引用格式：汪伟,陶海燕,卓莉,等.北京主城区伪基站时空规律分析[J].地球信息科学学报,2018,20(7):978-987. [Wang W, Tao H Y, Zhuo L, et al. Spatio-temporal analysis of pseudo base stations in Beijing downtown[J]. Journal of Geo-information Science, 2018,20(7):978-987.]
DOI:10.12082/dqxxkx.2018.170430

北京主城区伪基站时空规律分析

汪　伟,陶海燕*,卓　莉,李　敏,李旭亮,汪珂丽,史清丽

中山大学地理科学与规划学院 广东省城市化与地理环境空间模拟重点实验室/综合地理信息研究中心，广州 510275

Spatio-temporal Analysis of Pseudo Base Stations in Beijing Downtown

WANG Wei, TAO Haiyan*, ZHUO Li, LI Min, LI Xuliang, WANG Keli, SHI Qingli

Guangdong Provincial Key Laboratory of Urbanization and Geo-simulation/ Center of Integrated Geographic Information Analysis, School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China

Abstract: The rampant pseudo base stations have become a major public hazard. They undermine the normal telecommunications order, endanger public safety, seriously infringe the property rights of the masses, and violate citizen privacy. How to dig out the spatio-temporal patterns of the pseudo base stations' activities from massive spam messages, design effective prevention and control programs, and fight against the crime from the source, has become the focus of government agencies and researchers. The traditional methods for identifying pseudo base stations through the user terminal, however, face great challenges in terms of accuracy, comprehensiveness, and analytical ability, which no longer meet the requirements of identifying small-scale and mobile pseudo base stations. Utilizing data on the spam messages from February 23rd, 2017 to April 26th, 2017 in Beijing, this paper analyzes the spatio-temporal distribution of pseudo base stations through non-negative matrix factorization. We also constructed a classification model through TF-IDF (Term Frequency-Inverse Document Frequency) which compares types from different classifiers (k-Nearest Neighbors / K-Support Vector Machine / Random Forest/ Single- Layer Neural Network) and selects the most accurate random forest classification method. Combined with the land use data, we analyzed the spatio-temporal distribution of pseudo base stations that send different types of spam messages. The results of non-negative matrix factorization and spam message classification were analyzed in detail. The results show that most of the spam messages in Beijing are sent along the road network and in the central city. The number of spam messages during the day is much more than that during the evening. As time goes by in the day, the distribution of spam messages along the road network gradually shrinks inward. The pseudo base stations that send different types of spam messages differ in the spatio-temporal distribution, but all of them favor the traffic facilities and residential area within the Fourth Ring. The non-negative matrix factorization, which provides reliable results that match with traditional spam message classification, has shown simplicity in performing the analysis and interpretability in the form and result of the

收稿日期：2017-09-17；**修回日期**：2018-01-22.

基金项目：国家自然科学基金项目(41371499)；广东省自然科学基金团队项目(2014A030312010)。[**Foundation items**: National Natural Science Foundation of China, No.41371499; Guangdong Province Natural Science Foundation research team project, No.2014A030312010.]

作者简介：汪　伟(1996-)，男，安徽安庆人，本科生，主要从事时空数据挖掘。E-mail: wangw227@mail2.sysu.edu.cn

***通讯作者**：陶海燕(1966-)，女，江苏扬州人，博士，副教授，主要从事时空数据挖掘、空间流行病学、多智能体地理模拟研究。E-mail: taohy@mail.sysu.edu.cn

decomposition. It can help understand the spatio-temporal patterns of different types of spam messages and provide evident-based suggestions for government agencies to fight against the pseudo base stations effectively. By targeting the source of the spam messages, it is also beneficial for governments to combat the illegal behaviors based on pseudo base stations.

Key words: visualization analysis; non-negative matrix factorization; spam messages; spatio-temporal patterns; Beijing

***Corresponding author:** TAO Haiyan; E-mail: taohy@mail.sysu.edu.cn

摘要：随着公众移动通信的快速发展，伪基站的泛滥不仅破坏正常电信秩序，危害公共安全，而且严重损害群众财产权益，侵犯公民个人隐私，已成为社会一大公害。如何从垃圾短信大数据中挖掘出伪基站活动的时空规律，寻找有效的防控方案，从源头上进行打击和治理成为管理部门和研究者共同关注的焦点。本文基于北京市垃圾短信数据，利用非负矩阵分解的方法分析伪基站的时空分布规律；并利用TF-IDF构建垃圾短信分类模型，对垃圾短信进行分类，结合土地利用数据，分析伪基站在发送不同类型垃圾短信时的时空分布规律。结果显示：北京市垃圾短信多分布于路网和中心城区；白天垃圾短信数量远远多于晚上；垃圾短信的分布随时间的推移沿着路网逐渐向内收缩；发送不同类型垃圾短信的伪基站的时空分布具有一定的差异；通过非负矩阵分解得到的结果，与垃圾短信分类后得到的结果有很好的匹配。研究表明，非负矩阵分解具有实现上的简便性、分解形式和分解结果上的可解释性等优点，可以有针对性的为有关部门建言打击伪基站的有效方案，对于伪基站违法行为的治理具有一定的意义。

关键词：非负矩阵分解；垃圾短信；伪基站；时空规律；北京

1 引言

伪基站是由无线电收发设备和笔记本电脑组成的一种移动无线电通讯设备，能够搜集一定半径范围内的手机卡信息，利用GSM（Global System for Mobile Communication）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、广告等垃圾短信[1]。非法伪基站的出现不仅破坏正常电信秩序，危害公共安全，而且严重侵害群众财产权益，侵犯公民个人隐私，已成为社会一大公害。

随着伪基站设备的多样化，传统的伪基站排查识别手段在准确性、全面性、分析能力等方面遇到较大挑战，已不能满足小型化、易移动伪基站的排查识别要求[2]。伪基站有非常强的流动性，依据近似位置和传统数据分析方法，很难准确把握伪基站的活动规律，而且单独地从时间或者空间维度加以讨论，也往往会忽略时空的交互性，无法得到全面、有价值的结论。

目前伪基站防控的研究重点，是根据伪基站发送的短信内容，通过各种文本识别算法，对短信进行识别、分类，从而构建用户端的垃圾短信过滤系统。垃圾短信过滤方法主要有：安全认证法、基于统计的方法和基于规则的方法[3]3大类：①安全认证法主要通过客户端设置黑、白名单、关键字等过滤条件达到过滤效果；②基于统计的方法则是从统计学的角度探究垃圾短信的规律从而确定规则达到过滤目的。如竺吴辉等[4]基于某省的2000多万条短信记录，根据出入比、有效交互周期等特征建立垃圾短信的过滤算法。该算法针对垃圾短信的查全率达到99.51%，查准率为49.90%；③基于规则的方法则是通过文本识别算法构建过滤系统，主要使用的算法有SVM算法、KNN算法、贝叶斯分类方法等。如邓维维等[5]根据短信分词特征利用贝叶斯分类方法设计出一种移动环境下的垃圾短信过滤系统，综合考虑规则和长度信息，利用互信息减小单词属性的个数。颜世莹[6]则基于行为和内容协作分析，构建垃圾短信过滤系统。徐英慧等[7]基于统计学习理论，根据短信内容提出手机端的垃圾短信过滤策略。目前，国内鲜有学者从源头上研究和分析伪基站的时空分布规律。

传统的从客户端对垃圾短信进行过滤拦截这种被动式的治理，缺乏对伪基站移动规律的基本认知，无法从源头上对此类违法行为进行主动式防治。本文拟利用手机卫士应用软件收集的北京市垃圾短信样本数据，首先，利用非负矩阵分解得到伪基站总体时空规律；再基于TF-IDF(Term Frequency-Inverse Document Frequency)构建垃圾短信

分类模型，根据垃圾短信的文本内容对其进行分类；然后，利用非负矩阵分解结果与分类结果分析伪基站发送不同类型垃圾短信的时空分布规律以及伪基站的行为模式，为有效的打击伪基站的违法行为提供科学的决策依据。

2 研究区概况与数据源

本研究所使用的垃圾短信数据是来自QHNet公司的手机卫士应用软件收集的北京市被标记为垃圾短信的样本数据，时间跨度为2个月，从2017年2月23日到2017年4月26日，经过预处理后，共有3 341 678条记录。样本数据包含伪基站伪装的发送方电话号码、短信具体正文、垃圾短信接收时间戳、与伪基站的连接时间戳、伪基站发送此条短信时的近似位置经度和纬度等共7个字段，具体字段名称与含义如表1所示。

经分析，原始垃圾短信的97%分布在北京六环以内，而北京六环内的面积只占整个北京的14%，因此，本文将北京六环以内区域作为研究区，如图1所示。

3 研究方法

3.1 非负矩阵分解

传统的矩阵分解方法，例如主成分分析（PCA）、独立成分分析（ICA）、奇异值分析（SVD）等，其共同的缺点是分解后结果会出现负值，虽然从理论上分析其结果是正确的，但是在实际应用研究中无论是时间分量还是空间分量，负数均没有实际意义。

非负矩阵分解算法（Non-negative Matrix Factorization，NMF）是Lee和Seung的研究成果[8]，为处

表1 原始数据字段名称与含义

Tab. 1 The field name and definition of the raw data

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	伪基站发送短信时的近似位置经度
lat	伪基站发送短信时的近似位置纬度

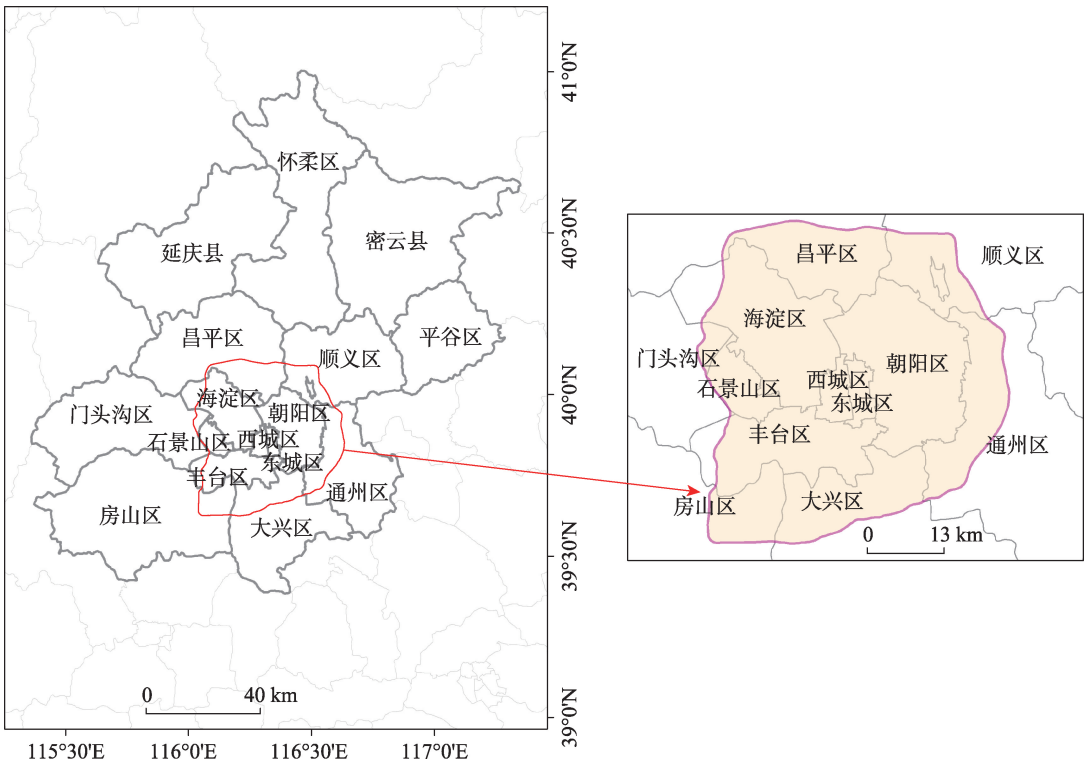


图1 研究区域

Fig. 1 The study area: Beijing, China

理大规模数据提供一种新的途径，具有实现上的简便性、分解形式和分解结果的可解释性，以及占用存储空间少等诸多优点[9]，可使数据的某种潜在结构变得清晰。其基本思想如下：

假设处理 n 个 m 维空间的样本数据，用矩阵 $\boldsymbol{V}_{m\times n}$ 表示，其中 $v_{ij}\geqslant 0(i=1,2,\cdots,m;j=1,2,\cdots,n)$。对 $\boldsymbol{V}_{m\times n}$ 进行线性分解，可以得到：

$$\boldsymbol{V}_{m\times n}\approx \boldsymbol{W}_{m\times r}\times \boldsymbol{H}_{r\times n} \tag{1}$$

NMF 将一个非负矩阵分解成 2 个非负矩阵的乘积，$\boldsymbol{W}_{m\times r}$ 称为基矩阵，$\boldsymbol{H}_{r\times n}$ 称为系数矩阵，其中，参数 r 一般小于 n 和 m。原矩阵 $\boldsymbol{V}$ 的一列向量可以解释为基矩阵 $\boldsymbol{W}$ 中所有列向量（基向量）的加权和，而权重系数为系数矩阵 $\boldsymbol{H}$ 中对应列向量中的元素。非负矩阵分解直接将分解问题作为带约束的非线性规划问题。

$$\begin{cases}\underset{W,H}{\mathrm{Min}}Q(V,W,H)\\ w_{ij}\geqslant 0,\text{where } l\leqslant i\leqslant m,\ l\leqslant j\leqslant r\\ h_{ij}\geqslant 0,\text{where } l\leqslant i\leqslant r,\ l\leqslant j\leqslant n\end{cases} \tag{2}$$

式中：$\boldsymbol{V}$ 是一个 $m\times n$ 维的矩阵；$\boldsymbol{W}$ 和 $\boldsymbol{H}$ 分别表示 $m\times r$ 和 $r\times n$ 维矩阵，$Q(V,W,H)$ 是 $\boldsymbol{H}$ 和 $\boldsymbol{W}\times\boldsymbol{H}$ 之间的距离函数；Min 表示求最小值。本研究采用 Lagrange multiplier 迭代方法使得 $\boldsymbol{V}_{m\times n}$ 与 $\boldsymbol{W}_{m\times r}\times\boldsymbol{H}_{r\times n}$ 之间的距离最小，也就是重构误差最小。

NMF 基于向量组合的形式具有很直观的解释，可以得到原始数据的潜在结构规律。该算法得到的基非负向量组 W 与 H 具有一定的稀疏性和线性无关性。通过选取合适的 r 值，能有力表达原始数据的特征及结构[10-11]。在本次研究中，在时间维度上，根据垃圾短信的 conntime 字段，将时间戳转化为日期，并将一天划分成 12 个时间段，每 2 h 作为一个时间段；在空间维度上，将北京六环划分为 1 km×1 km 的网格，共得到 2151 个网格；构建 2151×12 的空间×时间的矩阵（S×T）。由于基向量的数量 r 直接影响分解结果以及分解结果物理解释的合理性，然而目前并没有成熟的方法确定 r 的值，往往是依据分解结果的合理性进行人为调整，因此在本研究中，经过多次试验，选择 r=3 进行非负矩阵分解，得到时间维度和空间维度的分量。

3.2 垃圾短信分类模型

在现有研究[12]的基础上，将垃圾短信分为欺诈（银行名义、运营商名义、其他）、非法广告（违禁物品买卖、色情服务类、办假证假发票类）、骚扰（恶意骚扰、轻度打扰）和普通广告（房产中介类、金融理财类、其他广告）共 4 大类 11 小类，具体分类标准如表 2 所示。

表 2　垃圾短信分类

Tab. 2　The classification of spam messages

大类名称	大类编号	小类名称	小类编号
欺诈类	1	银行名义	1
		运营商名义	2
		其他	3
非法广告	2	违禁物品买卖	4
		色情服务类	5
		办假证假发票类	6
骚扰	3	恶意骚扰	7
		轻度打扰	8
普通广告	4	房产中介类	9
		金融理财	10
		其他广告	11

本研究首先使用 R 语言的 Rwordseg 包对训练集中短信文本进行分词，去掉标点符号、停用词、助词等得到词条 4858 个。根据 IT-IDF 确定单词的权重，提取短信文本的特征。根据特征使用分类器对测试集中短信进行分类。

其中，TF-IDF 是一种关键词自动提取算法，在计算词语的权值中应用较多且效果较好[13-14]，其主要的思想是某个词或词组的 TF 值在一个文档中高并且在其他文档中较小，那么就认为该词或者词组的类别区分能力强，和其他的词或词组相比，其更适宜用于分类[15]。词频 TF（Term Frequency ）表示某一给定的词语 i 在文档 d 中出现的次数。反文档频率 IDF（Inverse Document Frequency）是一个词或者词组的普遍重要性的度量。常用的计算公式如（3）、（4）所示[14]：

$$TF=t/s \tag{3}$$

式中：t 表示特征词在文档 e 中出现的次数；s 表示文档 e 在出现的总词条数目。

$$IDF=\log(D/d+0.01) \tag{4}$$

式中：D 表示总文档数目；d 表示包含特征词的文档总数。

模型分类流程如图 2 所示。所有实验采用五分交叉验证，即把数据集随机划分成 5 份，每次取其中 4 份进行训练，剩下 1 份进行测试，然后把 5 次分类结果的平均值作为结果，再整体迭代 5 次取平均值作为最终结果。本研究使用 KNN（k=1）（k-Nearest Neighbors）、线性 k-SVM（K- Support Vector Ma-

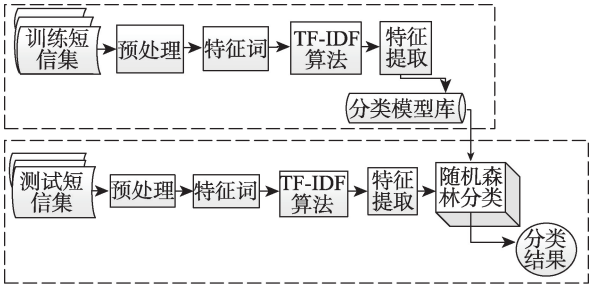


图2 垃圾短信分类模型流程图

Fig. 2 The flow chart for the spam messages classification model

chine)、随机森林RF(Random Forest)和单层神经网络nnet(Single-Layer Neural Network)4种分类器[16-20],得到11类短信分类结果的正确率p、召回率r、$F1$,并将准确率、kappa系数和p值作为整体评价指标,如表3与表4所示。通过比较不同分类器的分类精度,本研究最终选用了模型准确率95%和kappa系数93%的RF模型构建分类器进行分类。

4 结果与分析

4.1 伪基站时空统计规律

根据预处理后每条短信记录所代表的伪基站近似经纬度,计算北京六环内每个网格的垃圾短信发送量,生成垃圾短信分布密度图,如图3所示。从图中可以看出,伪基站主要集中在中心城区,距中心城区越远,伪基站分布越稀疏,其中朝阳区西部伪基站分布密度最大。密度较高的地区有沿着路网分布的趋势,四环路、京通快速公路、京开高速等公路覆盖区短信密度均较高,一定程度上说明垃圾短信较多是利用车载伪基站进行发送的。

将一天划分成12个时间段,统计每个时间段的短信数量,得到如图4所示的垃圾短信随时间的分布图。从图4可以看出,从第一天晚上20:00到第二天8:00,是垃圾短信发送最少的几个时间段。从8:00-20:00,垃圾短信的数量都比较多,尤其在8:00-12:00与16:00-20:00,垃圾短信的数量最多。这也比较符合人们作息的实际情况。

4.2 非负矩阵分解结果

非负矩阵分解结果的时间分量如图5所示,其中,横轴表示时间,纵轴表示对应时间的概率p,对应的空间分量如图6所示。从时间分量看,r=3时可以将一天分出3个不同的时间段,分别对应工作、傍晚、夜间模式。从不同时间分量分析对应的空间分量的分布情况:T2对应的S2,是工作模式(6:00-16:00)下对应的垃圾短信分布情况,可以看出此时

表3 分类结果及精度

Tab. 3 The classification result and its accuracy

分类器	指标	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	平均
RF	p	0.98	0.95	0.12	1	0.98	0.99	0.98	0.5	0.97	0.98	0.91	0.85
	r	1	0.77	0.06	0.69	0.91	0.99	0.85	0.98	0.96	0.93	0.69	0.8
	F1	0.99	0.84	0.08	0.8	0.94	0.99	0.91	0.66	0.97	0.96	0.78	0.81
KNN	p	0.99	0.9	0.58	0.83	0.99	0.98	0.96	0.16	1	0.95	0.99	0.85
	r	0.98	0.51	0.3	0.27	0.88	0.98	0.7	0.67	0.73	0.68	0.36	0.64
	F1	0.98	0.65	0.39	0.39	0.93	0.98	0.8	0.25	0.84	0.79	0.52	0.69
KSVM-linear	p	0.99	0.92	0.52	0.99	0.98	0.98	1	0.41	0.98	0.99	0.89	0.88
	r	1	0.83	0.3	0.73	0.91	1	0.85	0.74	0.96	0.94	0.61	0.81
	F1	1	0.86	0.37	0.83	0.94	0.99	0.92	0.52	0.97	0.96	0.72	0.83
nnet	p	0.98	0.77	0.13	0.87	0.92	0.98	0.96	0.49	0.91	0.94	0.87	0.8
	r	0.99	0.79	0.1	0.68	0.89	0.99	0.89	0.59	0.97	0.95	0.6	0.77
	F1	0.99	0.77	0.11	0.74	0.9	0.99	0.92	0.47	0.94	0.94	0.7	0.77

表4 分类评价指标精度

Tab. 4 The accuracy index of the classification

RF			KNN			KSVM-linear			nnet		
准确率	Kappa	P值	准确率	Kappa	P值	准确率	Kappa	P值	准确率	Kappa	P值
0.95	0.93	0	0.86	0.82	0	0.94	0.92	0	0.93	0.91	0

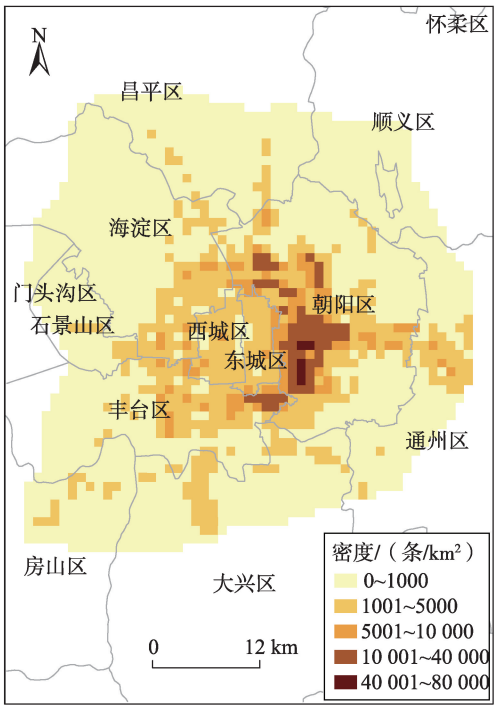


图3　垃圾短信空间分布图

Fig. 3　The spatial distribution of spam messages

垃圾短信是沿着路网分布的，四环路是分布密度最高的区域，在北京-拉萨高速公路、京通快速公路、北京-塘沽公路、京开高速等公路均分布较多。因东城区与朝阳区西靠近北京商务中心区，人流量较大，这些区域的垃圾短信密度最高；T1 对应的 S1，表示的是傍晚模式（16:00-20:00）下的分布情况，与之前的分布相比，短信分布更加集中，沿着四环路的内部密度更高，而外部沿着路网的分布更少，即呈现向内收缩的趋势。T3 对应的 S3，表示的是夜间模式（20:00-24:00，0:00-2:00）下的短信分布情况，垃圾短信数量明显下降，主要分布在朝阳区西以及四环路附近区域。

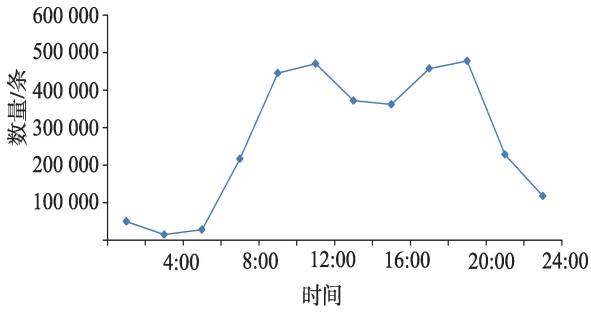


图4　垃圾短信时间分布图

Fig. 4　The temporal distribution of spam messages

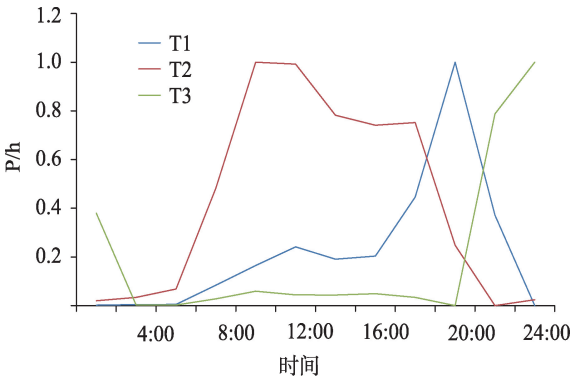
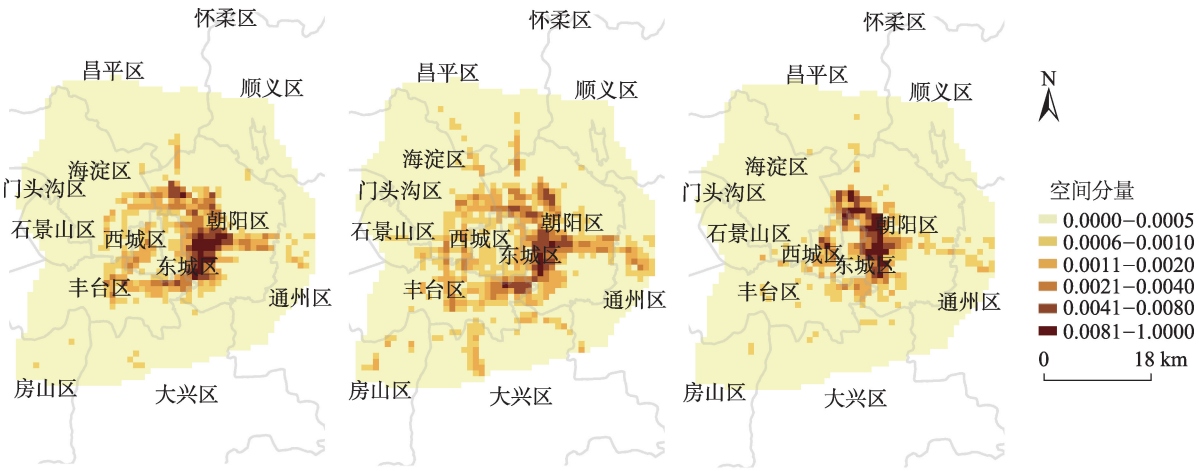


图5　非负矩阵分解时间分量

Fig. 5　The temporal component of NMF



(a) 空间第一分量S1　　(b) 空间第二分量S2　　(c) 空间第三分量S3

图6　非负矩阵分解空间分量

Fig. 6　The spatial component of NMF

4.3 分类型伪基站时空规律

考虑到不同类型短信的数量差异较大，本研究仅仅针对数量较多的假证假发票类、欺诈类、色情服务类、普通广告类、违禁物品买卖类以及骚扰类这6类垃圾短信，其比例分布如图7所示。其中，假证假发票类信息数最多，占总数的58.1%；其次是欺诈类和色情服务类，分别占比24.2%和10.0%。这3类占整个垃圾短信的92.3%。另外，普通广告类占总数的6.0%；违禁物品买卖类占总数的1.2%；骚扰类占总数的0.5%。这3类共计只占整个垃圾短信的7.7%。

（1）不同类型垃圾短信的空间分布特征

每种类型短信空间分布特征（图8）不同。各种类型的垃圾短信都是在中心城区的密度最大，有沿着路网向外分布的趋势。其中，朝阳区西和西城区是密度最大的区域。

利用北京六环内土地利用数据[21]来探究不同类型的垃圾短信在空间分布特征不完全相同的原因，数据主要包括交通设施、绿色地带、政府机构、商业用地、教育用地、公司、住宅以及其他共8类土地利用信息（图9）。试图解释不同类型垃圾短信空间分布的成因。

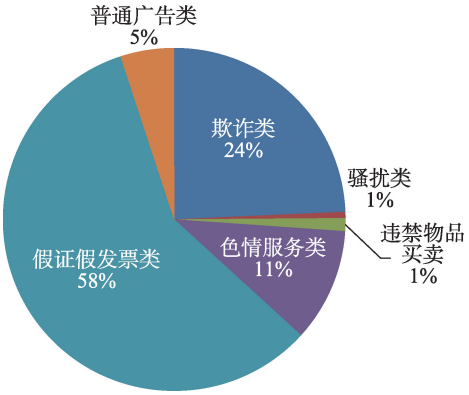


图7 垃圾短信分类类型及比例分布

Fig. 7 The proportion of spam messages by type

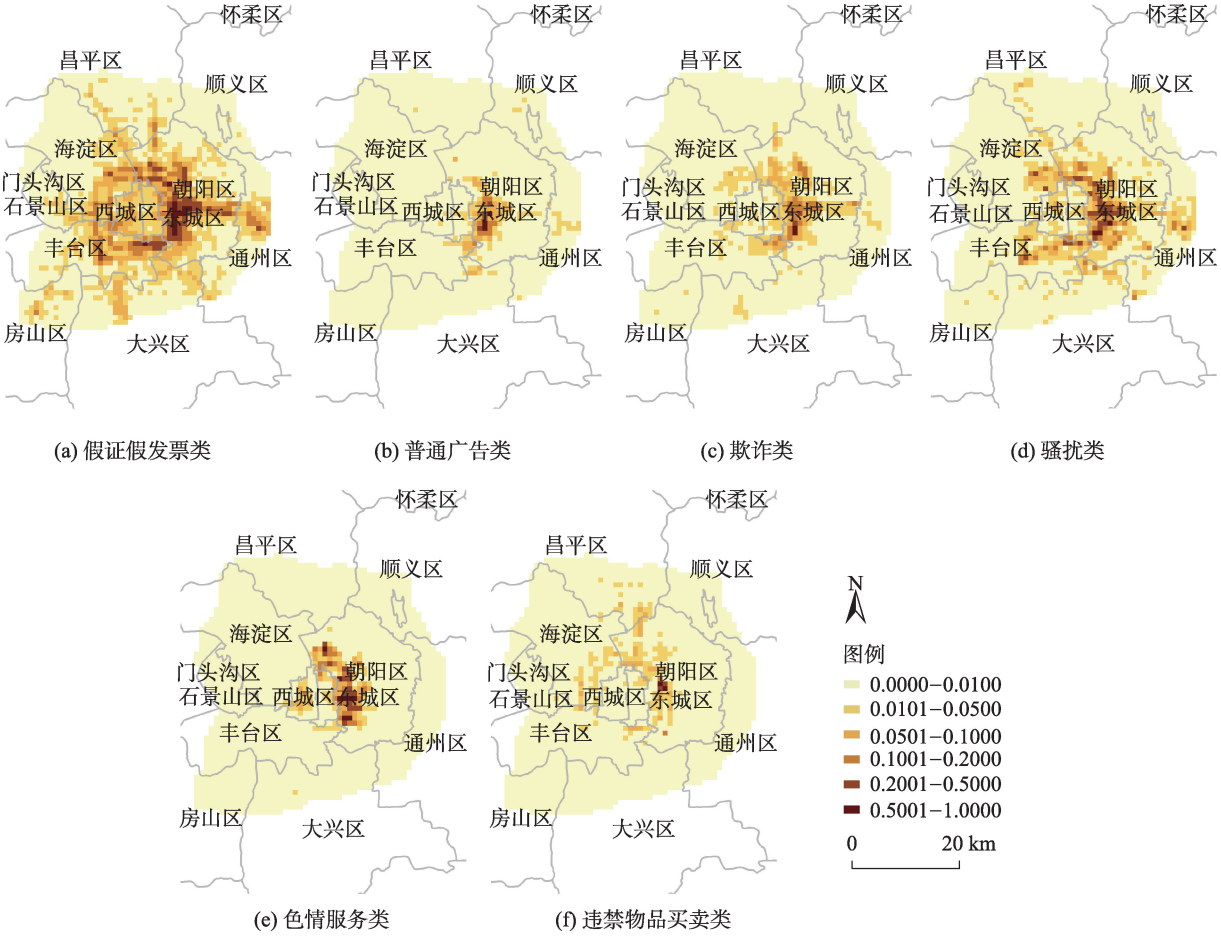


图8 不同类型垃圾短信空间分布

Fig. 8 The spatial distribution of spam messages by type

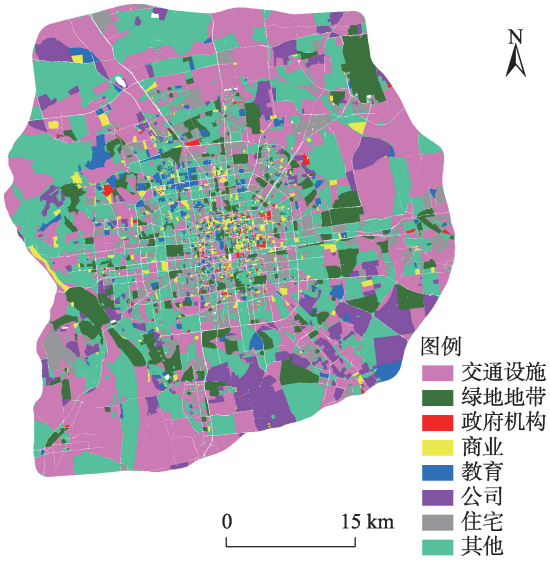


图9　北京六环内土地利用图

Fig. 9　The land use map of Beijing within sixth ring

从各土地利用类型中各类垃圾短信数量上可以看出(图10),伪基站偏好在交通设施与住宅区发送垃圾短信,在这两种用地类型中,欺诈类与假证假发票类垃圾短信占最大的比例,色情服务也占较大的一部分。另外,可以看出在政府机构,垃圾短信的发送数量是最少的,一定程度上说明伪基站更加惧怕在此类型的用地发送垃圾短信。

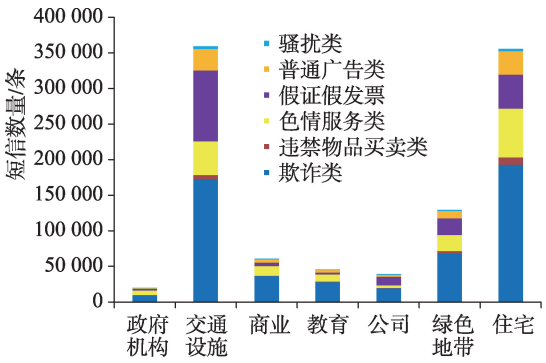


图10　各土地利用类型垃圾短信统计

Fig. 10　The spam message statistics by types of land use

从各土地利用类型中各类垃圾短信发送比例来看(图11),欺诈类、色情服务类及普通广告3类在各功能区发送比例基本相似,即大约30%的短信来源于交通设施地区,40%左右的短信在住宅区发送。违禁物品类短信在住宅发送量较高,而假证假发票类短信更偏爱在交通设施的地方发送,骚扰类在交通设施与住宅发送量基本持平。

(2)不同类型垃圾短信的时间分布特征

不同类型垃圾短信随时间分布如图12所示,根

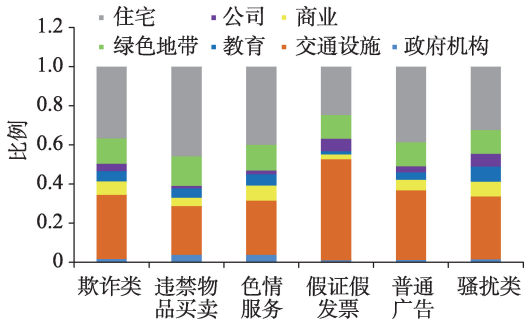


图11　各类型垃圾短信发送地区统计

Fig. 11　The sending area statistics by types of spam messages

据曲线的形状,将6类垃圾短信大致分成4类:第1类是假证假发票类,每天有2个峰值(8:00–10:00与18:00–20:00),在从早上6:00到晚上20:00的时间段内均有较高的分布;第2类是欺诈类,主要集中在每天的10:00至20:00,虽然数量比假证假发票类要少,但是也占有很大的一部分比例;第3类是色情服务类,主要集中在每天的20:00–24:00与0:00–2:00;剩下的是第4类,这类短信数量相对其他的较少,主要集中在每天的10:00–20:00。

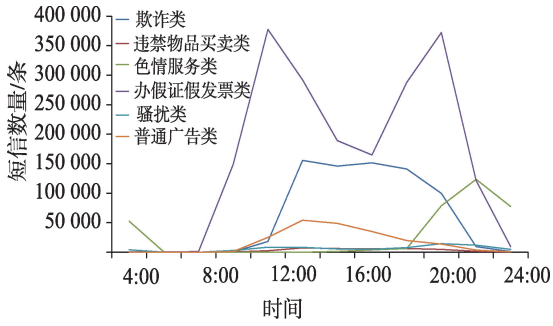


图12　不同类型短信随时间分布

Fig. 12　The temporal distribution of spam messages by type

(3)不同类型垃圾短信的时空分布特征

针对不同类型垃圾短信的时空分布情况,探究每类垃圾短信的时空规律。对于假证假发票类,主要集中在每天的6:00–20:00,在中心城区密度最大,沿着路网向外扩散。该类短信更偏向于在交通设施区发送;对于骚扰类,主要集中在每天的10:00–20:00沿着四环线分布。在交通设施与住宅区发送最多;对于欺诈类,主要集中在10:00–20:00城中心及周边地区。大约30%的短信来源于交通设施地区,40%左右的短信在住宅区发送;对于色情服务类,主要集中在每天的20:00–24:00与0:00–2:00的朝阳区西部以及西城区。大约30%的短信来源于交通设施地区,40%左右的短信在住宅区发送;对

于违禁物品类，主要集中在每天的10:00−20:00沿着四环线分布。多在住宅区发送此类短信；对于普通广告类，主要集中在10:00−20:00朝阳区西部。大约30%的短信来源于交通设施地区，40%左右的短信在住宅区发送。

（4）结合非负矩阵分解结果分析

对比图5所示的非负矩阵分解时间分量与图12所示的不同类型短信随时间分布图，可以发现，T1、T2主要由假证假发票类贡献而来，假证假发票类的两个峰值8:00−10:00和18:00−20:00与T1、T2对应，在这段时间，其他类型短信的数量很少（含有部分的欺诈类）；同理，可以看出T3主要由色情服务类贡献而来。

因此，非负矩阵分解得到的T1、T2以及对应的S1、S2，主要是假证假发票类的时空分布（包含小部分的欺诈类）。T3以及对应的S3，主要是色情服务类的时空分布特征。本研究非负矩阵分解的结果是在3个时间模式下的空间分布，即是在该时间段对应的垃圾短信类型的空间分布，也就是说，非负矩阵分解得到的结果与IT-IDF分类模型得到的结果能够很好的结合与匹配。

5 结论与讨论

本文以北京2个月的垃圾短信为例，首先构建时间乘以空间矩阵，利用非负矩阵分解分别得到时间与空间维度的分量；再通过IT-IDF提取短信文本的特征，利用随机森林算法对垃圾短信进行分类；对非负矩阵分解结果与垃圾短信分类结果进行具体分析。

研究发现，伪基站在空间上主要沿着路网发送垃圾短信，且越靠近北京中心城区，伪基站分布越密集。从时间上看，垃圾短信主要集中在白天8:00−20:00发送，晚上发送量较少。从空间上看，伪基站每天随着时间的推移沿着路网逐渐向内收缩；从不同类型垃圾短信来看，假证假发票类、欺诈类、色情服务类、普通广告类、违禁物品买卖类以及骚扰类这六类垃圾短信的时空特征均不相同。但是在空间上都偏向于在四环内的交通设施与住宅区发送，时间上则主要分布在每天的0:00−18:00。在这两种用地类型中，欺诈类与假证假发票类垃圾短信占最大的比例，色情服务也占较大的一部分。在政府机构，垃圾短信的发送数量是最少的，一定程度上说明伪基站更加惧怕在此类型的用地发送垃圾短信。从分类结果与非负矩阵分解结果结合来看，得到的非负矩阵的曲线T1、T2主要是假证假发票类，S1、S2可以看作主要是假证假发票类的分布。T3主要是色情服务类，S3可以看作主要是色情服务类的分布。了解每种类型的垃圾短信时空规律后，有关部门可以有针对性的制定政策从源头上打击发送某种垃圾短信的行为。

伪基站使用者会根据发送短信的内容选择合适的目标人群，并结合目标人群活动规律，城市交通状况，警方打击形势等选择合适的时间和位置发送相应的短信。从伪基站的移动发送垃圾短信过程中发现其行为规律面临海量数据、空间、时间、类型等多因素的综合作用以及因素之间复杂依赖、随机噪声大等巨大挑战。同时，通过大数据挖掘得到的特征和规律，难以进行定量化的验证，今后拟通过多源数据的集成开展进一步研究。

参考文献(References):

[1] 陈焕煜.使用伪基站群发短信的司法认定[J].人民司法(应用),2016(31):80-83. [Chen H Y. Judicial recognition of sending short messages using pseudo base stations[J]. People's Justice (Application), 2016(31):80-83.]

[2] Zhao M W, Lin-Zhou X U, Shi Z F, et al. A method for illegal pseudo base station site fast measuring and positioning[J]. Mobile Communications, 2016,40(8):18-21.

[3] 李辉,张琦,卢湖川.基于内容的垃圾短信过滤[J].计算机工程,2008,34(12):154-156. [Li H, Zhang Q, Lu H C. Junk SMS filtering based on context[J]. Computer Engineering, 2008,34(12):154-156.]

[4] 竺吴辉,王美清.基于短信发送模式的垃圾号码过滤算法[J].计算机应用,2012,32(12):3565-3568. [Zhu W H, Wang M Q. Span phone number method based on SMS sumission pattern[J]. Journal of Computer Applications, 2012,32(12):3565-3568.]

[5] 邓维维,彭宏.移动环境下的垃圾短信过滤系统的研究[J].计算机应用,2007,27(1):221-224. [Deng W W, Peng H. Research on junk SMS filtering system on mobile environment[J]. Computer Applications, 2007,27(1):221-224.]

[6] 颜世莹.基于行为和内容协作分析的垃圾短信过滤系统[J].电信工程技术与标准化,2011,24(9):54-59. [Xiao Z Y. New applications of IMS network in the future[J]. Telecom Engineering Technics and Standardization, 2011,24(9):54-59.]

[7] 徐英慧,刘梅彦.基于内容的手机端垃圾短信过滤策略研究[J].北京信息科技大学学报(自然科学版),2013,28(1):

51-55. [Xu Y H, Liu M Y. Content-based junk short message filtering for mobile phone[J]. Journal of Beijing Information Science and Technology University, 2013,28 (1):51-55.]

[8] Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999,401(6755): 788-91.

[9] 崔艳荣,何彬彬,张瑛,等.非负矩阵分解融合高光谱和多光谱数据[J].遥感技术与应用,2015,30(1):82-91. [Cui Y R, He B B, Zhang Y, et al. Fusion of hyperspectral and multispectal data using nonnegative matrix factorization [J]. Remote Sensing Technology and Application, 2015,30 (1):82-91.]

[10] 付仲良,刘进军,李金涛,等.加权端元约束非负矩阵分解的高光谱解混算法[J].测绘地理信息,2016,41(2):58-61. [Fu Z L, Liu J J, Li J T. Weighted Endmember Constrained non-negative matrix factorization method for hyperspectral unmixing[J]. Journal of Geomatics, 2016,41 (2):58-61.]

[11] Xie J, Douglas P K, Ying N W, et al. Decoding the encoding of functional brain networks: An fMRI classification comparison of non-negative matrix factorization (NMF), independent component analysis (ICA), and sparse coding algorithms[J]. Journal Neurosci Methods, 2017,282: 81-94.DOI:10.1016/j.jneumeth.2017.03.008

[12] 曾剑秋,杨光永,董豪.垃圾短信分类治理对策研究[J].北京邮电大学学报(社会科学版),2015,17(6):39-44. [Zeng J Q, Yang G Y, Dong H. Span SMS classification governance strategies[J]. Journal of Beijing University of Posts and Telecommunications (Social Sciences Edition), 2015, 17(6):39-44.]

[13] 李旭青,刘湘南,刘美玲,等.水稻冠层氮素含量光谱反演的随机森林算法及区域应用[J].遥感学报,2014,18(4): 923-945. [Li X Q, Liu X N, Liu M L, et al. Random forest algorithm and regional applications of spectral inversion model for estimating canopy concentration in rice[J]. Journal of Remote Sensing, 2014,18(4):923-945.]

[14] Beckschaefer P, Fehrmann L, Harrison R D, et al. Mapping Leaf Area Index in subtropical upland ecosystems using rapideye imagery and the randomforest algorithm [J]. Iforest Biogeosciences & Forestry, 2013,7(1):1-11.

[15] Statnikov A, Wang L, Aliferis C F. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification[J]. Bmc Bioinformatics, 2008,9(1):1-10.

[16] Zhang Y T, Gong L, Wang Y C. An improved TF-IDF approach for text classification[J]. Journal of Zhejiang Universityence A, 2005,6A(1):49-55.

[17] Tu S, Huang M. Mining microblog user interests based on TextRank with TF-IDF factor[J]. Journal of China Universities of Posts & Telecommunications, 2016,23(5):40-46.

[18] 周天宁,明冬萍,赵睿.参数优化随机森林算法的土地覆盖分类[J].测绘科学,2017,42(2):88-94. [Zhou T N, Ming D P, Zhao R. Land cover classification based on algorithm of parameter optimization randon forests[J]. Science of Surveying and Mapping, 2017,42(2):88-94.]

[19] 陈凯,刘凯,柳林,等.基于随机森林的元胞自动机城市扩展模拟——以佛山市为例[J].地理科学进展,2015,34(8): 937-946. [Chen K, Liu K, Liu L, et al. Urban expansion simulation by random-forest-based cellular automata: a case study of Foshan City[J]. Progress in Geography, 2015,34(8):937-946.]

[20] 黄承慧,印鉴,侯昉.一种结合词项语义信息和TF-IDF方法的文本相似度量方法[J].计算机学报,2011,34(5):856-864. [Huang C H, Yin J, Hou F. A Text similarity measurement combining word semantic information with TF-IDF Method[J]. Chinese Journal of Computers, 2011,34 (5):856-864.]

[21] Long Y, Liu X. Automated identification and characterization of parcels (AICP) with OpenStreetMap and Points of Interest[J]. Environment & Planning B, 2013,43(2):498-510.