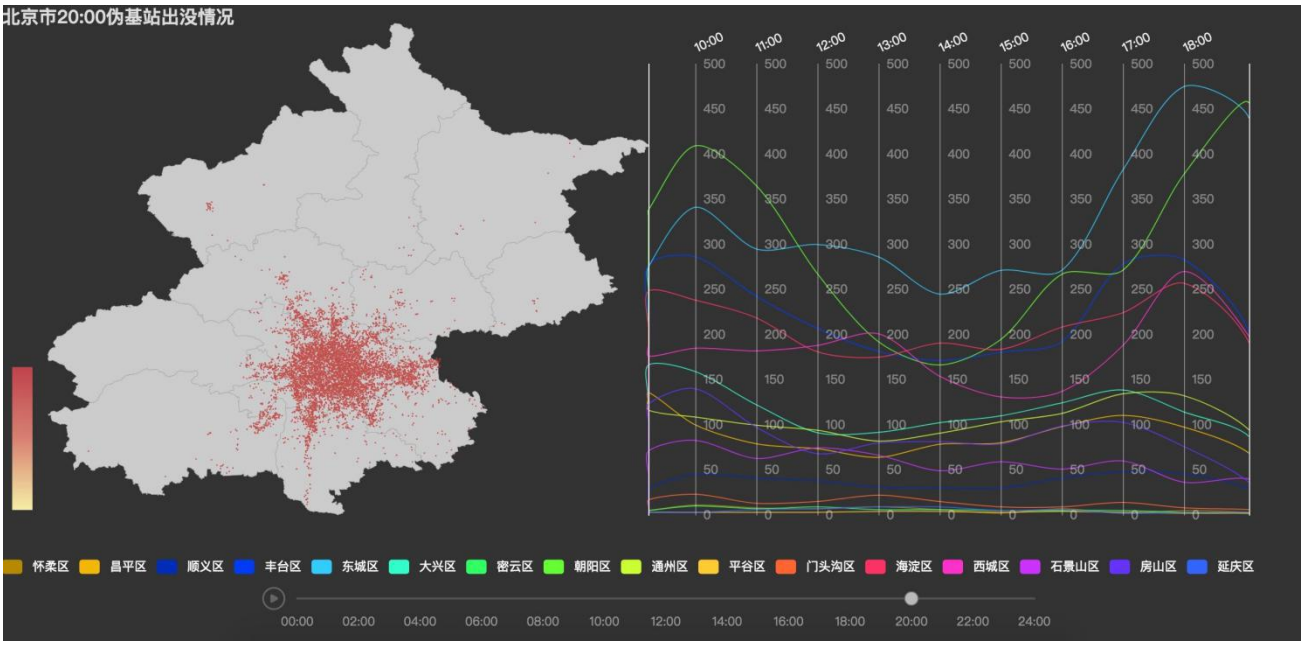

图 1-5 伪基站出没情况联动图

2、由图 1-2 可知，伪基站会频繁经过某些高频点，如果连接某个特定区域内的所有高频点，就能够在一定程度上反映出伪基站在该区域的活动规律。基于此思想，我们通过如下处理得到伪基站的活动规律。首先，根据短信所在商圈，得到北京市各商圈的短信数据。其次，计算出该商圈内所有的高

频点，对所有高频点做聚类，之后对聚类后的数据做曲线拟合，从而得到该商圈内伪基站频繁经过的运动轨迹。最后，由于曲线拟合的结果相对平滑，因此，我们结合实际交通路线，手动在 geojson.io 上细微调整伪基站运动轨迹，使其更符合交通道路走向，最终得出北京各地区伪基站的活动规律如图 1-6 所示。

图 1-6 伪基站运行轨迹拟合图

挑战 1.2：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）（请将回答尽量控制在 1500 个字和 8 张图片内）

一、垃圾短信分类：

1、垃圾短信类型分类

按照垃圾短信的类型分类，首先利用词云显示短信文本中出现频率高的关键词，如图 2-1 所示，关键词包含发票、银行、积分等，说明垃圾短信涉及假冒银行、代开发票、虚假积分、钓鱼网址等内容。

图 2-1　垃圾短信词云

分析垃圾短信具体正文，将垃圾短信分为四大类：广告推销、诈骗、非法服务及其他，其中每一大类都包含数量不等的小类别，具体分类及类别所占比例如图 2-2 所示。由图可知，非法服务类占有最大比重，小类中代开发票占比最高，诈骗类短信其次，色情服务、房地产、假冒机构紧随其后。

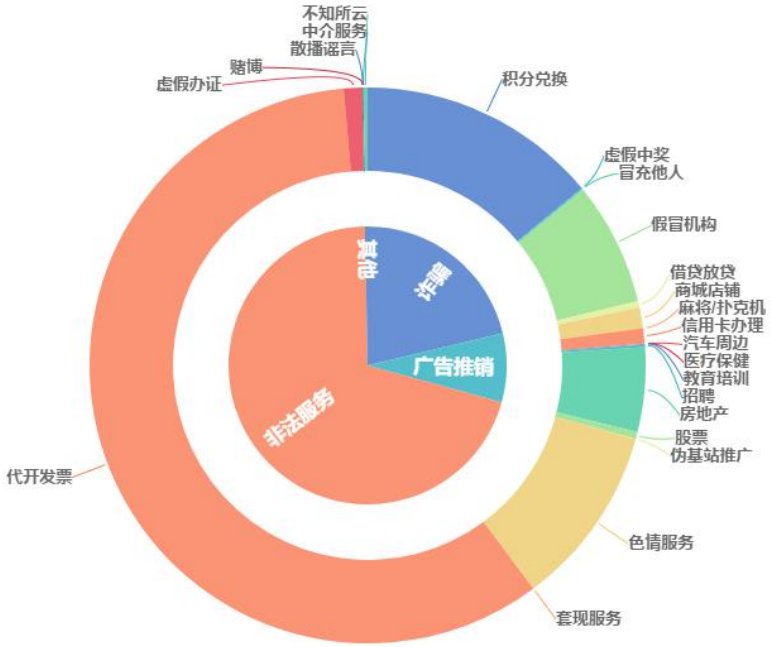

图 2-2　垃圾短信类型占比图

根据问题一得出的伪基站活动时间规律，分析常见类别的垃圾短信数量在 24 小时内的变化趋势，如图 2-3 所示，图中选出最具有代表性的六类垃圾短信。据统计，大部分类别的短信在早晨 6-7 点开始被发送，在中午 11-12 点达到第一个高峰，在晚上 6-7 点达到第二个高峰，但是色情类短信有别于其他类型短信，高峰时间在 19 点到次日 1 点间。

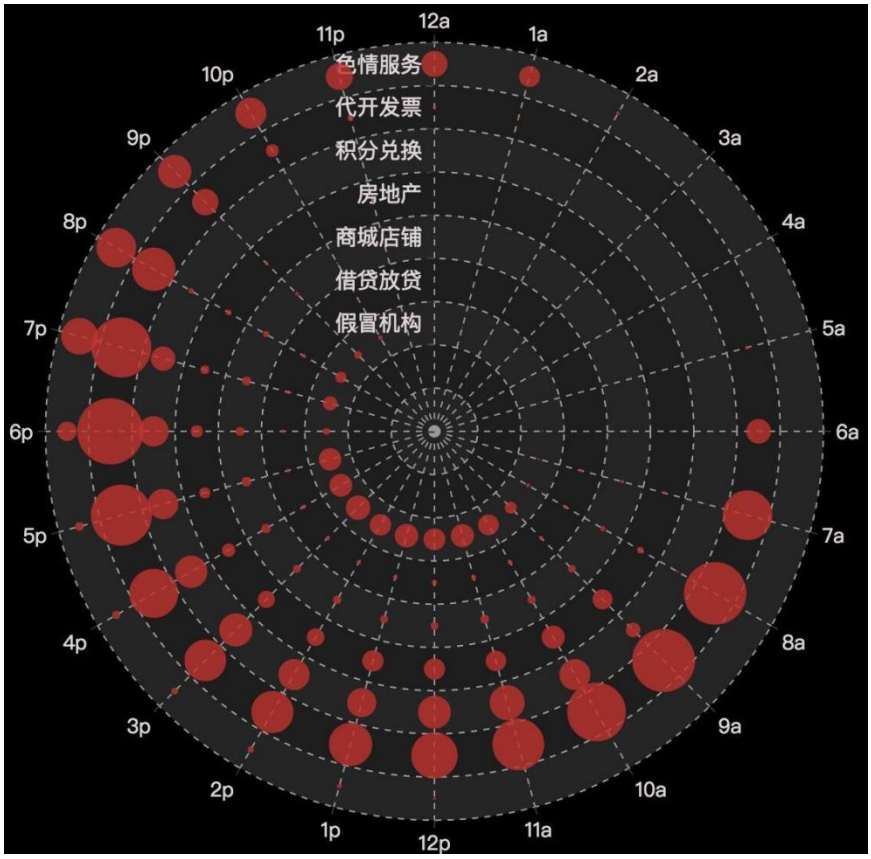

图 2-3　常见短信类别 24 时内数量变化图

2、垃圾短信经济危害程度分类

不同的垃圾短信也因为诱惑程度、推广方式等有着不同程度的经济危害，我们将垃圾短信按照经济危害程度进行分类，分类标准如图 2-4 所示。设定 5 个具有不同权重的因子，即商品、推广、财产、代理、引诱，每条短信按照因子所占比例进行分类，如此可分为产品交易，提升技能，违规服务，非法渠道，侵害利益六类。

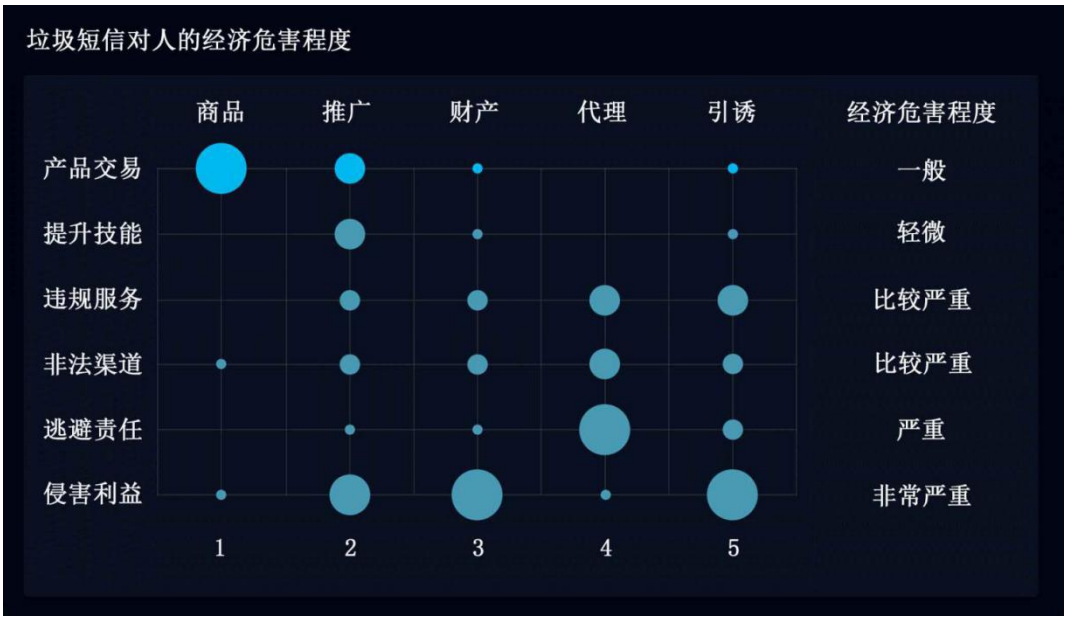

图 2-4　经济危害程度分类

各行政区内不同经济危害程度的短信数量如图 2-5 所示，其中朝阳区最为突出，各类短信数量远超其他行政区，同时由图可知丰台区和海淀区内伪基站情况也较为严重。

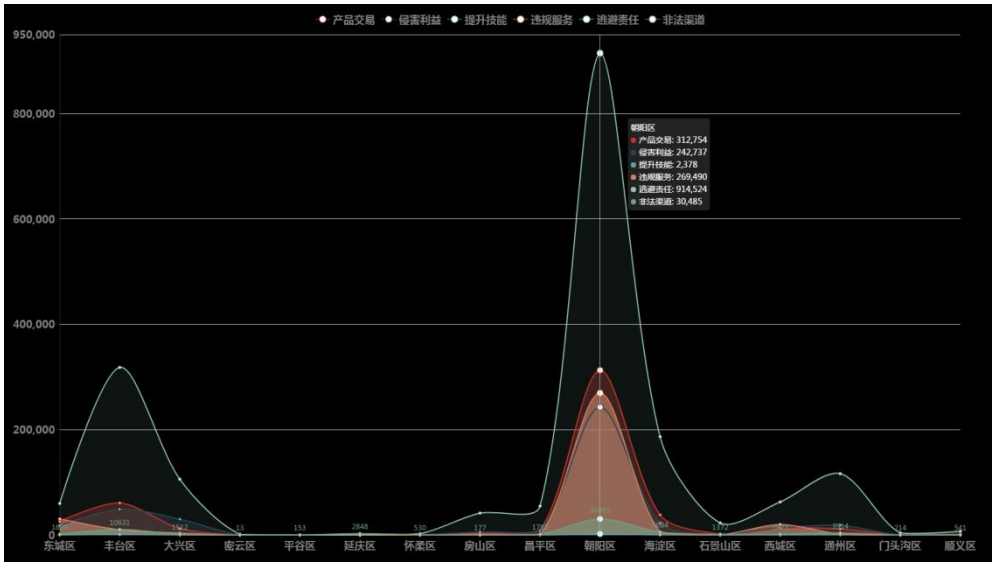

图 2-5　各行政区不同经济危害程度短信类别数量图

二、时空分布规律

短信类型因为自身类型特性在时间上呈现出了个性化特点。图 2-6 综合显示了伪基站出没的商圈与活时间和主要发送的短信类型之间的关系，图中左侧可以高亮显示伪基站活动时间，右侧可以高亮显

示商圈接收到的主要短信类型，中间部分则表示伪基站出没的商圈列表。当鼠标悬浮在某商圈上时，则会高亮显示该商圈内伪基站的活跃时间点和主要短信类型；当鼠标移到左侧某一时间点上时，则会高亮显示在此时间点内伪基站活跃的商圈列表；当鼠标移到右侧部分某一短信类型时，则会高亮显示此短信类型被较频繁发送的商圈。

图 2-7 中左图高亮的部分表示国贸内的伪基站活跃时间在 0、10、11、12、15、17、18、19 和 20 时，其中涉及到的主要短信类型包括积分兑换、假冒机构、商城店铺、房地产、代开发票、色情服务和赌博类。点击中间矩形，显示商圈相关的活跃时间和垃圾短信类型，如图 2-7 的右图所示。观察该图可知，中心圆表示选定的商圈，外围圆数字表示与此类型短信或此活动时间相关的其他商圈个数，点击此外围圆，则显示其他商圈列表，结果如图 2-8 所示，也可在主视图中点击垃圾特定类型和活跃时间。

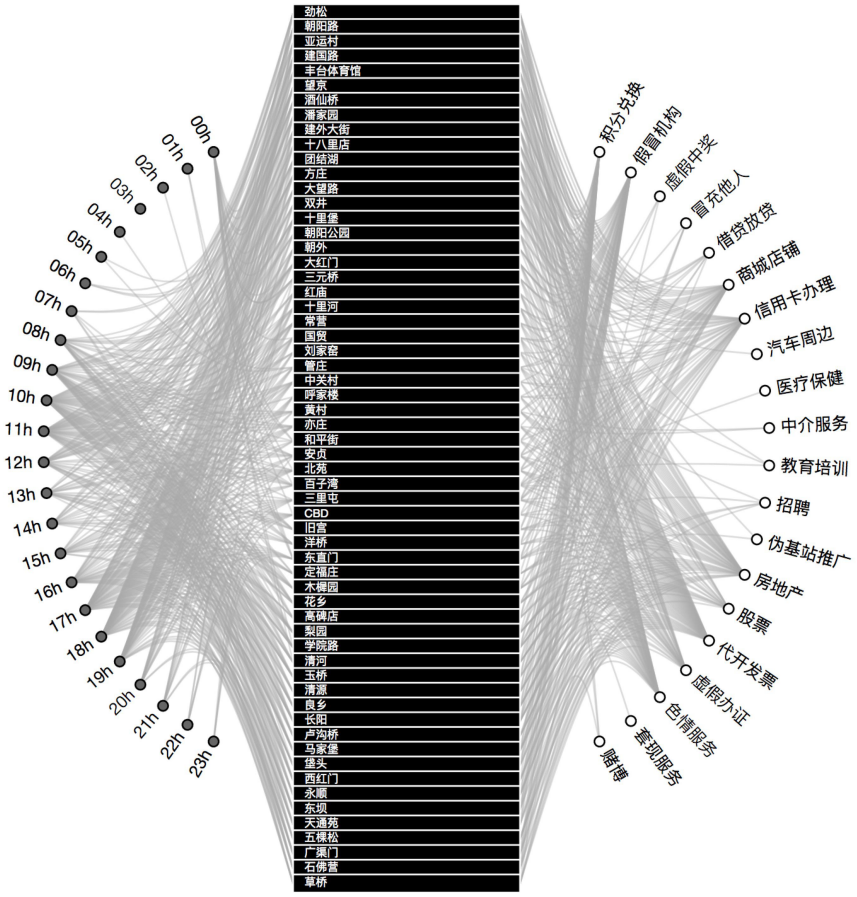

图 2-6　各商圈时间点与垃圾短信类型关系图

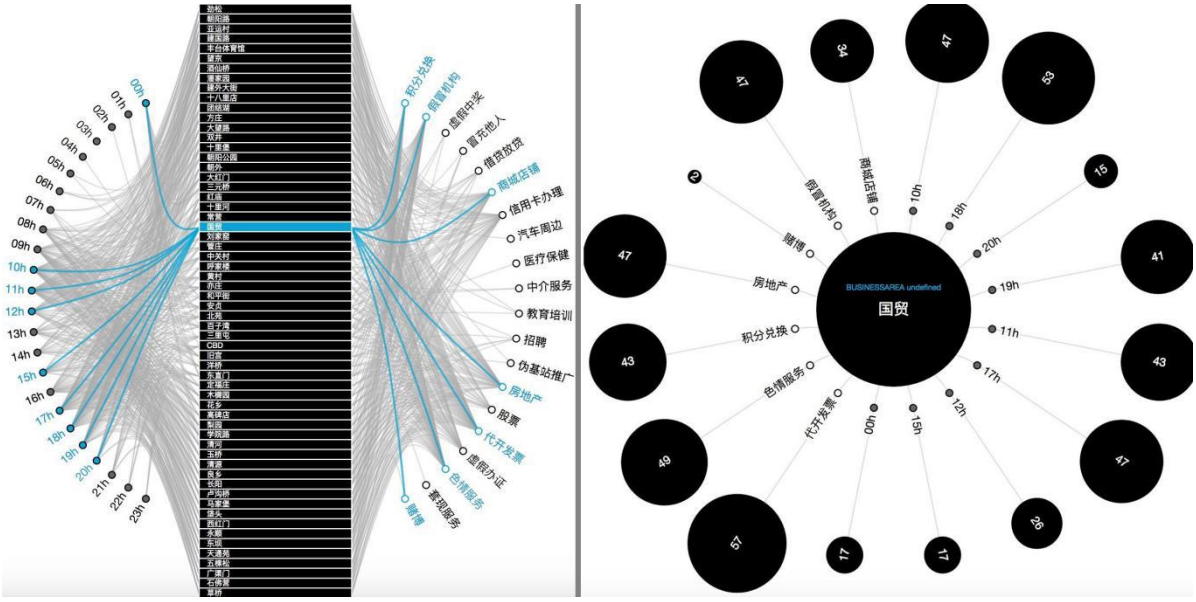

图 2-7　国贸商圈主要的短信类型和伪基站活跃时间图

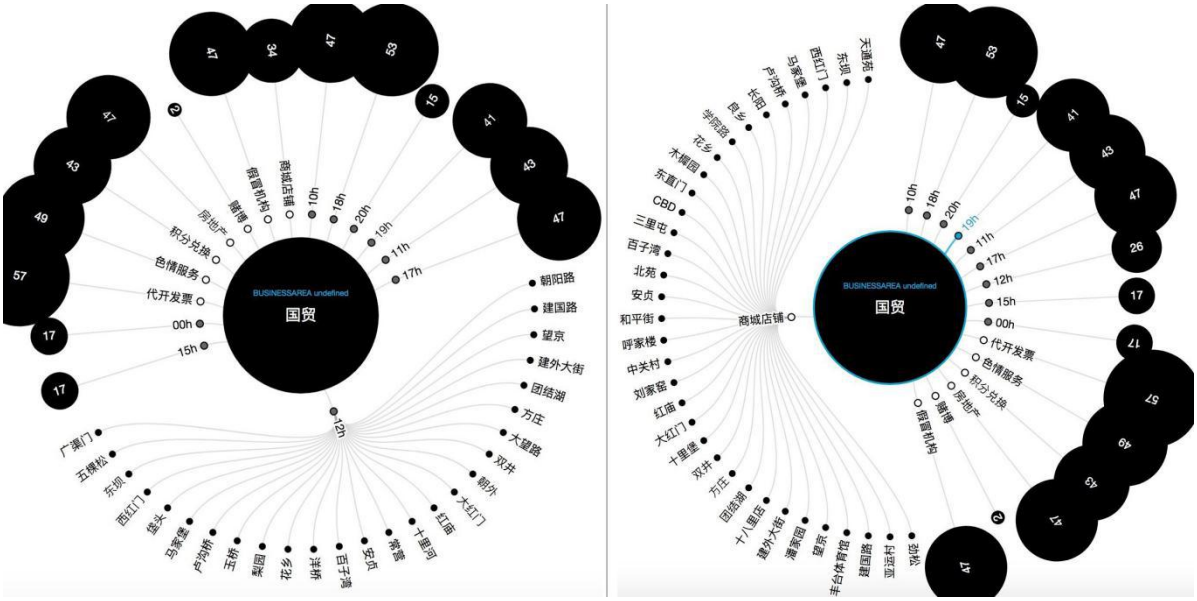

图 2-8 特定短信类型和活跃时间相关商圈列表

挑战 1.3：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。（请将回答尽量限制在 1000 个字和 5 张图片内）

一、分析：

综上分析，伪基站活动范围宽广，执法人员应避免大海捞针，需有更加明确的打击点，从大处着手，小处跟上，以提高整治伪基站的效率。因此，根据图 2-1，选取经济危害程度较高的短信类型，统计它们最泛滥的商圈，结果如图 3-1 所示，气泡的大小表示了数量的多少。由图可知，对侵害利益类别来说，劲松、建国路、大望路和朝阳路商圈的短信数量较多，对违规服务类来说，建国路、亚运村、建外大街等商圈的短信数量较多，对非法渠道来说，团结湖、丰台体育馆、亚运村商圈的短信数量较多，所以执法人员应该加大这些商圈内对伪基站的打击力度。

图 3-1 商圈内较高危害程度短信类别数量对比图

二、建议和方案

1、经过上述对伪基站行为模式的可视化分析，我们提出以下打击伪基站活动的建议：

1) 与运营商合作，找出打击伪基站活动的有利线索；

2) 与垃圾短信文本中广告推销类涉及到的商家进行沟通，挖掘出相应合作的伪基站的有关信息；

3) 结合问题一中我们对伪基站运行轨迹的分析，将伪基站行为模式与安全道路卡口系统结合，排查嫌疑车辆，并实施跟踪抓捕；

4)　各区执法人员根据分析得来的伪基站活动规律配置人员，合理出勤。

2、解决方案：结合问题一问题二中发现的伪基站活动规律和上述建议，以及北京市执法人员大致状况，我们为刚正不阿的执法人员订制出以下方案：

首先，执法人员根据时效性与运营商和涉事商家及时沟通，挖掘出潜在的有关伪基站的有效信息，进而结合图 3-1 中罗列出的危害程度较高的垃圾短信聚集商圈以及图 1-5 中拟合的伪基站活动路线，我们给执法人员制定了如下图的巡逻检测路线，其中上午和午夜的路线是从北京各区分局出发，下午的路线是从上午路线的结束位置出发，路线的数量表示执法人员分组数量。例如：某一条路线早晨从朝阳分局出发，途经朝阳公园、石佛营、十里堡、建国路，于高碑店结束上午的工作，下午从高碑店商圈出发途经朝阳路、定福庄和常营，于百子湾商圈结束下午的工作，午夜从朝阳分局出发，到达国贸检测伪基站的出没情况。

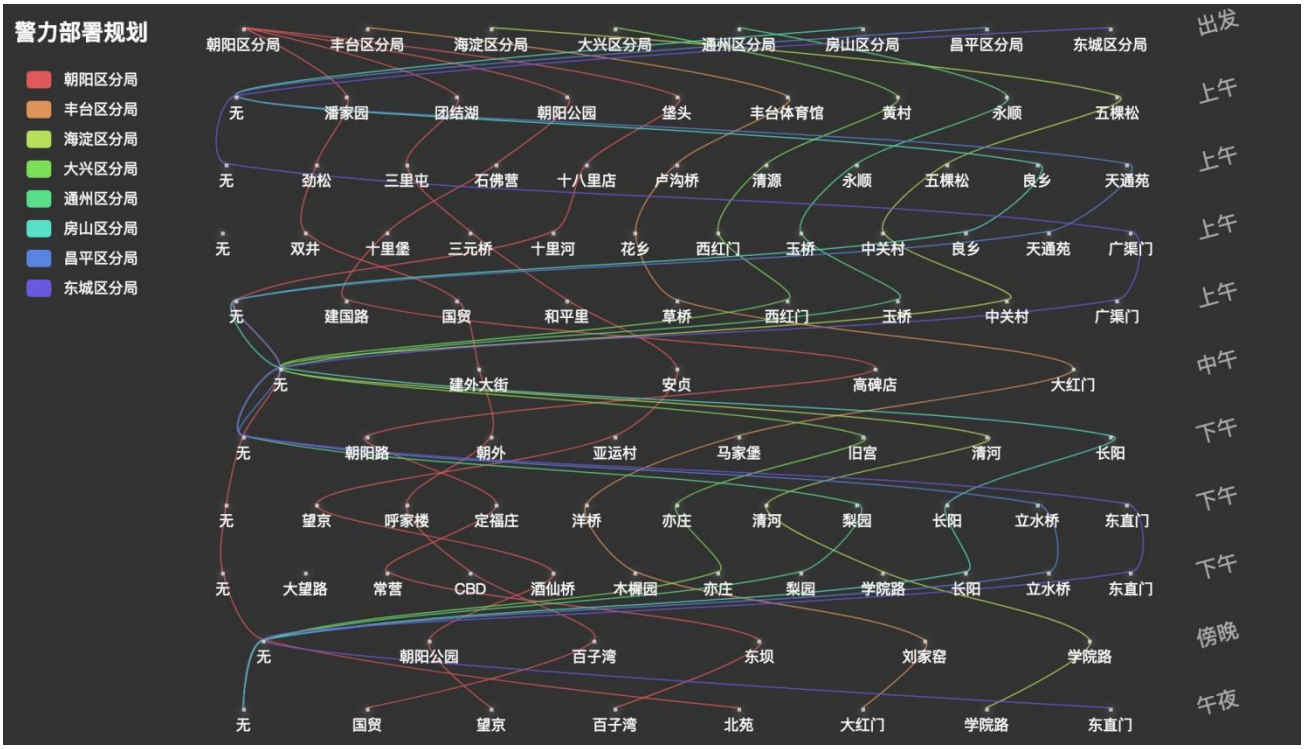

图 3-2　警力部署图

在图 3-2 的基础上，绘制各路线具体行走轨迹，如图 3-3 所示，路线规划的理念是在伪基站活动最猖獗的商圈和时间段内，经过那里，并且使得走过的路径规划得尽量短，这样可以花费最小的代价打击最多的伪基站。其中蓝色线路表示早晨从各区公安分局出发的路线，黄色线路表示下午的路线，黑色线路表示午夜从各区分局出发的路线。

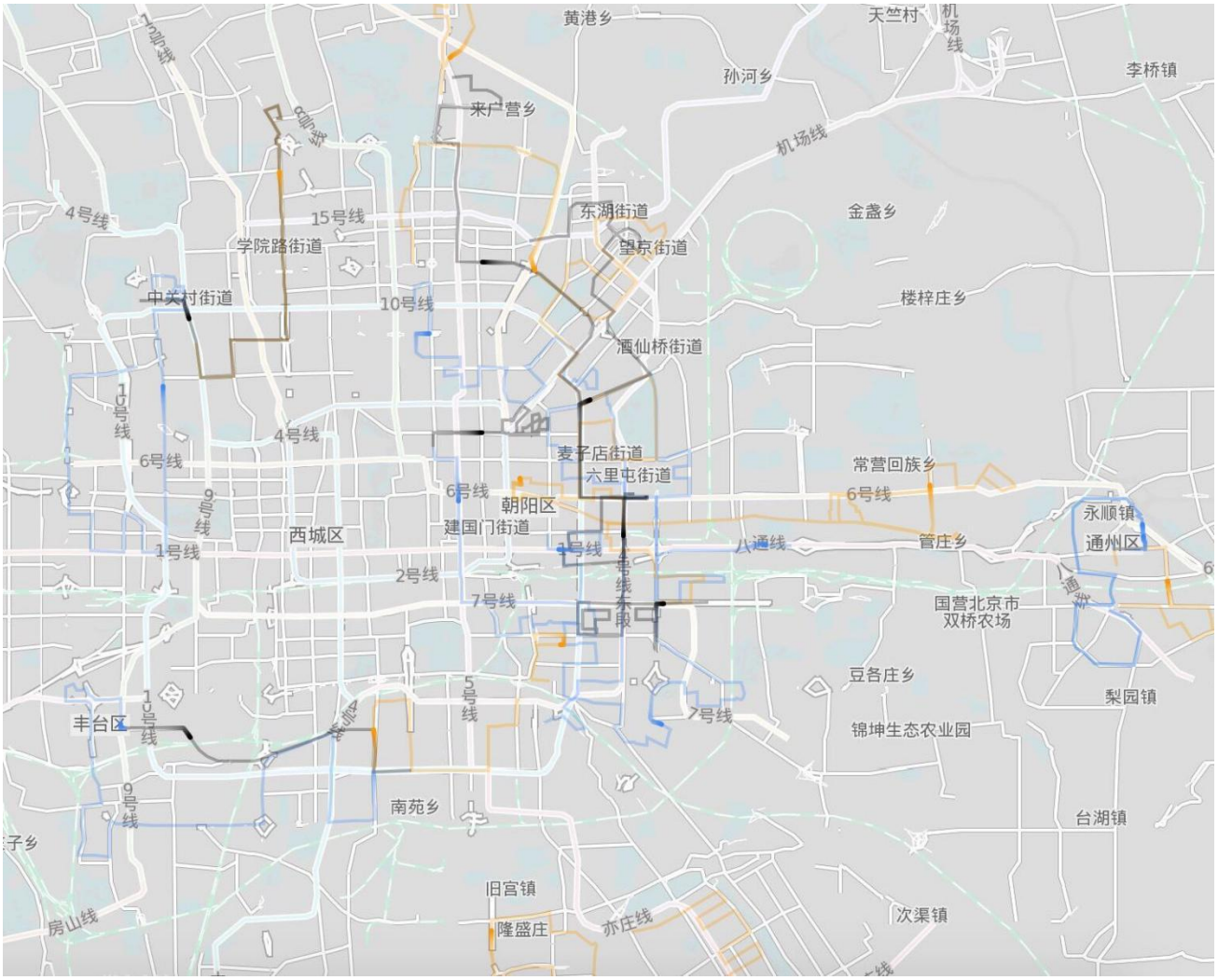

图 3-3　执法人员巡逻路线图