大数据环境下上海市综合交通特征分析

摘要：20 世纪 80 年代上海市已经开始建设智能交通系统，开展交通数据采集工作。历经 2010 年世博会，上海市交通信息化快速发展，交通大数据的种类越来越多，为基于大数据挖掘的综合交通调查与分析提供了良好的基础条件。基于 2014 年上海市第五次综合交通调查结果，综述上海市交通大数据资源现状和基于大数据的城市综合交通特征挖掘分析技术方法及主要成果。提出进一步拓展运用大数据分析城市综合交通特征的应用领域、优化大数据挖掘技术方法和完善大数据采集处理机制的相关建议。

关键词：综合交通，大数据，交通调查，技术方法，上海市

随着道路感应线圈、出租汽车 GPS、手机信令、轨道交通自动售检票等道路交通自动采集技术的逐渐成熟，利用自动采集数据挖掘城市交通特征已成为一种趋势。2010 年世博会后，上海市交通信息化快速发展，城市交通大数据的来源更多样、类型更丰富、数据量更巨大，包括车牌识别数据、土地利用遥感影像数据、移动通信数据等。基于大数据分析城市综合交通特征的资源条件日益成熟。2014 年，上海市开展的第五次综合交通调查已经广泛应用了新信息技术调查手段。新调查手段不仅继承了传统信息挖掘技术，还着重对新增信息化资源展开挖掘，有效弥补了传统调查手段的不足，已在某些领域成为主要调查手段。

1　交通大数据主要资源

上海市交通大数据来源于道路交通、公共交通等交通领域和移动通信、土地利用等相关领域。

1）全市用地数据。主要为约 23 万个用地单元的用地遥感数据（见图 1）和房屋建筑量统计数据库。遥感数据包括高分辨率航空遥感数据、卫星遥感影像、全市分类土地利用数据库。房屋建筑量统计数据包括单体建筑名称、占地面积、层数、坐落地址、房屋类型等建筑属性信息。这些数据主要用来支撑城市土地利用性质、开发强度分析等多种应用。

2）移动通信数据。主要指上海市域调查时段内出现过的移动手机用户（包括本地及漫游）信令数据，包括短信、通话、LAC 区（位置区，通常包含多个基站蜂窝小区）切换或每隔 1～2 h 定时与基站通讯的记录。经检测，出现在上海市域的日均手机用户规模约 1 800 万个，平均每个用户一天的轨迹点记录约为 60～70 条。这些数据主要支撑人口、职住分布、潮汐交通特征分析等多种应用。

3）车牌识别数据。上海市车牌识别系统覆盖全部 44 个市境出入通道和 343 个中心城快速路主要断面，数据内容包括车辆号牌编码、牌照类型、途经时间、途经车速、车辆属地及设备断面编号等，主要支撑出入市境、中心城快速路的车辆使用特征分析等多种应用。

4）高速公路收费流水数据。覆盖上海市域全部 104 个主线收费站和进出匝道收费站。数据内容包括驶入驶离收费站编号、车型、时间、流量等，主要支撑高速公路车辆使用特征分析等多种应用。

5）运营车辆 GPS 数据。包括约 2.9 万辆出租汽车、1 万辆集装箱卡车及普通货车的 GPS 数据。数据内容包括回报轨迹点位置坐标、车速、空重车状态（出租汽车）等，主要支撑对道路运行车速、出租汽车和货车出行特征分析等多种应用。

6）轨道交通自动售检票系统和交通卡自动刷卡计费系统数据。前者覆盖轨道交通全网进站、出站闸机的刷卡数据，数据内容包括进站和出站的车站名称、时间、乘客数量等，支撑对轨道交通系统客流分析等多种应用。后者全市日均约有 400 万张、1

000 万次刷卡数据，交通方式覆盖轨道交通、公共汽（电）车、出租汽车及轮渡，数据内容包括刷卡线路、刷卡时间、刷卡金额等，主要支撑对公共交通运行及乘客换乘特征分析等多种应用。

2 大数据挖掘技术在综合交通调查中的应用

数据挖掘技术在历次综合交通调查中都有应用，但第五次综合交通调查在以往调查和日常工作经验的基础上，充分利用了上海市长期积累的交通大数据资源，特别是更加广泛地应用了城市用地数据、移动通信数据，并新增了基于车牌识别系统数据挖掘的调查。交通大数据挖掘不仅在调查内容上与传统调查衔接，也为交通模型进行多样数据校核提供辅助。

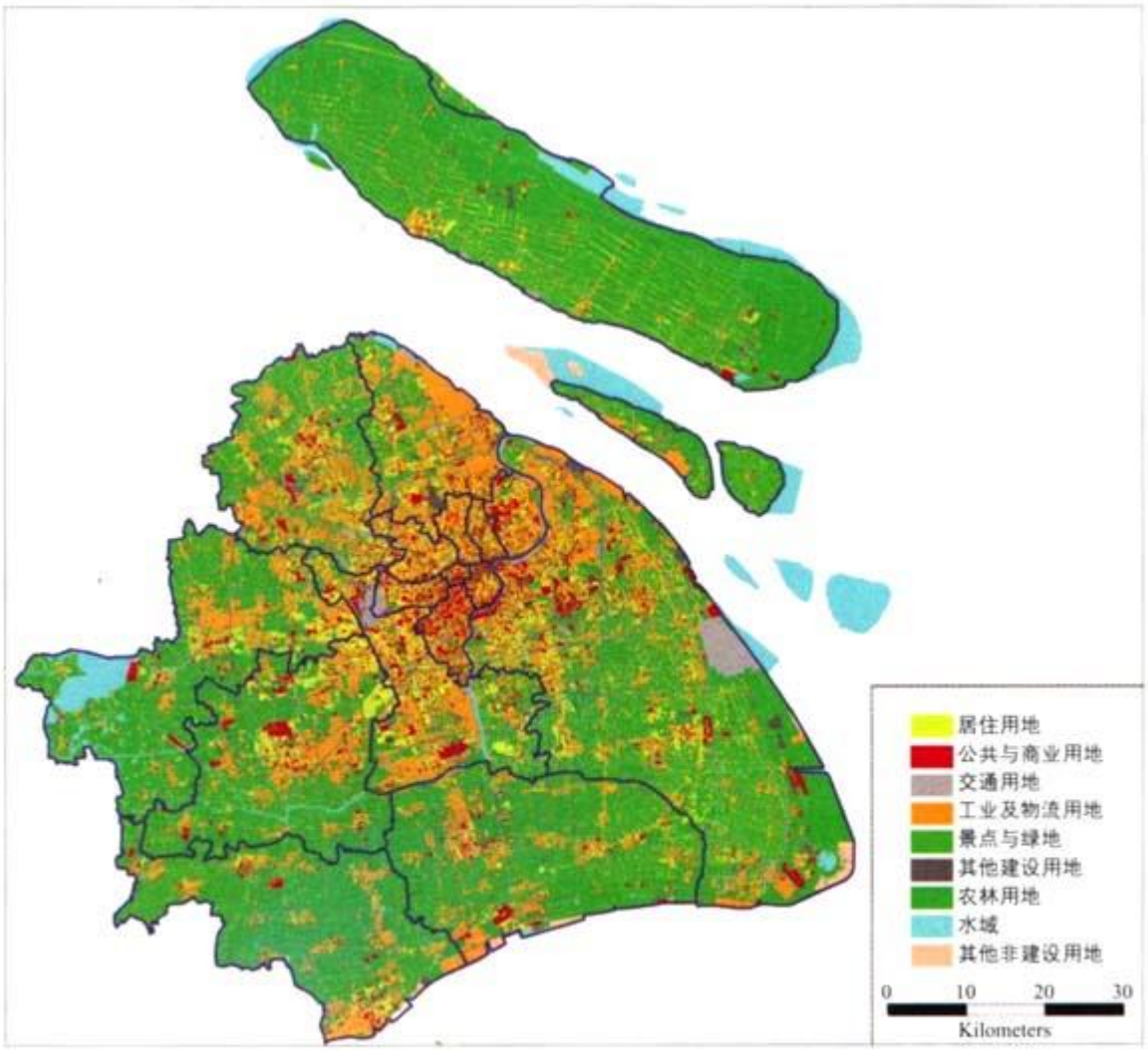


图1 2013年上海市分类遥感用地类型

Fig.1 2013 Shanghai land use patterns based on Remote Sensing Image

资料来源：文献[1]。

本次交通大数据挖掘对每项数据源均进行了原始数据质量分析和清洗工作，以保证数据挖掘结果的准确性，同时将挖掘结果与经其他渠道统计调查所获得的数据进行比对，也为验证数据挖掘结果的可靠性、固化数据挖掘技术方法提供了有力支撑。例如，在车牌识别数据挖掘方面，将市境出入道路关口识别数据与收费数据比对，将长期在沪使用外地号牌小汽车规模挖掘成

果与居民出行家访调查、夜间停车调查、在沪购买交通强制险记录数据进行比对；在移动通信数据挖掘方面，将夜间用户分布结构与常住人口普查数据比对，将轨道交通系统内部乘客换乘特征挖掘成果与轨道交通闸机及运营统计换乘量和线路乘距数据比对等。经检验，在城市用地、长期驻沪外地牌照小客车总量、轨道交通系统内部客流换乘特征、道路交通运行车速等方面，大数据挖掘可以成为调查的主要手段；在分析居民出行方式链特征、潮汐交通特征、职住分布特征等方面，大数据挖掘是辅助调查的重要校核手段。

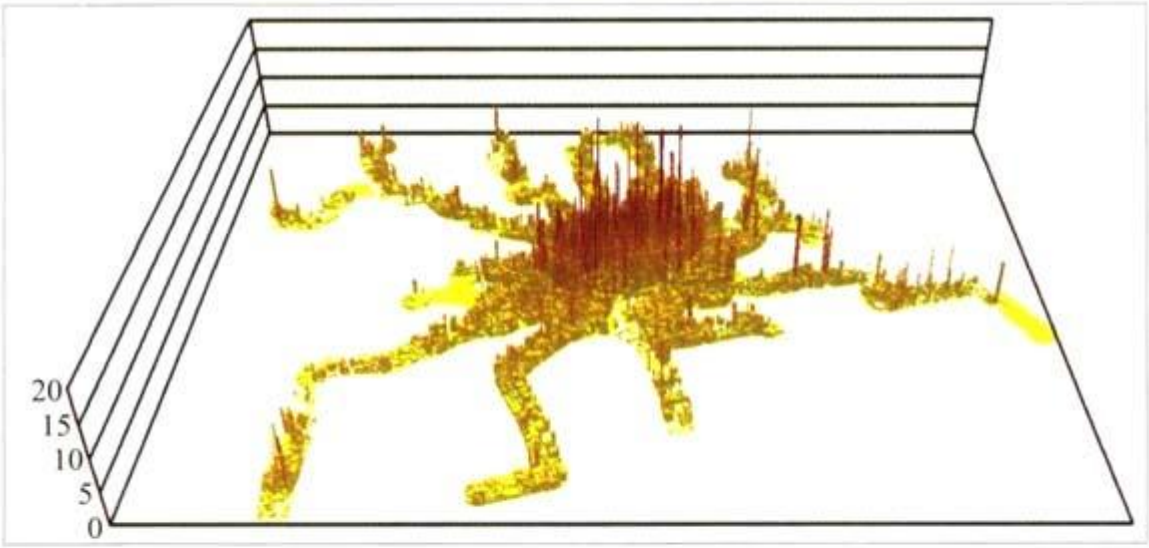


图2 2013年轨道交通沿线1 km范围内建筑容积率

Fig.2 Floor area ratio within 1 km buffer area along railway lines in 2013

资料来源：文献[1]。

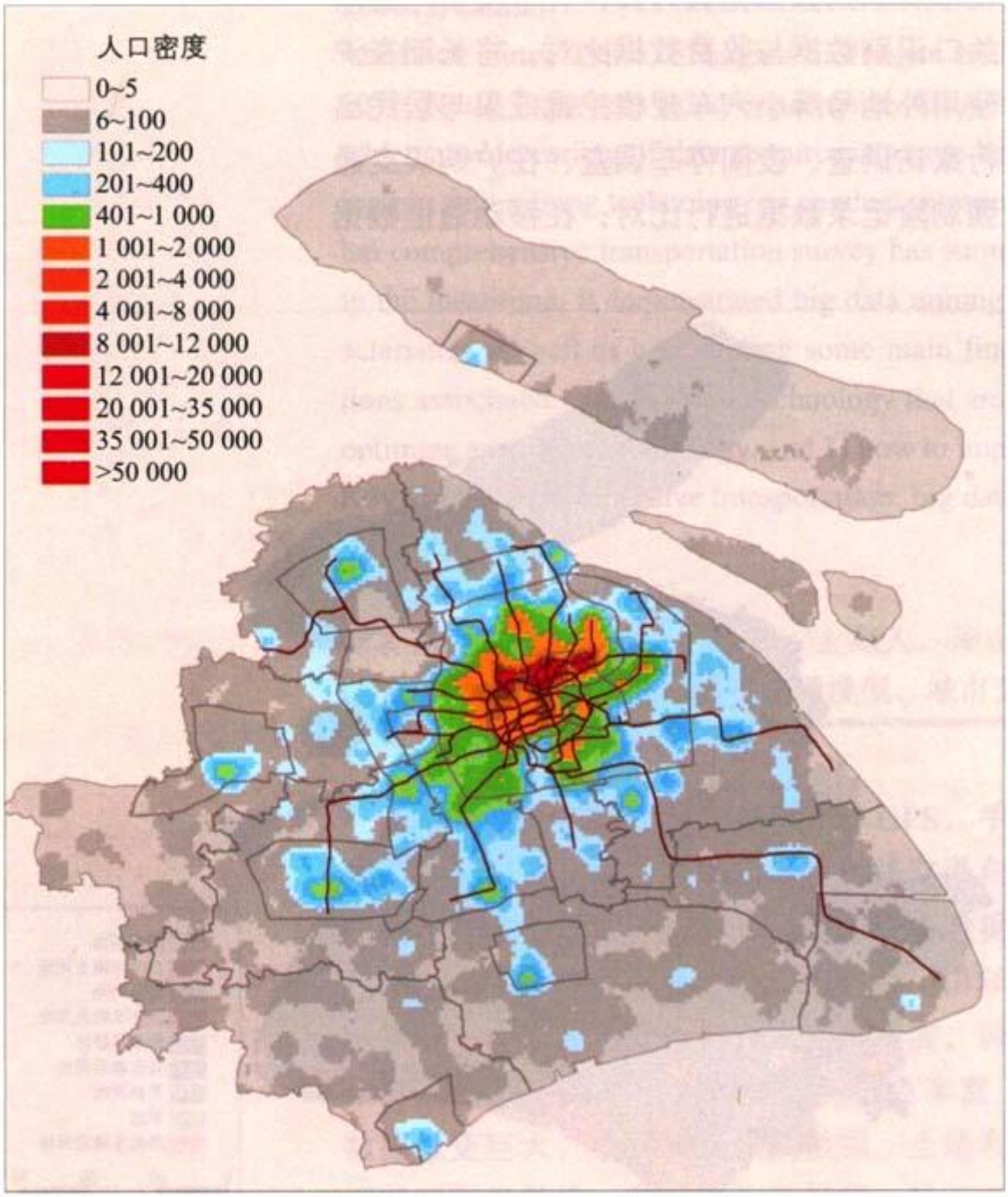


图3 2013年手机用户夜间分布

Fig.3 Distribution of mobile phone night users in 2013

资料来源：文献[2]。

2.1 利用遥感影像数据挖掘城市用地特征

通过内业解译与外业采样核对相结合，利用全市用地遥感影像和房屋建筑量统计数据库获得全市用地分类和建筑量分布，是土地利用信息获取的主要途径。用地数据是交通调查中不可或缺的基础资料，通过对交通相关用地的遥感数据进行调查挖掘，

可以掌握最新城市用地布局、土地开发强度及变化趋势，为评估用地总量与布局的合理性、优化完善交通设施布局提供定量依据。利用上海市市域红外航空遥感影像，按照规划用地分类标准，将用地类型解译细分至 23 万个用地单元（见图 2）。2013 年底全市建设用地 2 913 km^2，较 2008 年增长 8%。全市建筑量 12 亿 m^2，中心城区建筑量 5.7 亿 m^2。轨道交通对城市用地发展具有一定的引导作用，轨道交通车站 2 km 半径范围内容积率高于其他地区。以轨道交通 1 号线为例，中心城区、近郊区车站 2 km 范围内平均容积率分别为 1.17 和 0.51，其他区域平均容积率仅为 0.57 和 0.36[1]。

2.2　利用移动通信数据分析人口分布

通过跟踪手机用户的移动通信数据，分析日间、夜间手机用户分布规律性特征，是获得调查期实有人口分布的重要校核数据。这项调查技术曾尝试用于近年一些小样本量调查，但应用于全市性综合交通调查尚属首次，是弥补调查期实际人口统计误差和辅助校核就业岗位分布的重要手段。以一定时间窗内上海市域移动通信用户手机信号出现天数、累计出现时长以及通信信号出现和消失时间作为判断标准，研究特定区域内手机用户信号在不同时段出现的规律性特征，进而挖掘日间、夜间手机用户的分布情况。据分析，夜间手机用户分布密度由内向外逐渐降低，浦西内环线北段人口最为密集（见图 3）。另外，中心区日间固定出现的手机用户规模明显高于夜间，两者比值达 1.2～1.4，静安区尤为明显，日间、夜间手机用户比值达 1.5（见图 4）。

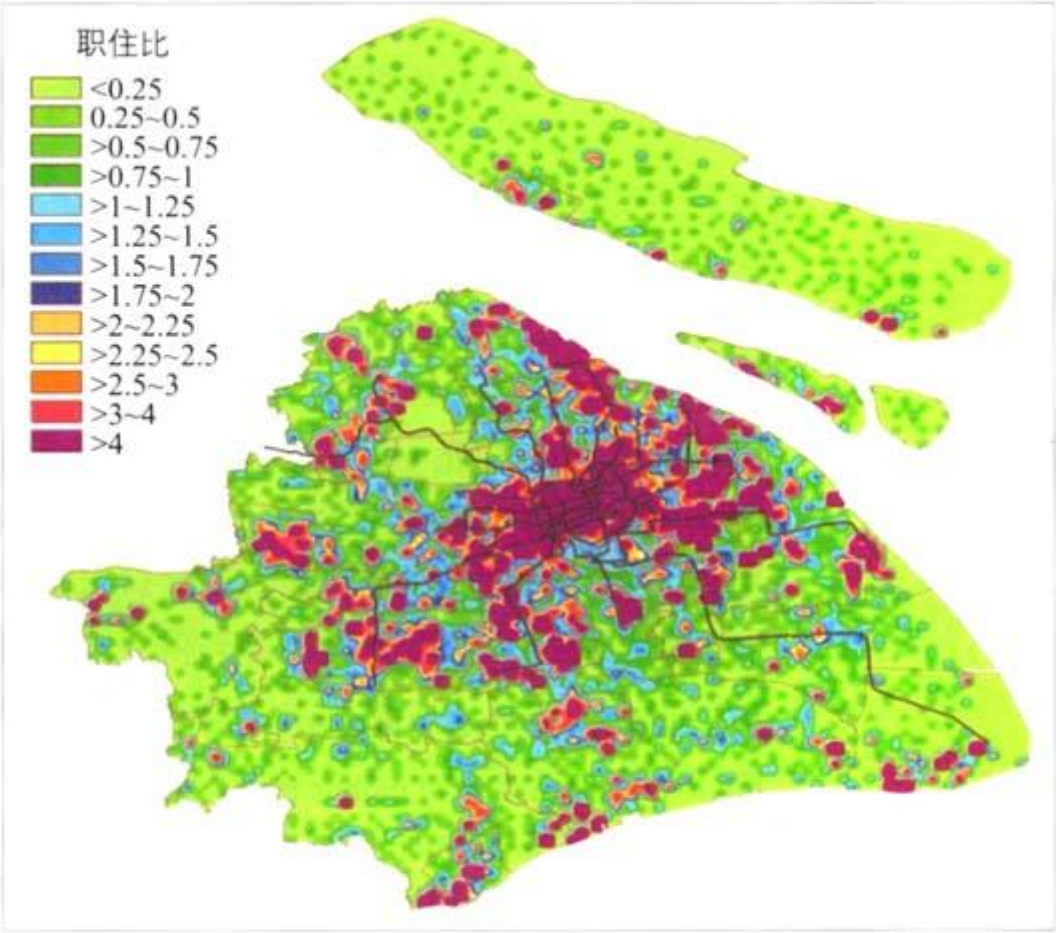
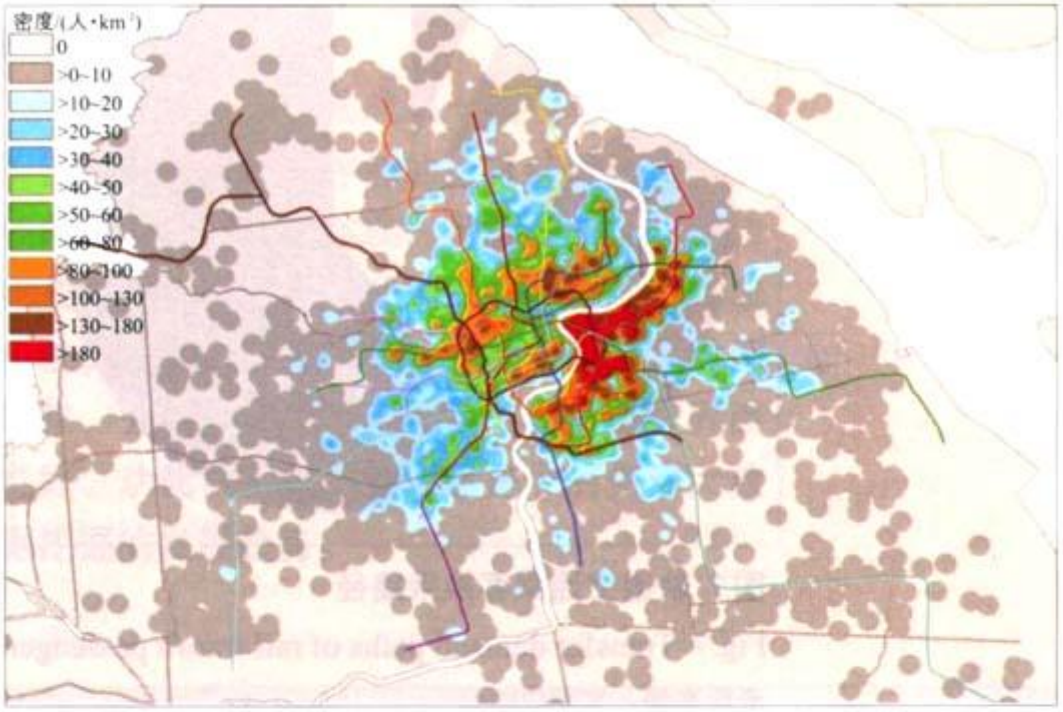


图4 2013年手机用户职住比分布

Fig.4 Distribution of job-housing rate for mobile phone users in 2013

资料来源：文献[2]。



a 陆家嘴

b 徐家汇

图5 特定区域工作人口居住地分布

Fig.5 Residential location distribution of employees at specific areas

资料来源：文献[2]。

2.3　利用移动通信数据分析通勤与潮汐交通特征

在判断移动通信用户日间、夜间分布规律的基础上，剔除日间、夜间长时间停留在同一地点的用户，进而推断居民通勤出行空间分布特征；同步考虑基站地理空间距离、实际路径距离等多重因素，获取职住出行距离特征指标。据分析，上海市平均通勤出行距离约为 8.5 km。郊区进出中心城区的通勤量占全市通勤出行总量的 12%，其中近 80%通勤量来自近郊区与中心城区之间。工作人口居住特征以陆家嘴和徐家汇地区为例，据统计约 90%的陆家嘴地区工作人口居住在中心城区，其中居住在浦东的略高于浦西，且主要集中在轨道交通 6 号线沿线（见图 5a）；约 80%的徐家汇地区工作人口居住在中心城区，其中近 90%居住在浦西（见图 5b）。

利用移动通信数据获取用户出行轨迹是反映人员全方式出行空间分布特征的重要数据。传统调查一般从系统流量反映某一种方式的潮汐交通特征，如轨道交通客流、道路交通车流等，利用移动通信数据可以反映全方式潮汐交通特征。以内环线为例，工作日早高峰时段进出中心城区断面的交通需求很不均衡，进出比约为 1.7（见图 6）。

2.4　利用移动通信数据分析轨道交通乘客换乘特征

利用地面和地下、不同轨道交通线路所在移动通信 LAC 区编码的唯一性和手机在跨越 LAC 区基站时必然会发生位置信令更新的特性，基于基站和地铁车站的对应关系，可以获得乘客在轨道交通系统内部的路径信息。以往主要通过在轨道交通车站进行人工问询的抽样调查方法，本次调查首次采用移动通信数据进行信息采集，完善了轨道交通系统模型分配算法的重要数据资源，使得移动通信数据作为获得乘客真实换乘路径的主要手段成为现实。基于上述数据也可获得不同车站间乘客的路径分布，以及换乘站分方向、分时段的客流量。经计算，使用地下轨道交通的移动通信用户中，只有单一路径的用户和有多路径选择的用户比例约为 6：4（见图 7）。在多路径选择情景下，约 70%的乘客还是以选择最短路径为主。此外，乘坐轨道交通的乘客中，约 50%无须换乘，44%仅需换乘一次（见图 8）。

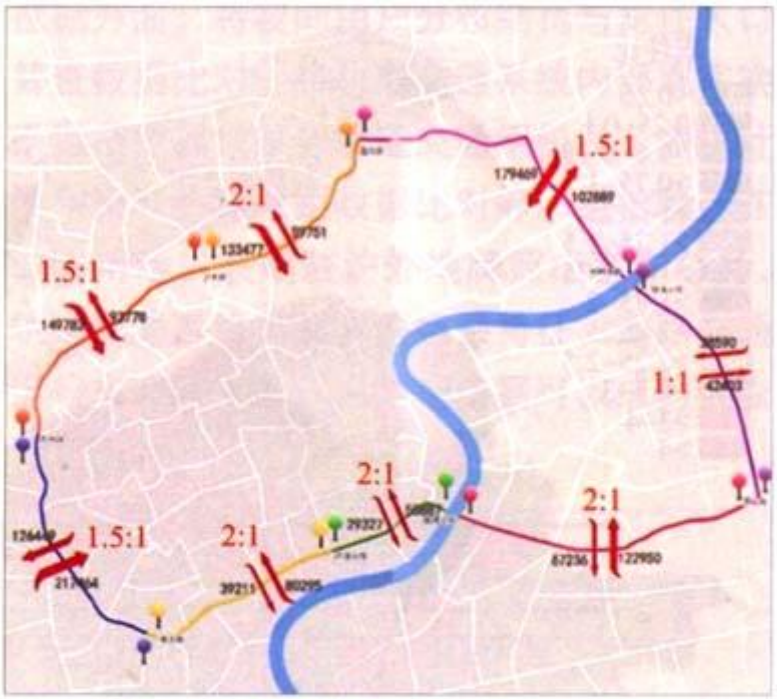


图6 早高峰时段内环各断面手机用户进出比值

Fig.6 Enter-exit ratio of mobile phone users on the inner ring road in the morning peak period

资料来源：文献[3]。

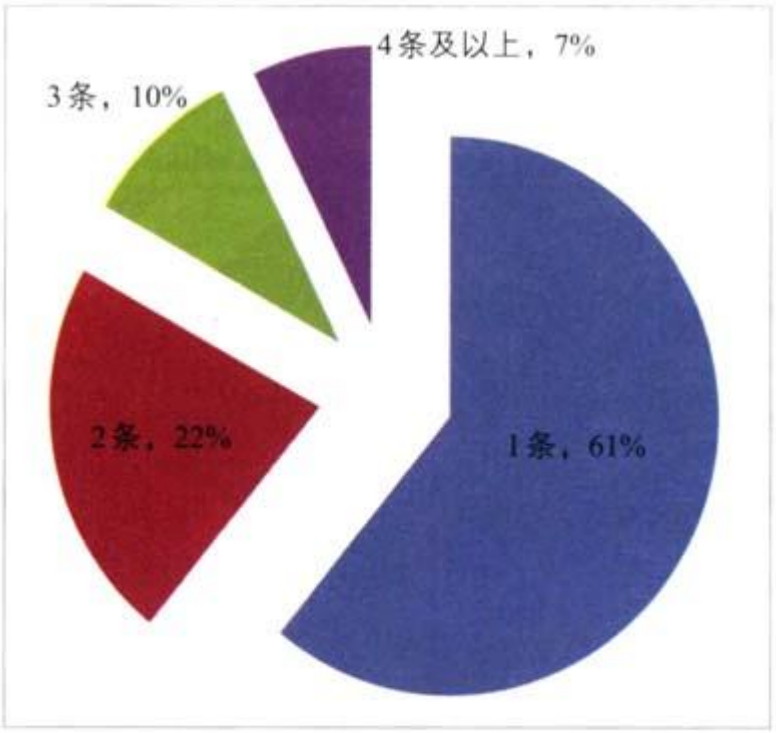


图7 轨道交通乘客换乘路径

Fig.7 Transfer demand paths of rail transit passengers

资料来源：文献[2]。

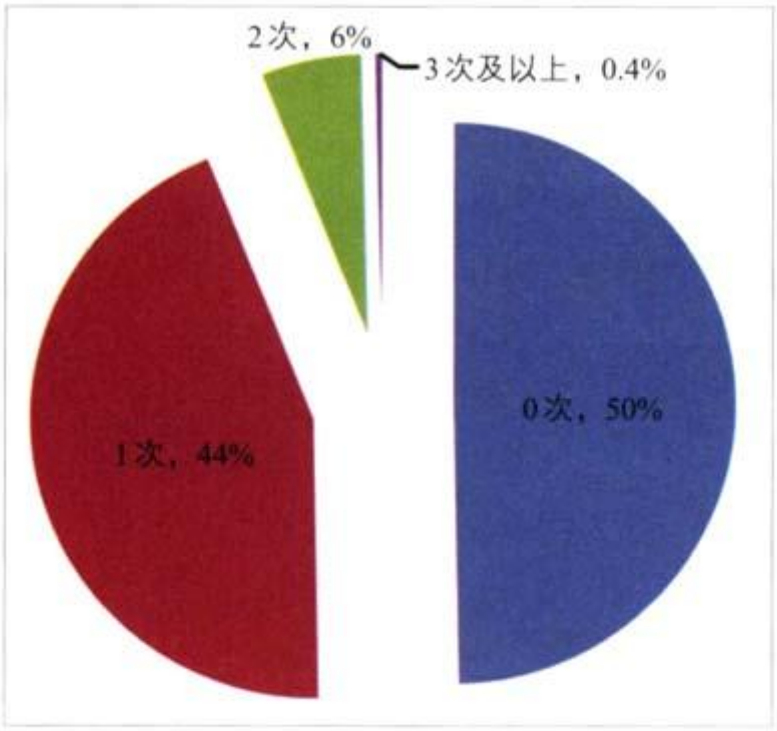


图8 轨道交通乘客换乘次数

Fig.8 Transfer times of rail transit passengers

资料来源：文献[2]。

2.5 利用牌照识别数据分析小汽车总量与快速路车辆特征

长期在沪外地牌照小汽车总量调查一直是历次调查的难点，夜间停车调查、居民出行家访调查等传统调查手段实施难度大、调查成本高。本次调查首次利用车牌识别数据获取长期在上海市使用的外地牌照小汽车总量。通过与夜间停放车调查数据、居民出行调查数据和外地车辆在沪保险数据相比对，验证了利用车牌识别数据作为外地牌照小汽车总量调查的可靠性。今后可利用牌照数据资源，捕捉有进出市境道路关口记录和使用过中心城区快速路的外地号牌小汽车，综合考虑累计在沪及单次在沪停留时间，判断是否属于长期在沪使用的外地号牌小汽车。经计算，2014 年上海市实有小汽车约 320 万辆，其中外地号牌小汽车近 100 万辆。

基于牌照识别设备的位置特点，结合高速公路收费数据，本次调查首次以车牌识别数据取代市境道路关口人工问询，调查进出市境车辆在上海市域内的出行分布特征。经计算，入境车辆进入上海市郊区和市区的比例约为 7：3，其中进入中心城区和外围城区的比例约为 2：8。

此外，本次调查还首次利用牌照数据调查中心城快速路系统的车辆使用特征，获取中心城区快速路各类车辆的驶入量、使用频率、行驶距离等特征指标，不同路段车辆的牌照结构以及车辆的流向分布（见图 9）。经计算，工作日中心城区快速路（不含外环线）全日驶入车辆中 25%为外地牌照，午间平峰外地牌照车辆比例达 32%；其中，中环线比例最高，午间平峰外地牌照车辆占驶入车辆总量的 40%。

3 大数据挖掘技术应用展望

3.1　拓展应用领域

1）利用互联网数据分析城市人口结构特征。

利用移动通信数据处理技术获取城市人口分布特征已成为现实。然而若要分析交通源出行特征，还需要掌握城市人口结构特征，该数据很难通过人口普查外的其他手段获取。多源数据融合分析将是一种必然选择。例如，利用线上购物、网站访问、电视收视等数据综合分析不同区域的家庭成员结构等。

2）利用移动通信数据分析居民出行特征。

在当前社会经济条件下，传统人工调查的样本规模和实施难度均有所增加。随着 3G，4G 移动通信数据的引入，用户的轨迹点信息将大幅增加，这为分析用户出行强度、出行空间分布及出行路径特征创造了有利条件。

3）利用车牌识别数据分析全市车辆出行分布特征。

随着地面高清卡口车牌识别数据的引入，车牌识别数据覆盖的空间范围将从市境出入道路关口和中心城区快速路断面拓展至行政区边界的道路。通过扩大地面车牌识别数据的采集可进一步掌握跨越行政区车辆出行分布特征。

4）利用车载 GPS 和公交 IC 卡刷卡关联数据分析公共汽（电）车线路客流 OD 特征。

公交 IC 卡刷卡数据仅能反映刷卡时间和刷卡线路信息，无法获得上下客车站信息。由于大部分公共汽（电）车都已安装了 GPS 系统，可将 GPS 数据和公交 IC 卡数据关联起来，再通过公交 IC 卡刷卡时间与公共汽（电）车 GPS 轨迹点回报时间的对应，获得公交 IC 卡刷卡的地点信息。最后通过 GPS 轨迹点与公共汽（电）车站的空间对应关系，分析公共汽（电）车线路客流的上客特征。在此基础上，通过对一段时间内公交 IC 卡刷卡数据的分析，获得乘客上客的规律性特征，利用一天内乘客本次出行的下客车站很有可能是下次回程的上客车站的一般规律，进一步分析公共汽（电）车线路客流 OD 分布特征。

3.2　优化技术方法

1）优化数据挖掘技术，增强交通大数据关联性分析。

优化、改进已有大数据挖掘技术方法，包括车辆 GPS 数据在高架道路、地面道路重叠区域的识别方法，移动通信数据一次出行的判断方法，移动通信数据交通方式的判断方法等。此外，交通大数据挖掘的价值不仅限于对单源数据的分析，各类数据的关联性分析可进一步提升交通大数据挖掘的应用价值，例如城市用地与移动通信数据的关联分析等。

2）形成一整套关于交通大数据分析与扩样的系统流程。

有些信息数据可以直接反映母体特征，例如中心城区快速路车牌识别数据。有些信息数据虽然数据量巨大，但本质上仍然属于抽样调查数据，例如利用移动通信数据分析城市人口分布，如何进行科学扩样是大数据挖掘需要解决的另一技术难点。根据本次调查经验，受移动用户市场占有率、手机持有率等多重因素的区域差异性影响，用同一系数对全市数据进行统一扩样可能会产生较大误差。目前移动通信数据以反映结构性特征指标为主，下一步需按区域确定扩样系数，以使移动通信数据分析结果可以反映总量特征。

3）丰富成果展现形式，形成大数据可视化产品。

传统大数据挖掘成果一般是基于信息分析技术的一整套程序包的运算结果，通常以成果数据库的形式体现，再进一步由人工转化成可视化图表，处理周期较长。如何突破传统调查成果的展示模式，基于固化的数据采集、分析处理、成果展示流程，形成界面友好、功能实用的软件产品是今后交通大数据挖掘的发展趋势。

图9 中心城区快速路本地牌照和外地牌照车流分布

Fig.9 Distribution of traffic volumes with local and non-local plates on expressways in the central district

资料来源：文献[4]。

3.3 完善长效机制

1）完善交通大数据的采集汇总机制。

上海市交通大数据资源已基本汇集到上海市交通综合信息平台。随着信息化快速发展，交通大数据资源种类日益多样化，仍然需要积极扩展信息资源获取渠道，形成数据采集汇总机制及统一的数据仓库，为交通大数据的分析挖掘提供基础。

2）建立交通大数据分层挖掘分析机制。

根据数据需求对交通大数据的挖掘进行分层设计，形成每日、月度、季度、年度的定期分析及结合重大节假日、重大事件及交通热点问题的不定期分析机制。例如，一些反映系统运行、计算周期短的分析指标可以进行每日数据挖掘，以满足日常运营管理需要；一些反映交通源出行特征、计算周期及时间长且通常需要累计数据进行计算的指标可以进行季度或年度数据挖掘，主要应用于反映交通运行规律及发展趋势等研究。

参考文献：

[1] 上海市城乡建设和交通发展研究院.上海市第五次综合交通调查：基于遥感技术的交通相关用地数据挖掘[R].上海：上海市城乡建设和交通发展研究院，2014.

[2] 上海市城乡建设和交通发展研究院.上海市第五次综合交通调查：基于手机信息的出行特征调查[R].上海：上海市城乡建设和交通发展研究院，2014.

[3] 上海市城乡建设和交通发展研究院.上海市第五次综合交通调查总体方案[R].上海：上海市城乡建设和交通发展研究院，2013.

[4] 上海市城乡建设和交通发展研究院.上海市第五次综合交通调查：基于牌照识别的车辆出行特征挖掘[R].上海：上海市城乡建设和交通发展研究院，2014.

[5] 何承，朱扬勇.城市交通大数据[M].上海：上海科技出版社，2015.

作者简介：陈欢（1981—），女，上海人，硕士，高级工程师，注册城市规划师，注册咨询工程师（投资），主要研究方向：城市交通模型、城市交通规划。E-mail：cathleen.ch@163.com