

基于手机信令数据的居住地人口分布辨识改进方法

黄 伟¹, 孙世超², 孙 娜¹

(1.北京清华同衡规划设计研究院有限公司, 北京 100085; 2.大连海事大学, 辽宁 大连 116026)

摘要: 针对传统的基于手机信令数据的居住地人口分布分析方法的不足, 通过建立一种基于手机信令数据和问卷调查数据的多源数据融合手段, 利用有监督机器学习方法, 实现对居住地人口分布现状的分析。首先通过问卷调查数据获取志愿者的实际居住地位置及其所使用的通信运营商相关信息, 并进行样本筛选。其次, 在通信运营商内部机房建立志愿者用户实际居住地位置与手机信令数据位置信息之间的对应关系。最后, 利用手机信令数据, 通过获取志愿者在居住地网格位置的停留特征以及非居住地网格位置的停留特征训练朴素贝叶斯分类器模型, 继而完成机器学习方法的建立并应用到其他手机用户实际居住地的识别。分析结果表明: 基于有监督学习方法的人口分布辨识方法较传统的阈值判断方法预测精度有明显提升。

关键词: 交通规划; 人口分布; 手机信令数据; 有监督学习方法; 朴素贝叶斯分类器

Improved Method of Population Distribution Identification Based on Cellular Signaling Data

Huang Wei¹, Sun Shichao², Sun Na¹

(1.Beijing Tsinghua Tongheng Urban Planning & Design Institute, Beijing 100085, China; 2.Dalian Maritime University, Dalian Liaoning 116026, China)

Abstract: Aiming to overcome the shortcomings of traditional population distribution analysis methods based on cellular signaling data, a multi-dimensional data fusion method is put forward as a new kind of method, combining cellular signaling data with questionnaire data and supervised machine learning method, to analysis the current situation of the population distribution. Firstly, the actual residential location of volunteers and the information about the communication carrier can be obtained by the questionnaire survey. Meanwhile, sample screening is performed. Secondly, the relationship between the actual residence location of volunteers and the location information of the mobile phone signaling data is established within the internal equipment room of the communication carrier. Finally, using the cellular signaling data, the Naïve Bayesian Classifier model is trained by acquiring the residential features and non-residential features of the volunteer users within the grid location. Furthermore, this method will contribute to identify the actual residence of other mobile phone users. The results show that the population distribution identification method based on the supervised learning method has a significantly improved prediction accuracy compared with the traditional threshold judgment method.

Keywords: transportation planning; population distribution; cellular signaling data; supervised learning method; Naïve Bayesian Classifier

收稿日期: 2019-07-18

作者简介: 黄伟(1969—), 男, 湖南常德人, 硕士, 教授级高级工程师, 主要研究方向: 交通运输规划与管理。E-mail: 15757371@qq.com

0 引言

人口的空间分布是城市社会经济活力、基础设施建设、公共资源配置以及城市交通、住宅、生态环境问题等方面的重要影响

因素之一。及时准确地掌握居住地人口分布是解决城市交通、城市空间布局等问题的基本要素, 是编制城市空间战略规划、国土空间规划以及城市综合交通体系规划等的重要基础, 是各级政府实施日常管理、制定相关

政策的主要依据。

针对人口分布特征的获取方法,国内外学者进行了大量的研究。传统的获取手段主要依托于人口普查、统计年鉴等人口统计数据以及问卷调查数据。人口普查数据采样率较高,但往往间隔周期长,难以获取基年同期数据;统计年鉴数据统计周期相对较短但人口数据精细化程度不够,往往以行政区为单位,且统计口径不同;问卷调查数据能够掌握目标区域人口分布情况,但调查成本较高,且精度受样本量影响较大。

近年来,随着手机的普及使用,手机信令数据作为一种新兴的可用于识别人口分布特征的数据被广泛使用。文献[1]利用手机信令数据,对广州市职住空间总体特征(职住密度、通勤特征等指标)进行刻画,探讨分析了广州四类典型区域的职住发展模式和优化方向。文献[2]提出了一种大数据背景下移动信令数据与空间地理数据应用于交通运输领域的联合分析方法,用于城市人口空间分布与活动规律分析。文献[3]以昆山市为研究对象,基于手机信令数据提取用户职住地信息,分析其职住空间分布特点,并在此基础上分析昆山市职住空间分布和通勤交通现状。文献[4]利用手机信令数据对上海张江科学城的职住不平衡问题进行了深入分析,并提出了内外联动提高住宅总量供给、增加多样化居住设施等规划建议,以解决大规模跨区域通勤出行的问题。

相比于传统数据,手机信令数据能够获得全天候个人出行轨迹,具有规模大、抽样率高、连续追踪采集的特点,上述最新的基于手机信令数据的研究成果正是利用这一特点,聚焦于城市职住平衡发展相关问题的应用研究。尽管基于手机信令数据的人口分布特征分析已经在多个城市得到应用,但仍未形成一套完整而有效的标准化分析手段^[5-7]。此外,上述研究成果更加注重利用手机信令数据获取特征人口分布特征后的实际问题应用,但对于基础数据处理过程中技术方法的应用合理性和准确性缺乏讨论。

造成这种局面的主要原因在于:1)手机信令数据的定位方式是以基站位置信息近似代替用户的实际位置信息,因此在定位精度上造成了偏差,而校正这种偏差的难度较大;2)通常只会采用一家通信运营商的数据进行居住人口分布特征的分析,由于不是全

样本,因此涉及较为复杂的扩样过程;3)通信运营商的信令数据采集标准各不相同,例如基站密度以及基站的服务半径差异较大,因此,即使同时获取了多家通信运营商的手机信令数据,仍无法建立数据之间的融合;4)目前利用手机信令数据进行人口分布的识别大多基于阈值判断法,主要通过判定候选居住地位置的单日停留时间以及停留天数是否超过所设置的阈值来进行筛选,然而通常使用经验法对上述阈值进行设定,并不具备一定的科学性,对于选取的阈值组合不能够提供相应的假设依据,对于判断结果也没有标准化的校核过程。因此,寻找合理选取相关阈值的科学化方法并且能够对分析结果进行有效的校核,是当前人口分布判断的重点任务。

本次研究对象,即居住地人口,指研究周期内在本地长期拥有相对固定夜间居住地的人口,也称为夜间人口。居住地人口与常住人口的定义略有不同,当地居住未满6个月但在研究周期内拥有相对固定夜间居住地的人口仍属于居住地人口。由于居住地人口通常长期生活在本地,其居住地分布情况同样是解决城市空间布局问题的基本要素。鉴于此,并针对单一数据源分析方法在人口分布应用中的不足,本文通过一种基于手机信令数据和问卷调查数据的多源数据融合手段,利用有监督机器学习方法,实现基于多源数据的居住地人口分布现状分析。具体包括以下几步:1)通过问卷调查数据获取志愿者的实际居住地位置及其所使用的通信运营商相关信息,并进行样本筛选;2)在通信运营商内部机房建立志愿者用户实际居住地位置与手机信令数据位置信息之间的对应关系;3)利用手机信令数据,通过获取志愿者用户在居住地网格位置的停留特征以及非居住地网格位置的停留特征,训练朴素贝叶斯分类器模型,继而完成机器学习方法的建立并应用到其他手机用户的实际居住地识别。

1 数据基础

1.1 手机信令数据

根据移动通信网络的覆盖特性,以及移动通信网络需具备为手机用户连续提供服务的功能,用户的手机终端会定期或不定期、主动或被动地与移动通信网络保持联系,这

些联系被移动通信网络识别成一系列的控制指令，即手机信令数据。将所有的手机信令数据整合起来，就能识别出一个手机用户的活动轨迹。

手机信令数据主要包括两类：原始手机信令数据表以及网格(500 m×500 m)基础信息表。原始手机信令是基于通信运营商的信令采集系统，采集匿名移动用户在发生主要信令事件时产生的手机信令数据；网格基础信息表由通信运营商提供，用于对应信令数据产生时的位置信息，本研究使用到的字段主要包括网格编码和标准网格中心点的地理空间坐标(见表1)。

1.2 问卷调查数据

相比于手机信令数据，问卷调查数据虽然样本量较少，但可靠性与抽样均衡性相对较高。为了能够更加准确地获取居住地人口现状分布情况，本研究首先提出了一种数据融合方法来建立问卷调查数据与手机信令数据之间的个体匹配，目的是想通过问卷调查的方式获取受访者真实的居住地位置，再结合手机信令数据所反映的各种时空特征，建立居住地结果判断与个体时空特征之间的关联性，最后用于手机信令数据总体样本的居住地分布结果判断。

为了建立手机用户与问卷调查受访者之间的关联匹配，最直接、简单的方法是通过获取问卷调查受访者的手机号并与手机信令数据中的手机号进行一一匹配。然而，由于手机信令数据的隐私保护特点，数据中所包含的个人手机号信息已经进行了相应的加密。为此，本研究设计了以下数据采集流程，既考虑对用户隐私的保护(确保无法根据所获取的数据实现对某一个体的轨迹追踪)，又能够建立真实居住地与手机信令数据特征分析结果之间的关联。具体流程包括以下三个步骤：

1) 按照抽样调查的基本原则，选择合适的志愿者样本，通过问卷调查获取志愿者的手机号码、所属的通信运营商情况以及实际居住地位置(精确到小区名称)；

2) 筛选所合作的通信运营商用户，告知相关受访者采集手机号码信息的用途，表明不会泄露或追踪其个体轨迹信息，与同意参与此项研究的志愿受访者签订知情授权书；

3) 利用GIS技术将各志愿受访者提供的真实居住地位置转换为地图坐标点，并将用户手机号、居住地坐标点信息提供给所归属的通信运营商，由其进行下一步数据匹配和集计处理，避免外部人员接触到受访者个人隐私数据。

2 数据预处理方法

2.1 手机信令数据预处理

受手机信令数据采集方法的影响，在海量的手机信令数据中，往往会产生数据冗余、数据漂移及数据缺失等问题。因此需要进行数据预处理，去除冗余及错误数据，修补缺失数据，最终得到有效且完备的数据。

1) 数据清洗。

数据清洗是在系统的各个处理阶段分别设置过滤器，根据不同原则将原始的信令数据一步步过滤去噪，对海量信令数据进行分析 and 筛选，去除数据噪声。数据清洗的原则包括：①字段有空的数据；②重复的记录；③数据标记为不正常数据；④事件类型错误的的数据；⑤用户字段异常数据；⑥其他错误数据。

2) 数据修补。

手机用户的数据采集受到信号连接不良、设备采集装备未升级、数据量溢出等影响，可能会出现数据的字段缺失现象。为保证手机用户轨迹的连续性，对于重要字段缺失的数据进行剔除，对于可根据轨迹信息进行修补的数据进行一定的修补。

修补的依据主要为同一用户的轨迹时间排列信令数据，首先将同一用户信令数据按

表1 信令数据及网格信息汇总

Tab.1 Cellular signaling data sheets

类型	字段名称	描述
信令数据	IMSI	手机识别号，IMSI或由IMSI单向加密的结果，唯一标识手机
	TimeStamp	时间戳，由厂商在采集卡上对成功发生的信令过程加上的时间标记，精确到秒
	LACID	网格编码
	EventID	事件类型(如开机、关机、打电话、发信息、上网或基站切换等)
网格信息	LACID	网格编码
	Lon	网格中心点位置的经度
	Lat	网格中心点位置的纬度

照时间进行排序,如果待修补数据的前后两条数据均完整,且需修补字段一致,则该条信息补充该字段且字段内容为前后两条相一致的内容。如果前后两条数据内容不一致,则根据时间差值进行判断,将与其时间差值小的一条记录的内容作为补充依据。

3) 有效停留信息的获取。

手机信令数据中存在乒乓效应、数据漂移等异常现象,需要处理这些现象以获取有效的轨迹点和停留信息。数据处理流程如下:

① 获取每个轨迹点的停留时长,即后一个轨迹点与前一个轨迹点出现的时间差被视为前一个轨迹点的停留时长;

② 删除停留时长小于5 s的位置;

③ 两次30度角修正,处理乒乓效应、数据漂移现象(对于同一用户的连续轨迹,若某一个轨迹点与其前一个出现的轨迹点以及后一个出现的轨迹点所构成的夹角小于30度,则删除该轨迹点,遍历所有轨迹点后再按照同样的方法重新遍历一次);

④ 得到修正后的有效轨迹信息以及停留时长信息。

2.2 多源数据融合处理

建立手机信令数据与问卷调查数据融合处理流程:

1) 通过所提供的志愿者手机号码,由通信运营商内部工作人员关联手机信令数据中相应的加密用户ID,建立问卷调查数据与手机信令数据之间的用户个体匹配和数据信息链接;

2) 以问卷调查数据采集的用户实际居住地坐标点位置为中心,利用GIS技术形成1 km边长的正方形缓冲区。通过对一段周期内的手机信令数据进行分析,获取关联用户在缓冲区范围内覆盖的各标准网格中心点的夜间停留时长特征,即统计每位用户在每日20:00—次日6:00缓冲区范围内各网格的停留时长,并将数据周期内累计夜间停留时间最长的网格作为其居住地真实位置的映射。

3) 以每日20:00—次日6:00为考察时段,委托通信运营商分别计算数据周期内志愿者居住地映射网格单日平均累积的停留时间以及出现的天数频次;此外,基于同样的数据,分别计算志愿者在考察时段内非居住地映射网格单日平均累积的停留时间以及出现的天数频次。

4) 根据步骤三的结果,将分别得到考

察时段内居住地映射网格的单日平均累积停留时间概率分布与出现的天数频次的概率分布,非居住地映射网格的单日平均累积停留时间概率分布与出现的天数频次的概率分布,以及居住地网格数量和考察时段内停留点总数等,此时所获取的概率分布中已经去除所有个体轨迹和隐私信息。

5) 根据步骤四中获得的各指标概率分布情况,进行朴素贝叶斯分类器模型的构建,并依次完成模型训练、测试和调试过程,最终应用到手机信令数据总体用户的居住地分布判断中。

3 朴素贝叶斯分类器

朴素贝叶斯分类器(Naïve Bayesian Classifier, NBC)是一系列在假设特征之间(朴素)相互独立的条件下运用贝叶斯定律的概率分类器。该模型具有算法表达形式简单、计算鲁棒性强并且应用过程中所需的存储资源要求少、时间成本小等特点,因此应用较为广泛,如风险链接的检测、文本的分类以及顾客营销资源分析等^[8]。与众多分类方法相比,朴素贝叶斯分类算法有如下特点:

1) NBC算法是基于贝叶斯概率的有监督学习分类算法,在理论层面上成熟完善。这种分类算法在数理上具有一定的优越性,分类结果的准确性较高^[9];

2) 可以处理各种数据类型的数据集,具有很强的算法适应性;

3) 算法结构较为简单,计算复杂度较低,可在开销较低的前提下处理数据量较大的集合^[8,10];

4) NBC算法是一种动态滚动的分类算法,随着有监督学习中训练样本的不断增加,训练过程可增量进行。

此外,相比于既有的基于阈值判断的手机信令数据处理算法,NBC算法在居住地人口分布识别中的应用同样具有以下优势:

1) 既有方法通常使用经验判断对上述阈值进行设定,对于计算结果缺乏标准化、科学化的判断依据。NBC算法属于有监督学习算法,通过训练和测试过程,能够一定程度上保证结果的准确性和可解释性。

2) NBC算法假设在给定类标记的条件下,参与判断的各特征属性是相互独立的。对于利用手机信令数据进行居住地人口识别情况,通常考虑居住地位置的单日停留时间

以及停留天数等特征,能够满足特征属性之间的独立性假设,因此算法适用性较强。

朴素贝叶斯分类器模型的基本原理为:假设获取的数据样本中任一对象都能够由 N 项属性特征量化描述,并且这些属性特征 $f_n(n=1,2,3,\dots,N)$ 之间都保持相互的独立性,可以用集合 F 对任一对象进行表示($F=\{f_1, f_2, \dots, f_N\}$)。假设获取的数据样本中任一对象都归属于某一个特定类别(例如居住地或非居住地),样本总体可以被划分为 M 个类别,分别为 $C_1, C_2, \dots, C_M$ 。现需要在已知某一对象的属性特征集合 F 的情况下,判断其隶属的对象类别 C (见图1)。

朴素贝叶斯分类器模型在分类识别问题上的解决方法是:利用未知类别对象的属性特征集合 F 与现有样本中所有已知类别对象的属性特征先验概率 $p(F|c)$,通过贝叶斯公式计算出新个体对象的后验概率 $p(c|F)$,即可以获得该未知类别对象属于某一特定类别的概率大小。该对象的最终分类识别结果由具有最大后验概率的类别所决定。算法的公式表达如下:

$$p(c|F) = \frac{1}{p(F)} p(c) p(F|c)$$

$$= \frac{1}{p(F)} p(c) \prod_{n=1}^N p(f_n|c), \quad (1)$$

$$\begin{cases} \hat{c}(F) = \arg \max_{c \in C} p(c|F) \\ C = \{c_1, c_2, \dots, c_M\} \end{cases} \quad (2)$$

在朴素贝叶斯分类器的具体实施过程中,将现有样本中已知类别的研究对象总体按照一定的比例分成两份。通过其中的一份样本数据对式(1)中 $p(c)$, $p(F)$ 以及 $p(f_n|c)$ 的概率分布情况进行模型参数的训练和估计;而第二份样本数据并不需要参与训练,其主要作用是测试训练后模型对其所属类别判断的准确性,并根据辨识结果对模型中的特征变量选取进行修正。最终,将待分类对象的属性特征集合 F 按照对应的条件概率分布分别带入式(2),即可得到其具有最大后验概率的隶属类别 C 。

4 案例分析

4.1 数据采集情况

为了验证上述方法的合理性和可行性,本文选取H省H市作为案例分析对象,利用上述提出的多源数据融合分析方法,获取居

住地人口分布现状特征。本次研究采集的手机信令数据由H省某通信运营商提供,包括H市市域范围内该运营商手机用户(包含漫游至该区域的手机用户)产生的手机信令数据。此次手机信令数据采集的时间周期为2018年9月1—30日(共30 d),经过数据预处理后得到手机用户的有效轨迹信息,继而提取用户停留点位置以及停留时长特征等。

问卷调查方面,于2018年9月14日完成调查任务,在H市市域范围内采用入户调查的方式共计发放问卷1200份。为了保证样本选取的合理性,根据每个街道的户数统计信息,按照比例分配问卷发放数量。问卷信息填写不完整、非指定通信运营商的手机用户以及在H市无固定/经常性居住的夜间居住地的问卷被视为无效问卷。最终,回收有效问卷895份,回收率74.6%。有效问卷中,18~65周岁受访者占84%,男女比例为1.02:1,与通信运营商提供的用户总体画像特征中的年龄结构相吻合,间接地说明了抽样的合理性。所有参与此次调查的相关受访者均被告知采集手机号码信息和居住地信息的用途,主动表明不会泄露或追踪其个体轨迹信息,并与同意参与此项研究的志愿受访者签订知情授权书。调查问卷主要包含以下五部分内容:个人社会属性信息(性别、年龄等);是否为指定通信运营商的手机用户(是/否);手机号码信息;当前在H市是否有固定/经常性居住的夜间居住地(是/否);夜间居住的家庭地址(精确到居住小区)。

4.2 数据融合及朴素贝叶斯分类器的构建

利用2.2节中提出的多源数据融合方法,建立有效调查问卷中的受访者与手机用户移动台识别码(Mobile Station Identifier, MSID)之间的精确匹配,并实现用户真实居住地理位置与基础网格之间的空间映射。因此,若以30 d数据周期为考察时间范围,以每日20:00—次日6:00为考察时段,针对每

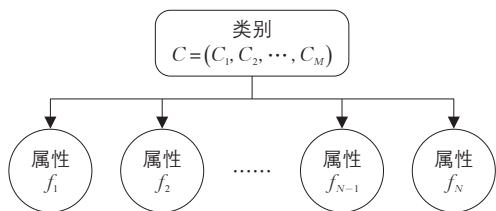


图1 朴素贝叶斯分类器结构

Fig.1 Structure of Naïve Bayesian Classifier (NBC)

位用户获取其在上述时间范围内的指标信息(见表2)。

其中, 以上述时间范围内用户停留的网格作为分析对象, 分别利用网格单日平均停留时长以及停留天数两个指标作为相互独立的基本属性 f_1 与 f_2 , 构建基本属性集合 F ; 以“是否为居住地网格”作为网格类别属性 C (是/否), 并建立集合 F 与类别属性 C 之间的贝叶斯概率关系, 训练得到朴素贝

叶斯分类器模型。经过测试验证后, 利用分类器模型, 对手机信令数据中计算得到的其他待分类用户的网格属性特征集合 F 进行应用, 即可得到其具有最大后验概率的“是否为居住地网格”的隶属类别 C 。

4.3 结果分析与应用

将895个匹配用户在30 d数据周期内(每日20:00—次日6:00)出现过的网格作为分析对象。其中, 属于居住地类别的网格为895个, 非居住地类别的网格为4 415个。为了构建有监督学习的训练集和测试集, 本次研究随机选择其中的530个居住地网格和3 100个非居住地网格, 共3 730个网格样本(包括网格真实的基本属性和类别属性)作为训练集, 约占总样本的70%。其余样本作为测试集, 包括居住地网格样本365个, 非居住地网格样本1 315个。经过训练集样本训练后的朴素贝叶斯模型被应用到测试集样本的类别预测中, 并将预测结果与真实结果进行比较。分别进行10次独立的随机分配试验, 10次试验结果的平均值如表3所示。

以居住地网格为例, 查准率(P)=预测正确值/居住地网格预测总值; 查全率(R)=预测正确值/居住地网格真实总值。通过表3的统计结果可以看出, 训练得到的分类器能够很好地辨识居住地网格。

上述结果反映了分类器对于居住地网格的预测准确性较高, 可以用于手机信令数据中待分类用户的居住地判断。此外, 为了进一步对比多源数据融合方法与传统阈值判断方法的分析结果, 本次研究还使用了多组阈值设置的传统阈值判断方法, 对同样的测试集样本进行居住地网格识别。其中, 对于传统方法中阈值参数组合的选取标准如下: 当地企事业单位工作日普遍的下班时间为17:30, 因此假设每天20:00—次日6:00为手机用户大概率出现在居住地位置的时段。考虑到手机用户存在下班后其他非基家活动出行, 这一情况将会影响手机用户在居住地位置的连续停留时间。因此, 假设手机用户当日在上述时段内连续停留4 h以上的网格位置即作为候选居住地网格。随后, 统计各候选居住地网格在30 d数据周期内符合上述要求的天数, 分别按照数据周期的1/3, 1/2和2/3设置三档天数阈值, 依次进行居住地网格的识别, 结果如表4所示。由于天数阈值设置的

表2 基于多源数据分析的用户指标信息

Tab.2 User metrics information based on multi-dimensional data

用户ID	网格编号	单日平均停留时长/h	停留天数/d	是否为居住地映射网格
100001	2354612	6.3	21	是
100001	5241321	0.4	12	否
100001	2334315	1.3	3	否
100002	2413523	7.3	17	是
.....				
100895	2134155	2.3	6	否
100895	2312311	8.1	19	是

表3 结果准确性分析

Tab.3 Analysis of the results accuracy

项目	预测值		查全率/%
	居住地网格	非居住地网格	
真实值	居住地网格	324	41
	非居住地网格	53	1 262
查准率/%		86	97

表4 传统阈值判断方法的分析结果

Tab.4 Analysis results of traditional threshold judgment methods

项目	预测值		阈值判断标准
	居住地网格	非居住地网格	
真实值	居住地网格	296	69
	非居住地网格	145	1 170
	居住地网格	219	146
	非居住地网格	89	1 226
	居住地网格	156	209
	非居住地网格	4	1 311

注: 阈值参数组合1, 候选居住地网格需要满足的条件为, 30 d数据周期内每日20:00—次日6:00连续停留4 h以上的天数不少于10 d, 若出现多个候选居住地网格, 则选择满足上述条件的天数最多的网格作为居住地网格; 阈值参数组合2, 需要满足的天数条件为不少于15 d, 其他条件同上; 阈值参数组合3, 需要满足的天数条件为不少于20 d, 其他条件同上。

要求越高,识别出的居住地网格查准率越高,但查全率会下降,因此传统阈值判断方法的精度介于最低档阈值参数组合的判断结果与最高档阈值参数组合判断结果之间。

随着需要满足的阈值条件要求的不断提高,能够被判断为居住地网格的数量也随之下降(测试集样本的预测结果中分别为441个,308个和160个网格被识别为居住地网格,见表4)。阈值参数组合1~3所对应的居住地网格查准率分别为67.1%,71.1%和97.5%。虽然第三种阈值参数场景下的查准率97.5%比本文提出的有监督学习方法的测试结果有所提升,但这一结果却牺牲了大量的查全率(阈值参数组合1~3所对应的查全率分别为81.1%,60.0%和42.7%)。由于查全率同样反映了模型对真实居住地的辨识效率,因此在综合考虑查全率和查准率的情况下,本文提出的朴素贝叶斯模型优于传统的阈值判断方法,查准率达86%,查全率达89%。

利用所提出的模型方法对H市的居住人口密度分布情况进行分析,结果如图2所示。H市人口分布主要集中于中心城区范围,配合人工经验判断对各交通小区的人口分布结果进行了抽样校核,基本符合现状特征。

5 结语

本文基于手机信令数据和问卷调查数据的多源数据融合手段,提出一种辨识城市居住地人口分布特征的有监督学习方法。通过两种数据的宏微观链接,建立真实居住地与手机信令位置数据的映射,并以此为基础进行贝叶斯模型的构建、训练以及测试,最终应用于辨识居住地人口分布特征。通过案例分析,验证了模型的可操作性。分析结果表明:基于有监督学习方法的人口分布辨识方法较基于阈值参数判断的传统辨别方法的精度有明显提升,此方法也可应用于工作地岗位分布特征的分析。

参考文献:

References:

- [1] 杨朗,周丽娜,张晓明.基于手机信令数据的广州市职住空间特征及其发展模式探究

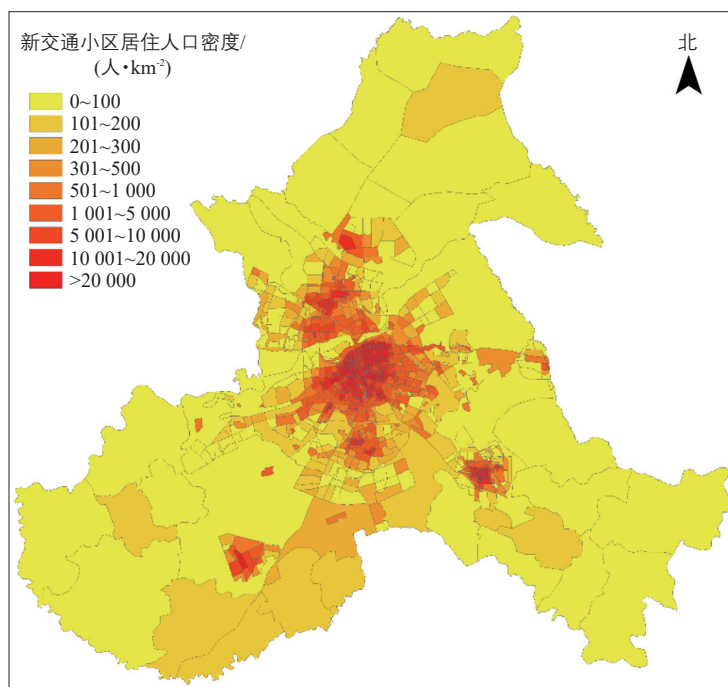


图2 基于手机数据分析的居住人口密度的空间分布结果

Fig.2 Spatial distribution results of residential population density based on cellular data

[J]. 城市观察, 2019(3): 87-96.

Yang Lang, Zhou Lina, Zhang Xiaoming. Research and Evaluation of Jobs-Housing Space Characteristics Based on Mobile Phone Signaling Data: A Case Study of Guangzhou[J]. Urban Insight, 2019(3): 87-96.

- [2] 汪瑞琪, 陈建均. 手机大数据在城市居民出行特征分析的应用[J]. 智能城市, 2019, 5 (15): 151-153.

- [3] 陆振波, 龙振, 余启航. 基于手机信令数据的昆山市职住分布与通勤特征分析[J]. 现代城市研究, 2019(3): 50-55.

Lu Zhenbo, Long Zhen, Yu Qihang. Analysis on the Job- Housing Spatial Distribution and Commuting Characteristics of Kunshan City Based on Cellular Signaling Data[J]. Modern Urban Research, 2019(3): 50-55.

- [4] 邹伟, 陈鑫, 秦战. 手机信令数据辅助下的张江科学城职住分析及对策[J]. 上海城市规划, 2019(1): 78-82.

Zou Wei, Chen Xin, Qin Zhan. Analysis and Countermeasure of Jobs-Housing Relationship on Zhangjiang Science City Under Mobile Phone Data[J]. Shanghai Urban Planning Review, 2019(1): 78-82.

(下转第12页)