面向人口洞察的手机信令数据产品质量校验方法研究

Research on Quality Verification Method of Mobile Signaling Data Products for Population Insight

梁　洁，张　岩，闫　嘉（中国联通智慧足迹数据科技有限公司，北京 100032）
Liang Jie，Zhang Yan，Yan Jia（China Unicom Smart Steps Digital Technology Co.，Ltd.，Beijing 100032，China）

摘　要：
随着数字化转型发展，基于脱敏后的手机信令衍生的数据产品也更加多样化，为解决其数据质量和准确性问题，提出了针对数据采集、数据处理与建模分析3个阶段的数据产品质量校验方法和质检流程，制定了17项质检关键指标，并通过案例进行了分析。结果表明，该方法可以实现用既定的规则代替主观判断，进而提高数据产品的交付质量。

关键词：
人口洞察；手机信令；数据产品；质检指标
doi：10.12045/j.issn.1007-3043.2020.11.0013
文章编号：1007-3043（2020）11-0065-06
中图分类号：TN919
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
With the development of digital transformation，the data products derived from desensitized mobile signaling are more diversified. In order to solve the problem of data quality and accuracy，it puts forward the data product quality verification method and quality inspection process in three stages of data collection，data processing and modeling analysis，and formulates 17 key quality inspection indicators，and analyzes them through cases. The result shows that the method can replace the subjective judgment with established rules to improve the delivery quality of data products.

Keywords：
Population insight；Cellphone signaling；Data product；Quality verification index

引用格式：梁洁，张岩，闫嘉. 面向人口洞察的手机信令数据产品质量校验方法研究[J]. 邮电设计技术，2020（11）：65-70.

0　前言

目前，社会经济形态经历了从传统经济到互联网经济，再到数字经济的演变。2018年，我国数字经济规模达到31.3万亿，占GDP比重为34.8%，数字经济在推动经济高质量发展中的战略地位和引擎作用不断凸显。同时，以网络化、信息化与智能化的深度融合为核心的第四次工业革命不断深化，全球逐渐进入以“万物互联”为显著特征的数字化时代。数据日益成为推动数字化时代发展的重要驱动力，如何最大化数据价值是当前数据服务商面临的新课题、新机遇、新挑战。

来自中国联通等移动通信运营商的手机信令数据是一种大规模采样、脱敏的移动位置数据，是手机用户使用移动通信网时留下的时空轨迹。手机信令数据的主要特点有：

a）手机的普及率高，根据工业和信息化部发布的电信业多项数据显示，截至2018年底，全国移动电话用户总数达到15.7亿户，人均拥有1.12张手机卡。

b）手机数据具有实时性，能连续记录居民活动的时空变化，无论是主动还是被动行为，在运营商网络内都会留下记录。

c）手机数据采集成本低，易于连续多日采集，方便挖掘居民多日行为的一般特征和活动规律。

收稿日期：2020-09-01

d) 被调查者不能干预手机信令数据实时采集,数据更为客观和有效。

目前政府机关、高校、科研机构和大数据企业已经利用手机信令数据开展了大量研究并衍生出各种类别的数据产品,例如利用手机信令数据能够有效把握城乡居民的行为轨迹、城市空间利用现状、交通运行现状、公共设施服务水平等,从而用于城镇体系等级结构、城市空间结构、城市中心体系、职住平衡、商圈活力、城市交通等方面的研究。

手机信令数据产品多种多样,其数据质量和准确性是产生价值的核心要素。但是数据采集存在不稳定性,底层数据清洗、处理过程参数多,数据建模阶段定制化需求多,如何保证算法的可靠性,提高数据产品的质量和准确性是亟待解决的问题。本文以基于手机信令的人口洞察类数据产品为切入口,研究数据质量校验方法和质检流程,用既定的规则代替相关人员主观上的判断,提高数据产品的交付质量。

1 数据质量校验整体框架

面向人口洞察的手机信令数据产品交付过程可主要分为数据采集、数据处理、建模分析共3个阶段,如图1所示。

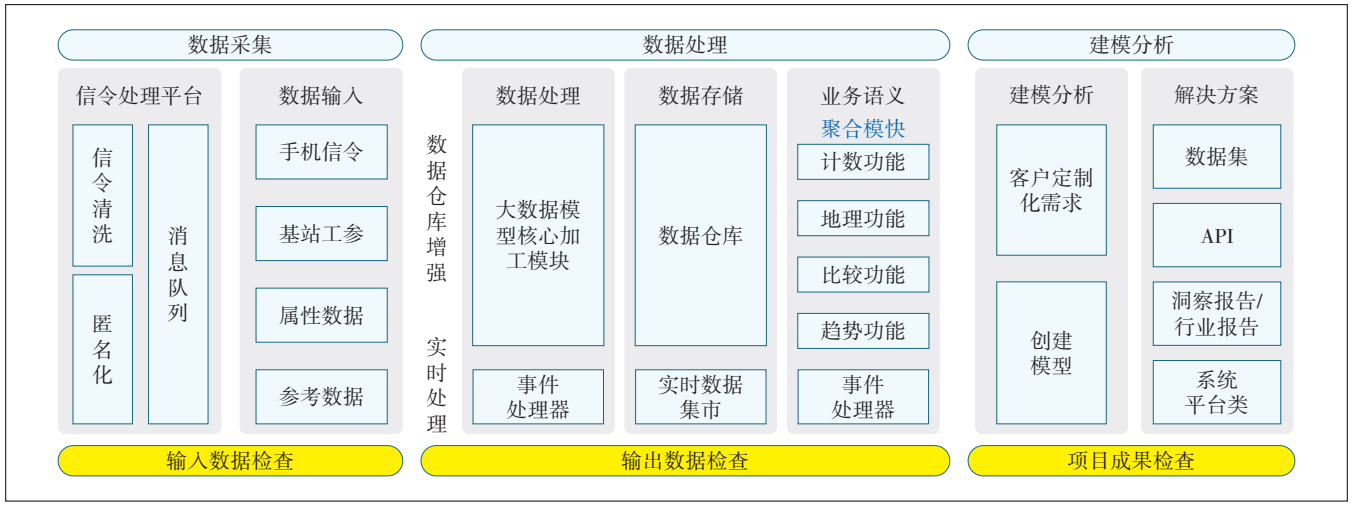

图1 数据质量校验整体框架

a) 数据采集:主要完成中国联通信令数据的接入、脱敏和清洗,以及不同来源数据(信令数据、基站工参等)之间的关联处理。

b) 数据处理:基于数据仓库,通过大数据模型的核心加工模块,将历史信令数据聚合计算为网格化标签数据,并基于GIS数据的路径拟合、基于预测算法的趋势分析等多种数据挖掘分析处理,生成人口洞察用的基础输出数据。

c) 建模分析与成果交付:根据客户定制化需求,对数据处理产生的数据结果进行建模分析,提供数据集、API、洞察报告/行业报告、系统平台类等多角度、多场景的交付成果。

数据质量校验必须贯穿数据产品的全过程,才能有效保障数据质量并提高产品交付一次成功率,实现提质增效的经营目标。

1.1 输入数据检查

输入数据检查主要包括手机信令、基站工参、属性数据、参考数据4部分。

a) 手机信令检查:以省为单位,对每日的2G/3G/4G信令数据进行检查,主要为分省数据的容量大小值。若某省信令数据量低于阈值,则标为异常并预警,人工排查原因,在问题修复或数据恢复后进行事件记录。

b) 基站工参检查:对每月灌入的2G/3G/4G基站工参数据进行检查,主要是分省统计2G/3G/4G各类基站工参总量,同时检查工参完整性,剔除基站经纬度为空或为0的工参数据。

c) 用户属性检查:对每月灌入的用户属性表进行统计检查,主要为:各省用户总量、正常状态用户量检查,其中正常状态用户指当月有手机信令的用户;年龄未知用户及占比、性别未知用户及占比检查。

d) 参考数据检查:参考数据主要为全国各区县的人口统计数据,为年度更新表,以各统计局/政府发布的统计年鉴数据为准。检查内容:按地(市)汇总区县

常住人口，与统计公报公布数字进行对比，检查是否吻合；男女总数是否等于常住人口数。

1.2 输出数据检查

通过大数据模型的核心加工模块处理，将手机信令数据加工为5大类核心数据表，分别为月点位表、月驻留表、月出行表、日驻留表和日出行表。对这5大类输出数据进行省、市、区县级等更精细空间粒度的数据质量检查，所采用的方法主要是统计学中的离散系数和离群值检验方法，检查项如表1所示。检查方法、检查逻辑及评判标准如表2所示。

表1　输出数据质量检查指标

校验表	检查指标名称	检查粒度	检查频率
月点位表	日总用户量	省/市/区	月
月驻留表	日驻留总量、人均驻留数量	省/市/区	月
月出行表	日出行总量、人均出行数量	省/市/区	月
日驻留表	日驻留总量、人均驻留数量	省/市/区	日
日出行表	日出行总量、人均出行数量	省/市/区	日

表2　输出数据质量检查逻辑

	月度	日度
检查公式	$G_n=(x\_((n))-\bar{x})/s$，其中 G_n 为格拉布斯检验系数，n 为统计样本数，$x\_((n))$ 为用户量，$\bar{x}$ 为用户量平均数，s 为标准差	
检查逻辑	以当月之前12个月计算平均值和标准差（如果12个月中有离群值，则抛掉离群值再计算平均），用当月统计数减去12个月平均值，再除以标准差，得到检验系数	以当天之前的30天计算平均值和标准差，再用当天的用户数计算检验系数
评判标准	检验系数绝对值>2.134，则标记为异常	检验系数绝对值>2.577，则标记为异常

1.3 项目成果检查

建模分析阶段的质量检查流程主要分为项目启动、项目成果提交、质检和内部/外部评审4部分，如图2所示。

a）项目启动：项目启动时需通知质检负责人，告知项目名称、城市描述、交付形式、项目描述、交付负责人、计划交付日期、计划提交质检时间。此时质检状态为“待提交”。

b）提交的项目成果主要为：

（a）合同+补充协议：指最终签订的项目合同，以及执行过程中因需求变更或者新增需求而增加的补充协议。该内容一般作为质检依据。

（b）成果包：包含执行代码、数据图层、数据集、图、报告。

（c）成果说明表：包含交付客户的成果列表、成果数据字典以及统计口径。

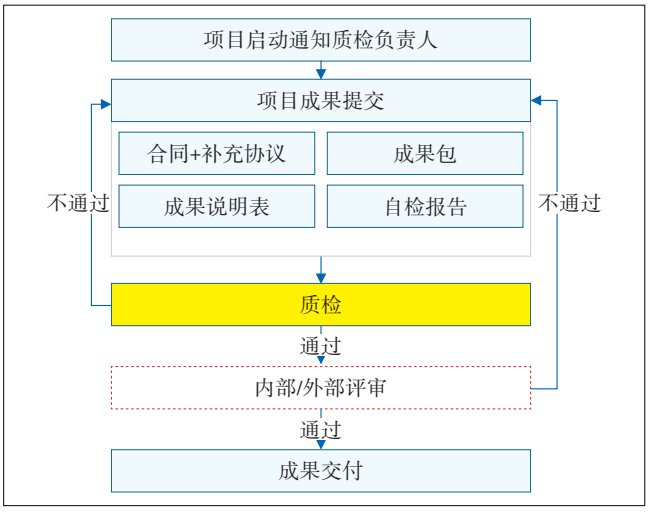

图2　建模分析与成果交付质量检查流程

（d）自检报告：交付负责人自检报告。此时质检状态为“待质检”。

c）质检：质检负责人根据提交内容，按质检指标进行质检；若质检无问题，则“质检通过”，并邮件通知交付负责人，通过质检可进行成果交付；若质检过程发现问题，则将质检结果以邮件形式反馈给交付负责人，此时质检状态为“修改中”；交付负责人进行修改，直至质检通过。

d）评审：对于大额项目或有重大战略意义的项目，由质检负责人组织内部评审。交付负责人作为主讲人，讲解项目需求、制作方案，展示项目成果和质检结果，内部专家对以上内容进行评审，评审过程记录到《评审记录表》进行留存和问题追踪。如果内部评审后对结果意见不一致或把握性低于50%的，由项目经理邀请相关的行业专家，召开外部评审会。

2 质检关键指标体系

面向人口洞察的手机信令数据产品的质检，分为人口分布类、出行类、属性类、热力分布4大类，具体细化成17项关键指标，如表3所示，该部分指标均为必查项，且问题等级均为最高级。

3 案例应用

以2019年9月商洛市人口分布与出行大数据服务为例，详细介绍质量检验流程。该项目的统计指标为：人口分布特征研究，含居住人口分布，性别、年龄特征分析；人口出行特征研究，含全用户出行空间分布、出行距离、出行时间。下文主要从输出数据检查

表3 质检指标

质检项	质检指标
人口分布类	极值合理性检查:居住人口top5%检查,不存在明显的数值断层
	落水检查:水域中不应该存在居住人口、工作人口
	人口分布合理性检查:居住人口top5%检查,查看空间分布,呈集聚状态,无零散独立分布
	居住人口类检查:特一线、一线城市不超过20万人/km²,二线城市不超过15万人/km²,其他城市不超过10万人/km²
	常住人口检查:与统计年鉴公布的常住人口相比,上下浮动超30%需确认是否符合城市发展趋势
	城市职住比检查:与统计年鉴公布的数据相比,上下浮动不超过10%
	日流动人口:不超过居住人口的25%
出行类	通勤出行:上班出行与下班出行人口量差异值不超过50%
	出行距离:超50%用户应集中在15 km以下
	平均出行距离,平均出行时耗特征,与交通年报校核,不存在太大差异
	分小时出行:工作日有明显的早晚高峰
属性类	居住人口男女性别比例接近1:1
	APP排名,应与第三方公布的排名基本相符
	终端品牌排行,应与网上公布的排行榜符合
	数据安全检查:不输出明细数据,交叉属性的汇总结果仅输出≥15的记录
热力分布	分小时人口热力曲线,早晚两头低,中间时间段人口数量高,不出现明显的缺失或跳动
	相邻日的首尾人口变化规律应有延续性

和成果检查2方面进行具体说明。

3.1 输出数据检查

本项目分析过程均采用月度增强模块处理后的月表进行统计分析,因此对2019年9月商洛市输出的月点位表(见图3)、月驻留表(见图4)、月出行表(见图5)进行检查,可以看出:

a) 日总用户量、日驻留总量、日出行总量,均在均值上下小范围浮动,无明显缺省异常情况,格拉布斯检验系数范围分别为(-0.88,1.01)、(-0.88,1.55)、(-0.15,0.29),最大值均小于2.134,正常。

b) 中秋节前后、国庆节前日,出现了比较大的用户量的增长现象,符合节日特征。

3.2 成果检查

3.2.1 人口分布特征结果检查

据《2018年商洛市国民经济和社会发展统计公报》显示,"2018年末,全市总户数85.70万户,户籍人口251.03万人。总人口中,男性133.17万人。据1%人口抽样调查结果显示,2018年末全市常住人口238.02万人,比上年减少0.11万人。"根据中国联通用户外推计算的2019年9月商洛市居住人口198.56万人,差值16.58%。查看商洛市2019年1月至11月中国联通总用户量的分布特征(见图6),可以看出商洛市从2019年1月21日春运开始到2月底,人口出现大幅增长,且从公报看出户籍人口大于常住人口,表明该城市主要为人口输出型城市,因此出现平常月居住人口略低于统计口径的常住人口,属于正常现象。

从各区县的居住人口分布密度来看,中心城区商州区的人口分布最密集,与其经济发展为全市第一相符。从人口属性来看(见图7),商州市男女比例为0.52:0.48,与统计公报发布的0.53:0.47基本吻合;年龄结构上,25~34岁人口为主,其次为35~54岁,符合正态分布。

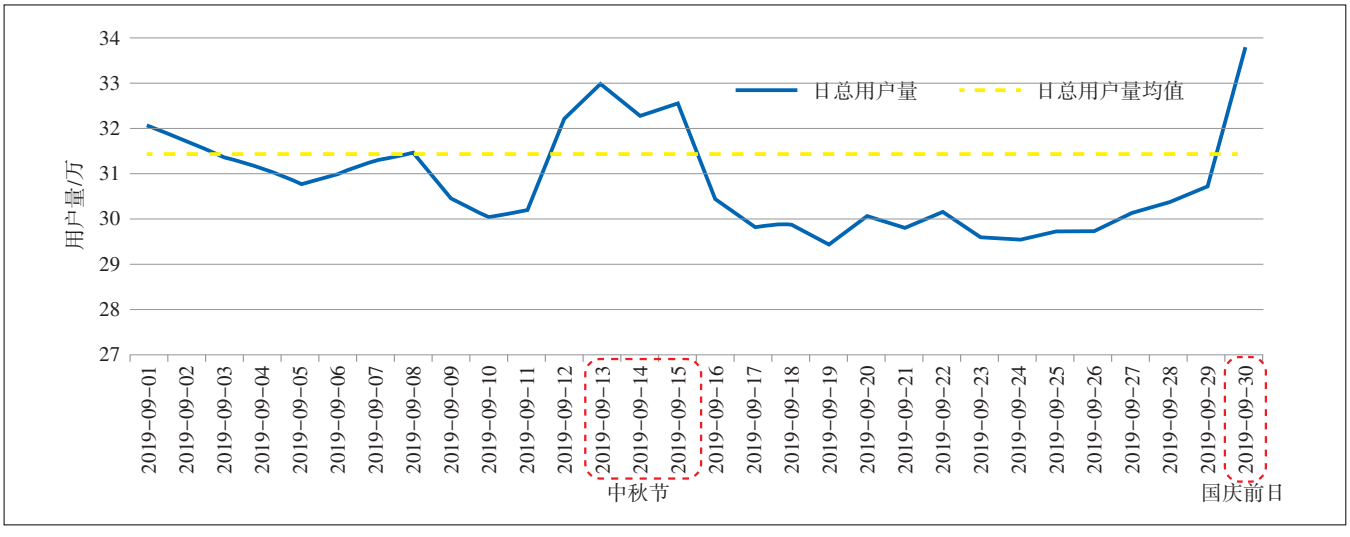

图3 点位表输出结果检查

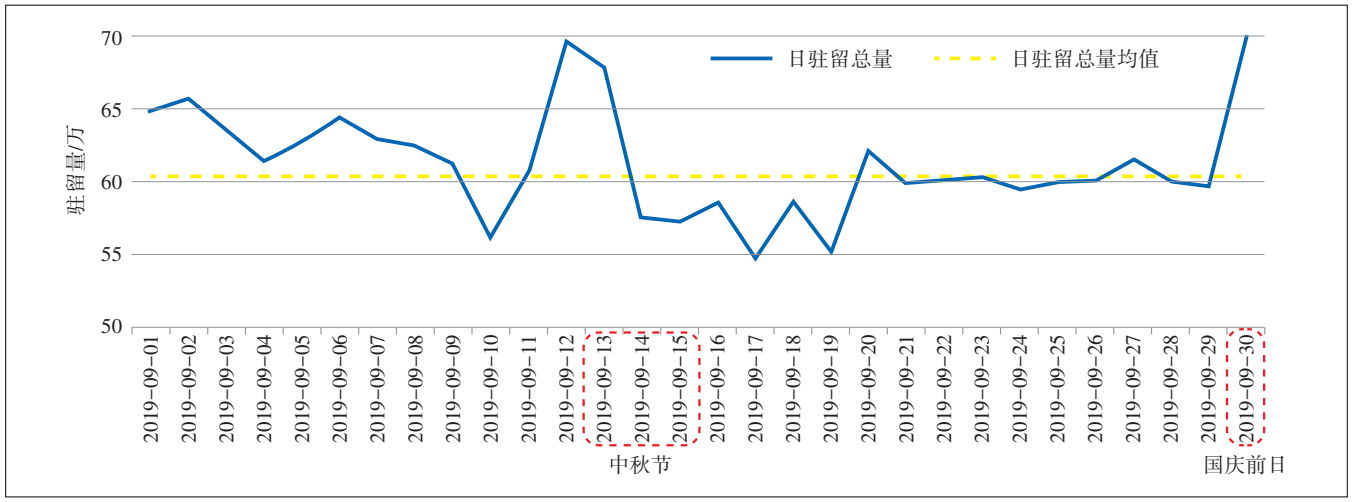

图4　驻留表输出结果检查

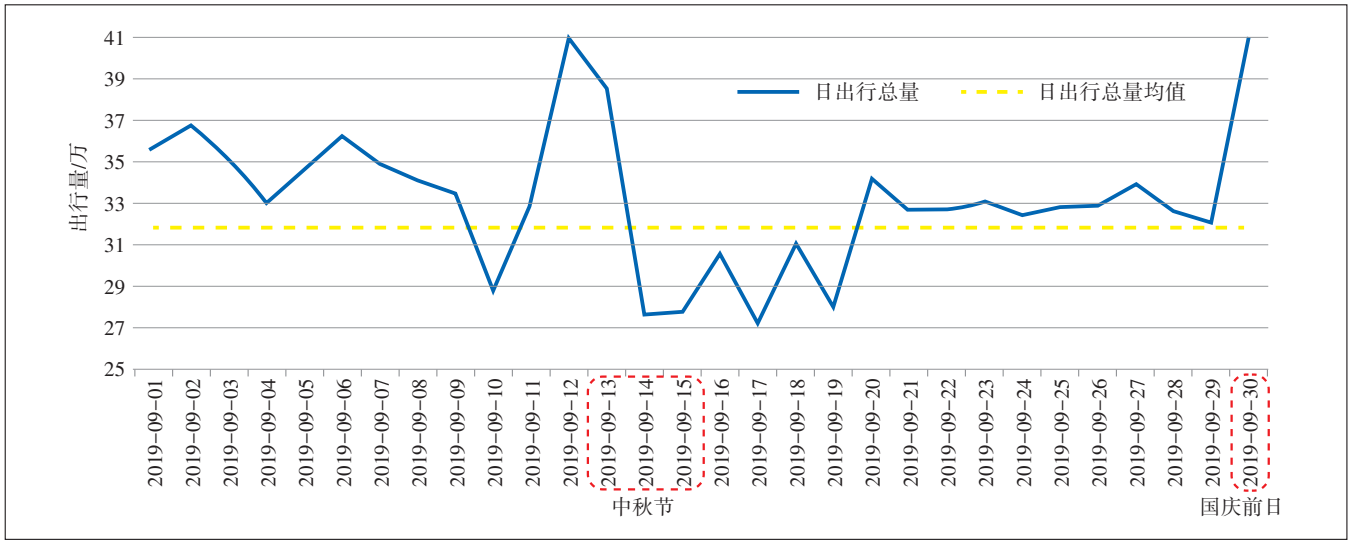

图5　出行表输出结果检查

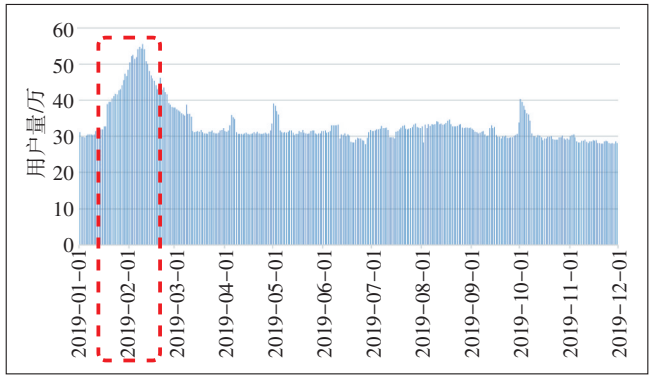

图6　商洛市2019年1月至11月中国联通总用户量分布特征

3.2.2 人口出行特征结果检查

从工作日、周末全方式出行数据统计结果来看：

a）各区县均以区县内出行为主，除丹凤县外，其他区县区内出行：跨区出行均大于7:3（见图7和图8），符合当地经济发展现状。

b）跨区出行，以中心城区商州区为中心，其与周边区县活动最频繁，符合中心城区的出行特点（见图8和图9）。

c）工作日出行有明显的早晚高峰，工作日主要为晚高峰，符合分小时出行特征（见图10）。

d）工作日和周末出行距离在15 km以下的用户分别为58.14%、58.24%，均高于50%，正常（见图11）。

4 结束语

手机信令大数据是开展人口洞察的主要数据源，基于手机信令的数据产品种类繁多，时空尺度多样，

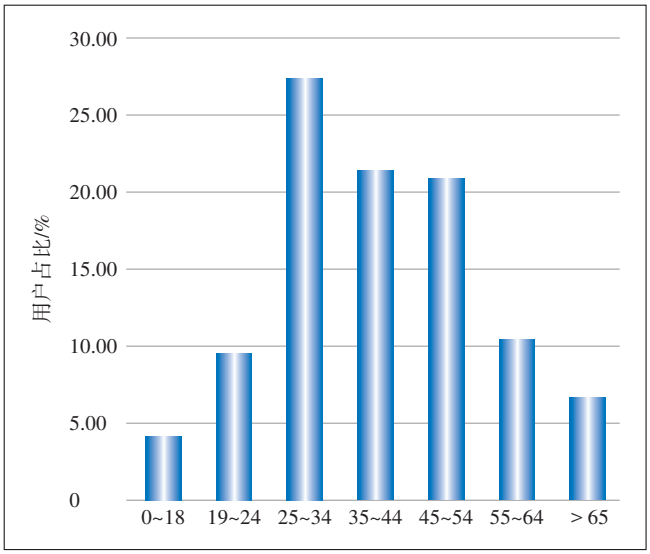

图7　商洛市居住人口年龄分布

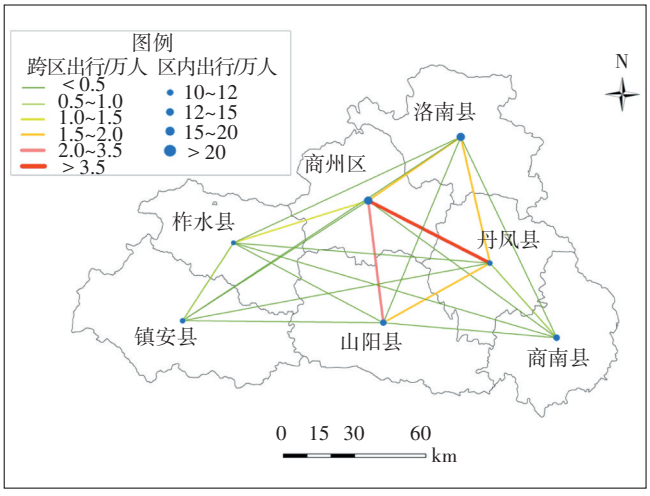

图8　工作日出行分布

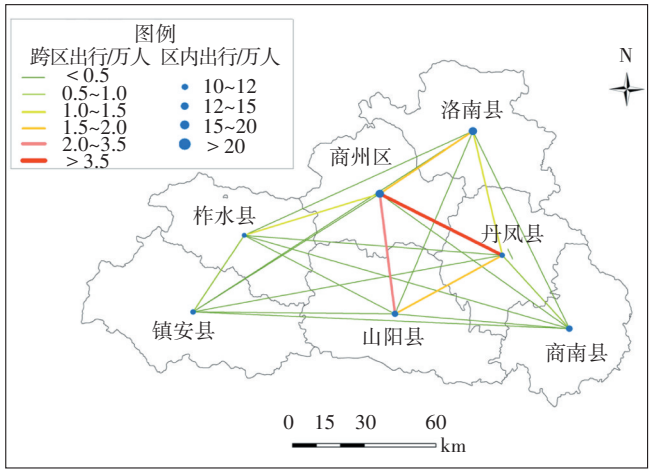

图9　周末出行分布

其数据质量和交付效果是政府及企业客户的关注重点。制定可信的质量校验规则,替代分析人员的主观判断,将有利于交付质量的提升。智慧足迹聚焦“人口+”,以位置为核心,长期进行人的职住、出行、行为、消费、健康等人口洞察类数据产品的研发,在产品交付和质量控制方面具备大量的经验,通过总结数据分析过程中的各种问题,将面向人口洞察的手机信令数据产品的质检划分为4大类共17项指标,形成了可行的质量校验方案和质检流程,提高了数据产品的数据质量和交付品质。

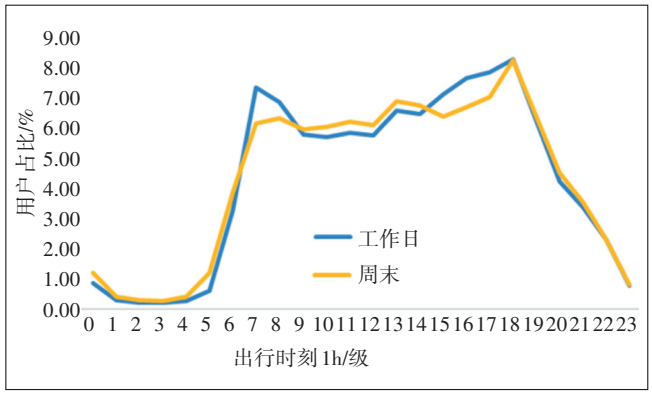

图10　分小时出行分布

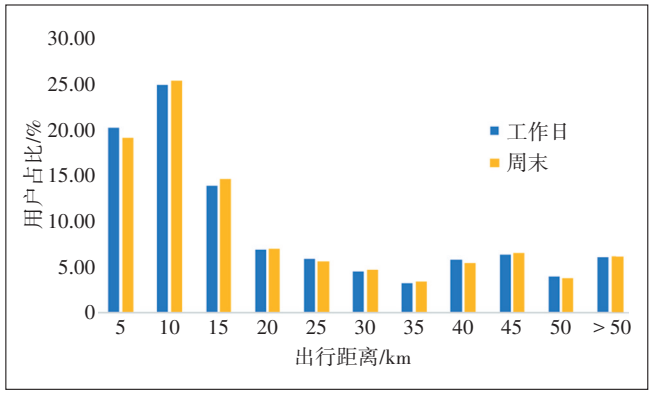

图11　出行距离分布

参考文献:

[1] 赵鹏军,胡昊宇,海晓东,等. 基于手机信令数据的城市群地区都市圈空间范围多维识别——以京津冀为例[J]. 城市发展研究,2019,26(09):69-79.

[2] 钮心毅,王垚,丁亮. 利用手机信令数据测度城镇体系的等级结构[J]. 规划师,2017,33(1):50-56.

作者简介:
梁洁,高级数据分析师,测绘工程师,主要从事手机信令数据在政府城市规划领域的业务应用研究及手机信令产品的数据质量校验工作;张岩,智慧足迹数据科技有限公司COO,教授级高级工程师,长期从事手机创新业务及大数据业务的规划设计、研发管理和运营支撑工作;闫嘉,高级产品总监,项目管理专业认证(PMP),主要从事电信运营商时空大数据应用场景的规划设计和项目落地可行性研究工作。