电信运营商大数据在交通运输行业旅客联运场景下的应用研究

Research on Scene-oriented Application of Telecom Operator Big Data in Passenger Intermodal Transport

许致远，张　慧，张　鹤（联通数字科技有限公司，北京 100032）
Xu Zhiyuan，Zhang Hui，Zhang He（Unicom Digital Technology Co.，Ltd.，Beijing 100032，China）

摘　要：
在分析交通运输行业中旅客联运业务发展现状和对大数据需求的基础上，结合电信运营商大数据的5V特性和应用现状，梳理了电信运营商大数据与旅客联运业务场景的结合应用思路，并总结了其中的5项关键技术。最后，以中国联通全量数据的应用实践为例，对以上思路进行佐证。

关键词：
电信运营商；大数据；旅客联运；场景应用
doi：10.12045/j.issn.1007-3043.2022.04.013
文章编号：1007-3043（2022）04-0064-09
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：
Based on the analysis of the development status of passenger intermodal transport business in the transportation industry and the demand for big data，combined with the 5V characteristics and application status of telecom operators' big data，it sorts out the application ideas of telecom operators' big data and passenger intermodal transport business scenarios，and summarizes five key technologies. Finally，the application practice of China Unicom's full-volume data is taken as an example to prove the above ideas.

Keywords：
Telecom operator；Big data；Passenger intermodal transport；Scene-oriented application

引用格式：许致远，张慧，张鹤. 电信运营商大数据在交通运输行业旅客联运场景下的应用研究[J]. 邮电设计技术，2022(4)：64-72.

0 前言

旅客联程运输（以下简称“旅客联运”）是建设现代化交通强国的重要举措，也是单一客运方式发展到一定阶段后的必然结果，随着海量旅客跨域活动的逐渐增多，旅客对高效、安全的联程出行提出了越来越紧迫的需求，如何引入数据化手段把握整体发展状况，统一规划旅客联运体系，成为亟需解决的问题。而电信运营商所承载的用户数据，天然具备时空连续、真实可靠、实时鲜活等优势，为解决旅客联运业务研究和发展中的部分问题提供了有效路径。

收稿日期：2022-02-16

1 旅客联运业务现状及需求

1.1 旅客联运术语解释

根据JT/T 1109-2017《中华人民共和国交通运输行业标准》中的定义，旅客联运指的是通过2种或2种以上对外运输方式完成的旅客连续运输。其发展目标是由单一旅客联运承运人或代理人为旅客及其行李全程负责，旅客全程使用一本票证。

通俗来说，旅客联运是一种客运组织模式，即通过对同一旅客不同运输方式的多段行程进行统筹规划，提高旅客出行效率和满意度。同时，该组织模式

可充分发挥现行多种客运方式的比较优势，对于加快推进旅客运输服务的供给侧结构性改革，推动现代综合交通运输体系建设具有重要意义[1]。

1.2 我国旅客联运业务发展现状

2017年2月，国务院印发《“十三五”现代综合交通运输体系发展规划》，其中第4项明确提出要“推进旅客联程运输发展”，要求各省开展专项行动，推进跨运输方式的客运联程系统建设，实现不同客运方式间的有效衔接，以此提升客运服务安全便捷水平。2020年11月，新华社发布《中共中央关于制定国民经济和社会发展第十4个五年规划和二〇三五年远景目标的建议》，其中也指出“发展旅客联程运输”仍是“十四五”时期交通强国建设工程的重要举措之一。

随着近年来我国交通运输行业的快速发展，以民航、高铁、高速公路等为代表的主干旅客运输方式的整体运能和服务水平也在不断提高，现代化、智能化、一体化的旅客综合运输体系正在加速形成。已有不少人员聚集度较高、出行频率较高且经济发展水平较高的地区，如京津冀、长三角、珠三角等比较成熟的城市群，积极发展完善了空铁联运、空巴联运、公铁联运等旅客联运业务，并取得了一定的成效，满足了部分跨域旅客对于联程出行的日常需求[1]。

但从总体上来看，我国旅客联运业务发展仍处于起步阶段，存在发展水平较低、发展基础薄弱、市场主体不成熟、重点环节服务不到位、行业发展缺乏共识、基础设施不适应、信息共享难度较大、法规标准不协调以及体制机制缺乏协同等诸多困难和问题，亟待进一步的科学研究、市场优化和新技术支撑。

1.3 旅客联运业务对大数据的需求

旅客联运是现代综合交通运输体系的重要组成部分，其服务核心是旅客，业务需求比较复杂，仅从对于大数据的需求方面来看的话，主要集中在数据源、平台、融合建模以及创新应用4个层面上。

a）数据源层面：旅客联运是典型的跨部门、跨市场主体的组织行为，其涉及的数据源较为分散，既有监管部门汇总数据，也有运营单位实时数据，还有第三方企业的业务数据，多源数据的采集、采购、存储、计算是其后续应用的基础。

b）平台层面：区域性乃至全国性多式联运数据平台一直是有缺失的，对于铁路、公路、航空、水运等联运主体间基础数据共享困难，导致衔接时“最后一公里”不畅通等问题，需要通过平台的共享开放、标准化治理等手段来辅助实现。

c）融合建模层面：旅客的联程出行需求同联运基础资源的匹配程度是旅客联运业务发展优劣的衡量指标之一，融合多源数据构建出行预测模式，对于统筹规划相关交通资源会有很大助力，比如旅客出行方式、各方式客流分担率、旅客驻留时间等常用模型。

d）创新应用层面：“出行即服务”是旅客联运的重要发展方向，但在既有技术条件下，单纯依靠公路、铁路、航空等公共承运方是很难完成旅客端到端的联程服务需求的。因此，以数据为媒介提升对旅客全出行链的服务能力，打造提升旅客体验的创新应用显得尤为重要。

交通运输部已筹建全国综合交通运输标准化技术委员会（编号TC571），对于旅客联程运输场景下涉及到的电子客票信息系统互联互通技术规范进行了编制，一定程度上解决了数据源对接和平台层面信息共享的部分难题。

2 电信运营商大数据特性及应用现状

2.1 运营商大数据的5V特性

通信运营商大数据一直被认为是大数据界的金矿，按照业界当前的通用评价方法，从体量（Volume）、种类（Variety）、价值（Value）、速度（Velocity）和质量（Veracity）这5个维度分析通信运营商大数据的特性。

a）体量大：在体量方面，电信运营商数据具有“大而广”的特点，即数据量巨大且覆盖面广。截至2020年底，全国移动电话用户总数达15.94亿户，普及率为113.9部/百人，固定互联网宽带接入用户总数达4.84亿户。数以亿计且不断增长的用户体量构成了电信运营商数据应用的基石，其采集基本不受地理环境、空间分布和经济发展等外部因素影响。

b）种类多：由于用户的通信行为、网络行为等依赖于电信运营商，故其数据类型包括身份、位置、社交、消费、终端、上网等多个维度，涉及结构化数据、非结构化数据以及半结构化数据等，种类极为丰富，而且各数据维度间存在以用户ID为主键的强关联特征。从数据来源区分，电信运营商大数据来源涵盖O域（Operation support system，运营支撑系统）、B域（Business support system，业务支撑系统）和M域（Management support system，管理支撑系统）3大支撑域的内部数据，以及众多生态合作伙伴的业务数据。

c）价值密度高：一般来说，大数据虽然体量巨大，

但其价值密度却远远低于传统的关系型数据库中的标准化数据,需要提炼和深度挖掘才可体现其价值。但由于国内已实现100%的通信用户实名制入网,使得电信运营商大数据的价值密度相对较高。电信运营商可以使用用户号码为唯一ID来整合各类行为数据,无需进行漫无目的的大规模挖掘实验,其刻画用户、洞察行为的完整性和便捷性是其他任何行业数据都难以企及的。

d)存取速度及增速快:电信运营商具备天然的数据属性,在多年处理用户信息以及消费数据的过程中,通过计费、管理、服务、运营等业务平台的多维度建设,已经建成可以快速处理上百PB海量用户数据的计算分析平台,有丰富的内部实践经验,能够快速地对外提供数据加工、建模计算和挖掘服务。而且,每时每刻都在不断产生的海量通信数据,使得其体量增长速度极快,充分满足大数据价值挖掘的鲜活性要求。

e)优质且真实可靠:不同于大部分互联网数据,电信运营商数据皆是实名用户基于实际行为的信息记录,且为满足核心业务需要,电信运营商内部对于各业务系统的数据标准化、质量稽核、运维管理等皆有较长时间的规范化实践,使得该部分数据可信度和可靠性有很大程度的提升,初步构建了干净高品质的数据资产库,这也为数据对外服务奠定了坚实基础。

2.2 电信运营商常用数据类别

数据是电信运营商的立身之本,对外用于行业服务的数据主要涵盖身份、位置、上网、社交、支出、通信、终端和时序8大类[2]。

a)身份数据:具备实名制的客户资料数据,覆盖完整且真实准确,还可基于实际行为进行验证,可用于判定用户的信用程度等。

b)位置数据:手机CS域信令具有基站位置更新、开关机动作以及位置区切换等信息,相比PS域和话单更容易作为基础数据进行分析,可用于洞察用户出行特征、迁移动态、停留时长等。

c)上网数据:基于用户访问哪些网址、下载哪些应用、访问哪些内容等的日志数据,可了解用户的出行偏好、线上喜好等。

d)社交数据:基于通信交往圈的大小、主被叫和时间序列,可用于分析用户的社交特征。

e)支出数据:基于用户通信消费数据,比如流量费用、短信费用、语音费用以及新业务费等,了解用户消费特征及基本信用情况。

f)通信数据:通过用户的通信使用情况,比如本地通话、长途通话等,了解用户通话行为特征。

g)终端数据:识别记录用户的移动手机终端型号,了解用户手机使用特征、换机周期以及消费水平、消费偏好等。

h)时序数据:将用户上网、位置、通话等行为按照时间排列,了解更多规律性特征以便进行更深度的行业洞察。

2.3 电信运营商大数据在交通运输行业的应用情况

运营商数据在交通运输行业的应用由来已久,早在2009年就有利用运营商提供的手机信令数据开展交通规划相关研究的探索,但当时受限于数据量、网络和技术水平等原因,还仅限于理论研究,2015年开始出现高速增长,至今热度不减。从产学研用的角度来看,运营商大数据在交通运输行业的应用和实践已经逐步进入成熟期,可用且能发挥较大效用的场景也已经较为明确,其应用场景可基本总结为以下几个方面(见表1)。

随着电信运营商与行业客户的进一步磨合,对于交通运输行业市场会越来越清晰,逐步聚焦可规模化复制的应用场景,最大限度地挖掘和发挥自身大数据的内在价值。

表1 运营商大数据在交通运输行业应用场景分类

序号	类别	具体应用场景
1	辅助交通规划	区域性综合运输规划
2		综合客运枢纽规划
3		多模式旅客联程运输规划
4		城市线网、路网规划
5		站点枢纽规划
6		物流线路规划
7	基础设施开发建设服务	交通基础设施建设立项依据
8		轨道站点建设(客流、服务半径等)
9		公交站点建设
10		机场建设评估(流量、吞吐量等)
11		停车场、充电桩等建设选址
12	交通监测及运营管理	高速公路流量监控
13		旅游交通监测及预警
14		城市轨道交通人流监测及运营管理
15		综合交通枢纽客流分析
16		城市道路交通控制(信控等)
17		突发事件、重大活动等应急管理
18	评估评价及改进	基础设施建设的可行性评估及后评价
19		交通设施安全评估

3 旅客联运场景下电信运营商大数据应用思路

3.1 总体思路

旅客联运场景本身比较复杂，一段完整的联程出行一般涉及多个城市、多个交通枢纽以及多种交通运输方式，时间和空间跨度较大，难以进行全方位的分析。因此引入电信运营商大数据之后，便可利用其时空连续、覆盖范围广的优势，从旅客本身的行为特征入手，反向研究联程运输这一组织方式的现状、问题和发展方向等，用以辅助宏观洞察、缺陷发现、规划设计以及线路优化等，最终实现“现象—数据—决策—行动—评估—更好的现象”这样的良性应用闭环，让数据助推现实发展，总体思路如图1所示。

3.2 关键技术及应用解决路径

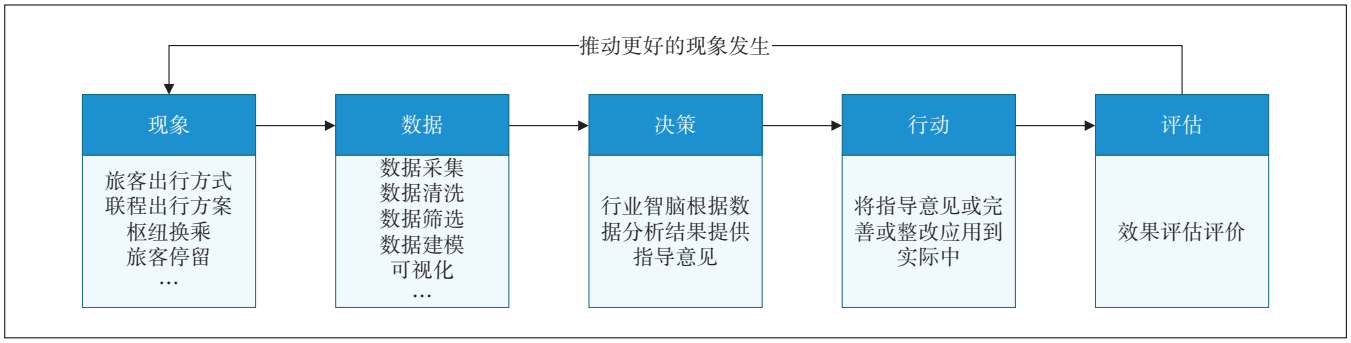

图1　总体思路示意图

根据上文提到的总体应用思路，其核心点之一在于如何能用可获取的数据更精准地反映业务现象，这也是体现运营商数据资源价值以及数据挖掘水平的地方。通过项目实践和行业理解，总结了这一过程中涉及到的5个关键技术点，并提出了可落地的解决路径，具体阐述如下。

3.2.1 旅客出行链提取

运营商大数据要想在旅客联运场景下应用，首先要做的便是将每个旅客完整的出行链（Trip Chain，TC）拆分出来。

按照交通规划业务中的常用定义，旅客从出发地到目的地的移动过程被称为一次出行。一次出行有2个端点，开始的端点叫起点（即出发地Origin，简称O点），结束的端点叫讫点（即目的地Destination，简称D点），中间过程可能有城际出行、市内出行、枢纽换乘等多种交通行为，所以业界把出行分析也叫作起讫点分析（即OD分析）。如果把某个旅客个体在一段时间内所有的起讫点按照时间顺序连接起来，可以形成由出行起讫点构成的序列，这种“时间+位置”组成的有序序列被称为出行链[3]。出行链中的每个点代表出行的一个端点，每条连线代表一次出行，图2给出了一次完整的出行链示意。因此要提取旅客的出行链，必须约定好时间和空间范围。

按照指定的空间尺度（省、市、县或圈定的某个区域均可）和时间周期，可以获取手机终端用户在这段

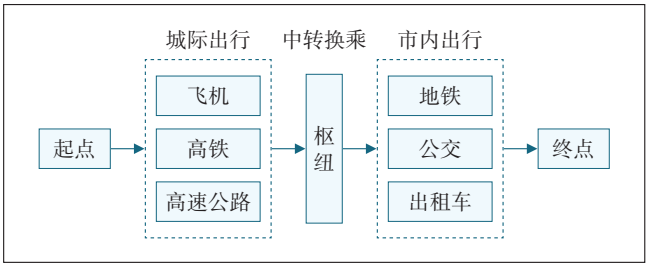

图2　一次完整的出行链示意图

时间内的起始位置，然后根据该手机终端在运营商基站之间的信令切换数据推导出其在空间位置上的移动轨迹，持续跟踪至其在某一个驻留点持续时间大于约定的阈值（如市域一般约定为6 h，县域约定为2 h等，可根据实际空间尺度情况进行调整），最终确定其交通轨迹为一次完整的TC，并将该时空序列数据提取出来。在此过程中，可以结合各POI点的土地属性，进一步分析用户轨迹特征，根据事先设定好的空间、时间和频次阈值，将从信令数据中提取出来的各轨迹点按照驻留点、移动点等进行划分，为简便起见，驻留点之间的变化即可判断为出行。

3.2.2 旅客群体划分

对于旅客属性的识别可用于对特定的群体进行观察分析，比如学生群体、外来务工群体、旅游偏好群体等，结合用户入网实名制资料及出行特征，可以构建初步的群体识别模型，对具备共性行为特征的用户群进行筛选，条件具备的场景下可进行二次核验，提

高模型的准确性。

运营商信令数据中具备丰富的数据维度来帮助解决此类问题,比如针对学生群体(主要指成年学生群体)的判断,可以聚焦在年龄(身份证第7~14位)、常驻POI是否为高校、常驻时间和天数等明显特征;针对外来务工群体的判断,可以聚焦在籍贯(身份证前6位)、职驻地、籍贯地与职驻地之间的往返规律等特征;针对旅游偏好群体的判断,可以聚焦在职驻地、外地驻留时间、外地景区POI驻留时间、旅游类APP或网站浏览次数等特征。有了这些基本的判断依据之后,便可以根据数据的统计特征进行加工逻辑设计,选取置信度较高(一般为95%)的置信区间节点作为临界点,并进行多次拟合验证。按照拟合通过后的加工逻辑对生产数据进行加工,生成相应的属性标签。当需要对这些特殊群体进行专题分析时,便可以根据属性标签进行快速提取。

重点人群在特定时间段内的出行需求受到越来越多的关注,如每年的春运期间,铁路部门就针对学生群体和外来务工群体制定了相关的服务政策,一定程度上对这些群体的集中出行进行了疏解。

3.2.3 出行方式识别

旅客出行方式的识别一直是重点也是难点,理论上来说,基于手机信令数据,结合飞机、高铁、汽车等常见交通方式的筛选原理和判定准则,在算法设计时预设好判定条件,并根据核验情况调整阈值标准,便可以进行比较精准的划分。但由于运营商基站数据本身存在误差较大、不够精细、脏数据多等问题,需要较为繁杂的数据清洗和业务验证过程。

根据旅客联程出行时表现出来的驻留位置变动、行驶路径特征以及行驶速度等,可以将联程联运中最常用的公路、铁路、航空3种主干方式先区分出来,主要技术路线可以分为4步[4]。

a)交通场站POI点归集。由于旅客出行的3种主干方式必然会跟相应的交通场站发生关联,公路对应汽车站或高速收费站、铁路对应火车站(含高铁站)、航空对应机场,且换乘驻留跟过站经停也会呈现出完全不同的基站信息,这些特征对基于交通枢纽的联运情况分析可以起到很好的辅助作用。因此,可以建立专门的场站POI表单,将所需空间尺度内的汽车站、高速收费站、火车站和机场的经纬度信息进行统一归集和动态更新。

b)旅客位置与场站中心位置匹配。通过经纬度信息,将旅客出行链中的端点信息与目标场站中心点进行拟合,在有效范围内的保留,有较大出入的非有效出行数据暂时剔除。在此过程中需要根据不同场站的实际规模进行换算,得到其最佳阈值半径,比如汽车站有效半径阈值在150 m~500 m、火车站有效半径阈值在500 m~2 000 m、机场有效半径阈值在1 000 m~3 000 m,特大或特小的场站根据实际情况调整。

c)根据匹配的场站类型初步判断出行方式。航空方式判定最为简单,一次出行的2个端点场站皆为机场,且中间过程无任何行驶路径,即可判定;铁路出行的2个端点皆为火车站;其他起讫点中有汽车站或收费站的都可暂判定为公路出行,更细致一点可以区分公共汽车和自驾车,但区分可信度不像航空和铁路方式那么准确。

d)根据行驶速度校验出行方式判断结果。旅客选择这3种主干交通工具的行驶速度差距较大,一般来说,民航最快、铁路次之、公路最慢,因此最后一步可以根据起讫点间距离跟花费时间,大致推算出行驶速度,同经验阈值进行对比,以此作为二次核验依据。

通过以上方法,可较为简便且准确地从海量的信令数据中将联程旅客的出行方式先一一甄别出来,运算量和成本较低,基本可以满足大规模空间尺度下的数据分析要求。

3.2.4 热门路线计算

在点到点之间总有一条线路因受欢迎或需求量大等原因聚集大部分的旅客,分析这个问题可以为联运设施规划和调配提供参考。基于运营商大数据进行热门路线的发现和识别需要对轨迹数据进行较深度的挖掘,目前主要采用DBScan这种较为成熟高效的聚类算法来实现[5]。

一般识别、计算过程可简单分为4个步骤。

第1步:从提取的出行链中截取所有满足条件的出行轨迹数据。

第2步:使用最短描述轨迹方法合并连续时间相同经纬度的数据,得到用户的驻留点$P\sim(p_1,p_2,\cdots,p_n)$。

第3步:按照时间顺序依次连接用户的驻留位置点,得到用户的出行轨迹段数据Tr_j,如$\mathrm{Tr}_1=\left(p_1p_2\right)$、$\mathrm{Tr}_2=\left(p_2p_3\right)$。

第4步:采取分布式DBScan聚类方法对大量轨迹段进行聚类计算,具体计算公式如下:

$$\mathrm{dist}(\mathrm{Tr}_1,\mathrm{Tr}_2)=3\times d_{\mathrm{per}}(\mathrm{Tr}_1,\mathrm{Tr}_2)+d_{\mathrm{par}}(\mathrm{Tr}_1,\mathrm{Tr}_2)+3\times$$

$$d_{ang}(\mathrm{Tr}_1,\mathrm{Tr}_2) \tag{1}$$

式中：

d_{per}——轨迹段Tr_1与Tr_2的垂直距离

d_{ang}——轨迹段Tr_1与Tr_2的角度距离

d_{par}——轨迹段Tr_1与Tr_2的平行距离

最终得到聚类后的轨迹段合集并进行最短续接轨迹端的递归计算，热门轨迹计算过程示意图如图3所示。

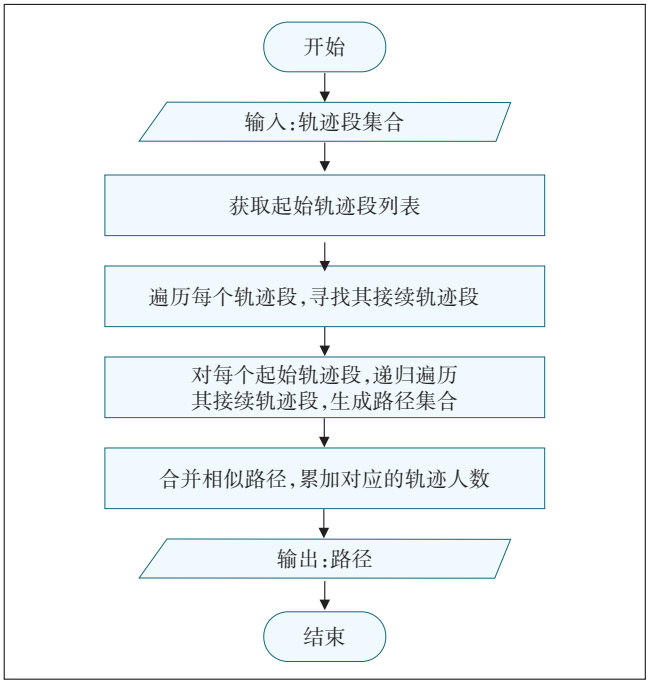

图3　热门轨迹计算过程示意图

根据各最短路径的相应人数即可进行TOP排序，涉及场站为起讫点的热门路径计算也类似。

3.2.5 旅客流量预测

电信运营商大数据具有很强的时空属性，所以利用该类数据进行客流量预测时，更趋向于进行时间序列预测。但时间序列很容易受到各种偶然因素的影响，呈现出不规则的波动状态，其分析方法只适用于近期以及短期的预测，而且预测效果并不尽如人意，比如平滑法、自回归模型、移动平均模型、自回归移动平均模型、差分自回归移动平均模型等，需要在生产环境中不断地调参优化。

目前常用的是差分自回归移动平均模型，也叫ARIMA模型，它是自回归移动平均模型的升级版，表达式如式(2)，记为ARMA(p,q)，主要针对非平稳时间序列进行预测[6]。

$$Z_t = a + \sum_{i=1}^{p}\varphi_i X_{t-i} + \sum_{i=1}^{q}\beta_i \varepsilon_{t-i} \tag{2}$$

式中：

Z_t——t时刻的预测值

a——均值

β_i——第i个权重

p——自回归阶数

q——移动平均阶数

ARIMA(p,d,q)为ARMA(p,q)结合有限次差分运算而来，以便将非平稳序列转化为平稳序列，其中d是差分阶数，通过修正这些参数构成多元线性回归函数，差分法的计算公式如下所示。

一阶差分计算：$\nabla x_t = x_t - x_{t-1}$ (3)

二阶差分计算：$\nabla^2 x_t = \nabla x_t - \nabla x_{t-1}$ (4)

所以可以基于ARIMA(p,d,q)模型，经过时间序列平稳性检验、时间序列随机性检验和模型识别与定阶3个基本步骤，将相关的3个参数计算出来，并进行模型拟合验证。该模型在准实时人口数据预测方面具有较高的精度，也因此在交通运输行业应用比较广泛，但仍具有一定的局限性，需要同具体场景下的其他数据进行联合检验使用，以提高预测的准确性。

4 中国联通全量数据应用实践

作为通信运营商，中国联通早在2010年就提出了数据大集中战略。2012年开始组建全国数据中心，实现了31省全网数据的统一集中汇聚与管理。其“一点集中，服务全国”的优势为各个空间尺度下交通运输行业的相关分析和应用提供了便利条件。

基于以上应用思路，本文进行了以下几个层面的应用探索和实践，以此来验证运营商数据赋能旅客联运场景的可行性。

4.1 三级空间尺度下的旅客出行特征分析

按照空间尺度划分标准，将旅客出行按全国、省、市进行3个级别的流动分析和指标监测，帮助各级交通运输管理部门宏观掌握旅客出行的数量和发展趋势，而且可以基于信令数据统计T-1小时时段内的实时出行人口数量，相比传统统计方式更加即时、鲜活。

以全国范围为例，管理部门一般主要聚焦在旅客出行数量和热门排行2个层面上，以区县流动OD为基础，对于整体的旅客出行链进行特征分析，为后续进行联程联运规划或问题发现打好基础。

a）旅客出行数量层面：主要关注在当日旅客出行

总量、实时出行人数、按小时人数趋势以及省内、省际出行分布情况等方面，按不同的时间范围和空间范围统计分析旅客出行数量，将这几个数据指标所代表的现实意义进行了详细分析，结合信令数据内容构建了区县级的小时轨迹表、日轨迹表等，对应各指标维度进行口径判定和持续矫正，旅客出行数量统计表说明见表2。

表2　旅客出行数量统计表说明

序号	数据指标		口径定义
1	当日出行总量		有跨区县轨迹且满足停留时长要求的所有用户总数，并进行全网扩样反推
2	实时出行人数		*T*-1小时时间段内各区县迁入旅客数量之和
3	按小时人数趋势		整点时间段内实时出行人数按时间排序
4	省内出行数量		31个省份区域范围内有跨区县轨迹且满足停留时长要求的用户总数，且起讫点区县均在省内
5	省际出行	出行旅客数量	31个省份区域范围内有跨区县轨迹且满足停留时长要求的用户总数，且起讫点区县有其一在不在省内，起点不在省内为迁入旅客，讫点不在省内为迁出旅客
6		常住人口迁入	迁入旅客中职住地在省内的用户总数
7		常住人口迁出	迁出旅客中职住地在省内的用户总数

b）热门排行层面：从全国范围上看旅客出行轨迹会呈现一个很大规模的宏观态势，很难快速地进行更深的研究，那么对于一些旅客聚集度较高、重合度较高的OD进行二次分析便会有事半功倍的效果，因此对全国热门的出行目的地、热门交通枢纽（主要指火车站、汽车站和机场）、热门省际出行OD、热门城市出行OD以及热门中转城市等数据指标的拆分可以提高对旅客出行特征的认知水平，同样本文也进行了数据口径的研究，热门排行统计表说明见表3。

任意选取某一日数据，以2021年4月1日为例，将上述数据指标按照梳理好的口径定义进行了计算，原始数据来自中国联通数据中心，为全量中国联通用户数据，具体计算结果如表4所示。

将计算结果同有关部门统计数据及其他渠道获取的数据进行了对标分析，除少数指标有量级出入外，大部分数据拟合度较好。当然，某一天的数据不能完全说明模型的可靠性。本文建立了几个核心指标的稽核机制，以自动化对比方式按天进行核对，以此反推实时数据口径的合理性，并根据异常数据情况进行模型口径的调整，目前已经可以较为准确地为需

表3　热门排行统计表说明

序号	数据指标	口径定义
1	热门出行目的地	实时的热门目的地以*T*-1小时时间段迁入旅客实时数量按序排名；按天/月的热门目的地以对应时间段迁入旅客总量进行排名
2	热门交通枢纽	一定时间段内，剔除职住人群后各火车站、汽车站、机场POI范围内旅客总量排名
3	热门省际出行OD	一定时间段内，按出行旅客数量进行归属省份排名，可按职住模型区分是否为通勤人员出行（OD起讫点均不是职住地为非通勤出行）
4	热门城市出行OD	一定时间段内，按出行旅客数量进行归属城市排名，可按职住模型区分是否为通勤人员出行（OD起讫点均不是职住地为非通勤出行）
5	热门中转城市	当日内在3个或以上区县产生出行轨迹的用户，非OD起讫点区县所在城市作为中转城市，按该类轨迹旅客数量进行排名

表4　2021-04-01旅客出行数量及热门排行榜

数据指标	计算结果示例		
当日出行总量	52 574 051		
省内出行数量	45 269 269		
省际出行旅客数量	14 609 566		
省际常住人口迁入	2 789 496		
省际常住人口迁出	3 589 361		
热门出行目的地TOP5	郑州		915 828
	合肥		754 742
	成都		662 930
	泉州		620 222
	苏州		547 651
热门交通枢纽TOP5	北京大兴国际机场		388 875
	成都双流国际机场		376 027
	重庆北江国际机场		347 745
	广州白云国际机场		336 344
	北京首都国际机场		309 965
热门省际出行OD TOP5	安徽省	江苏省	279 778
	河南省	安徽省	213 163
	江苏省	安徽省	189 731
	江苏省	上海市	180 348
	湖南省	广东省	179 688
热门城市出行OD TOP5	咸阳市	西安市	121 880
	西安市	咸阳市	107 996
	廊坊市	北京市	98 140
	佛山市	广州市	96 203
	中山市	珠海市	94 832
热门中转城市TOP5	长沙市		19 318
	成都市		16 039
	甘孜藏族自治州		14 866
	广州市		13 816
	郑州市		12 018

求方提供服务。

4.2 客运方式及旅客联程出行洞察分析

优化旅客联运的基础是对其所依赖的运输方式现状的掌握，因此对全国及各省铁路、道路客运、水运、航空展开了分方式客运量分析与预测，结合出行旅客属性以及出行时间、距离等规律分布情况，展开多方式联程出行分析，实现全国联程情况概览。

该部分应用主要聚焦在公路、铁路、水路等主干客运方式概况和旅客联程出行现状分析2个方面，依赖于出行方式识别、旅客流量预测等算法模型的性能和准确度，是从旅客出行链出发进行研判的重要内容。

a）主干客运方式概况。以旅客轨迹数据为基准，根据出行方式识别算法构建相关模型，推算出不同的主干运输方式下旅客的属性特征和行为特征等，帮助需求方对单一客运方式进行分析，主干客运方式统计表说明见表5。

表5　主干客运方式统计表说明

序号	数据指标	口径定义
1	各出行方式旅客数量占比	区分铁路、道路客运、航空、自驾及其他，由出行方式识别算法得出
2	各出行方式旅客基本属性	包括性别和年龄维度，年龄分布：<18，19~30，31~40，41~50，>50
3	各出行方式旅客出行目的	区分返乡、差旅和旅游，由旅客群体划分算法得出

b）旅客联程出行现状。基于对各单一主干客运方式的分析，对空空、空铁、空巴、铁铁、铁空、铁巴等联合使用2种出行方式的旅客情况进一步展开分析，将其存在联程出行特征的轨迹拆分出来，统计其距离、时长、速度等核心指标数据，并对旅客联程过程中的换乘次数、时长等影响旅客满意度的指标单独进行统计计算，旅客联程出行统计表说明如表6所示。

任意选取一日数据，以2021年4月1日为例，我们将上述数据指标按照梳理好的口径定义进行了计算，具体计算结果如表7所示，基本能与其他渠道相关数据保持一致。

4.3 春运期间客运专题分析

2019年春运大数据分析工作中，相关研究机构首次增加了旅客联程运输场景下的分析并得到了初步的研究成果，对于当年以及后续的春运工作安排也起到了一定的辅助作用。

2021年春运从2021年1月28日到2021年3月8

表6　旅客联程出行统计表说明

序号	数据指标	口径定义
1	联程出行旅客占比	当日产生2种及以上出行方式的人口总量/全国客运总量
2	联程方式占比	区分空空联运、空铁联运、空巴联运、铁铁联运、铁空联运、铁巴联运等常见联运方式，并进行比例计算
3	联程出行距离	符合联程出行特征的出行链起讫点空间距离
4	联程出行时长	符合联程出行特征的出行链起讫点时间差
5	联程出行平均速度	联程出行距离/时长
6	联程出行换乘次数	产生多种出行方式的次数-1
7	联程出行换乘时长	枢纽场站迁入迁出信令切换时长
8	热门联运路线	按主流联运方式的旅客数量进行线路排名

表7　2021-04-14主干客运及旅客联程出行数据分析

数据指标	计算结果示例					
各出行方式旅客数量占比/%	铁路	6.30				
	道路客运	31.40				
	水运	0.50				
	航空	1.50				
	自驾	60.30				
航空方式旅客基本属性	<18岁	19岁~30岁	31岁~40岁	41岁~50岁	>50岁	
	30 423	353 281	220 855	117 024	81 400	
各出行方式旅客出行目的		返乡	旅游	差旅		
	铁路	48%	5%	47%		
	客运	37%	5%	58%		
	水运	48%	5%	47%		
	航空	12%	8%	80%		
联程方式占比/%	空铁	7.49				
	空巴	11.65				
	铁巴	80.86				
联程出行距离区间分布/%		<300 km	300 km~600 km	600 km~1 000 km	1 000 km~3 000 km	>3 000 km
	空铁	0	9.5	18.7	69	2.8
联程出行时长区间分布/%		<2 h	2 h~6 h	6 h~10 h	10 h~14 h	>14 h
	空铁	15.9	62	21	1.1	0
	空巴	5.1	53	38.5	3.4	0.1
	铁巴	36.5	45.2	16	2.2	0.1
联程出行平均速度/（km/h）	空铁	517				
	空巴	296				
	铁巴	142				
联程出行换乘时长区间分布/%		换乘2 h	换乘2 h~6 h	换乘大于6 h		
	空铁	64	32.1	3.9		
	空巴	52.7	44.5	2.8		
	铁巴	75.2	20.8	3.9		

日，期间利用中国联通全量数据对客运量等指标进行了统计与预测，并针对春运期间返乡、返程、复工、联运等专题性指标展开分析，辅助春运期间交通运输行业管理部门进行客运监管与交通资源分配调度决策，同时根据出行者属性及偏好特征区分农民工、大学生等不同群体，实现特定群体的出行分析。该部分数据分析结果也得到了主管部门的大力支持和认可，春运旅客联程运输分析统计指标说明见表8。

表8　春运旅客联程运输分析统计指标说明

序号	数据指标		口径定义
1	累计客运量		春运开始截至当日，全国出行人口乘坐交通工具出行的人口总量
2	客运总量预测		基于旅客流量预测模型，根据2019年、2020年春运数据以及春运走势进行预测
3	返乡客运	返乡热门通道	春运开始截至当日，抽取有返乡出行特征的人群OD，对起讫点按旅客数量进行排名，对学生、农民工特殊群体可单独统计
4		返乡运输方式	区分铁路、客运和航空
5		返乡联运方式占比	区分空空、空铁、空巴、铁铁、铁空、铁巴等联运方式，按旅客量进行比例计算
6		反向春运城市排名	关注跨域OD中讫点为一线及新一线城市，年龄<18岁或>50岁的用户
7		城市空城率排名	关注一线及新一线城市，春运期间减少的常住人口与春运前常住人口之比
8	返程客运	返程热门通道	以用户轨迹在除夕当天的最后驻留轨迹为起点，轨迹讫点为常住地城市
9		返程运输方式	返程期间区分铁路、客运和航空
10		返程联运方式占比	返程期间区分空空、空铁、空巴、铁铁、铁空、铁巴等联运方式，按旅客量进行比例计算
11		返程率	返程人群与常住人群之比
12		城市复工率	统计当天常规办公区AOI范围内人群与春运前工作人群之比

以2021年3月8日春运最后一天的数据为例，将上述数据指标按照梳理好的口径定义进行了计算，具体结果如表9所示，为主管部门及时了解春运进展提供了多维数据参考。

5 结束语

电信运营商大数据天然的时空连续性是很多行业大数据所无法比拟的，也是其在交通运输行业中受到多方关注的主要原因，基于时空特性的研究也取得了很多突破性的进展，并得以在实际项目中应用和验证，旅客联运场景便是其一。但由于旅客联运业务本身发展仍处于初期，且电信运营商数据本身和相关数据处理技术也有一定的局限性，双方都有待进一步发展完善。相信在具体应用场景下的不断磨合探索，必然会对2个行业的数字化转型进程有所进益。

表9　2021-03-18旅客联程运输数据分析结果

数据指标	计算结果示例	
累计客运量	850 546 951人	
客运总量预测	1 503 243 329人	
返乡热门通道	武汉市—黄冈市 344 056人	
返乡运输方式	铁路	52.10%
	道路客运	36.10%
	水运	5.20%
	航空	6.60%
反向春运城市排名	成都市	225 372
	广州市	144 652
	深圳市	139 459
	上海市	135 727
	重庆市	135 141
返程热门通道	茂名市—广州市 1 826 128人	
返程运输方式	铁路	53.00%
	道路客运	35.20%
	水运	5.10%
	航空	6.70%
返程率/%	86.6	
城市复工率/%	77.2	

参考文献：

[1] 苏田田. 我国旅客联程运输发展初探[J]. 交通世界，2018(9)：158-159，168.
[2] 施扬峥，李及言，吴殿义. 电信运营商："大而全"数据的价值探索[J]. 国际品牌观察，2020(33)：57-61.
[3] 杜娇. 基于手机通讯记录的居民出行信息提取方法研究[D]. 石家庄：石家庄铁道大学，2015.
[4] 闫超，龚露阳，李达标，等. 基于手机信令数据的旅客联程出行时空特征分析方法[J]. 交通运输研究，2019，5(6)：36-42，49.
[5] 曹晓蕊，赖丽娜，孟品超. 基于手机信令数据的用户出行方式识别[J]. 长春理工大学学报(自然科学版)，2021，44(3)：134-142.
[6] 肖志权. 基于手机信令数据的受灾人口快速计算方法研究[D]. 济南：山东建筑大学，2020.
[7] 龚露阳. 我国旅客联程联运发展关键问题及思路[J]. 交通标准化，2014，42(15)：100-102，108.
[8] 陈琳. 京津冀城市群枢纽间旅客联程出行行为研究[D]. 北京：北京交通大学，2020.

作者简介：
许致远，毕业于北京邮电大学，高级工程师，硕士，主要从事创新业务数字化运营管理工作；张慧，毕业于北京航空航天大学，解决方案工程师，主要从事行业洞察与解决方案支撑工作；张鹤，毕业于西南大学，主要从事交通大数据产品规划与设计工作。