文章编号：1672-5328(2019)03-0053-14　中图分类号：U491.1+2　文献标识码：A　DOI:10.13813/j.cn11-5141/u.2019.0307

大数据时代的交通模型

Cuauhtemoc Anda[1], Alexander Erath[1], Pieter Jacobus Fourie[1]　著，宗　晶[2]　译
(1.苏黎世联邦理工大学未来城市实验室，新加坡ETH中心，新加坡 138602，新加坡；2.中国城市规划设计研究院，北京 100037)

摘要：通过新的大数据来源诸如手机通信记录、智能卡数据以及社交媒体地理编码记录，可以前所未有地观察和了解出行行为的细节。尽管有如此庞大的大数据来源，但在规划实践中使用的交通需求模型，其数据源仍大多来自交通调查和人口普查等传统方法。对近期利用大数据研究交通行为，以及使交通规划师可以进行假设情景分析的交通需求模型的最新进展进行梳理。从出行识别到出行活动推理，回顾和分析现有数据分析方法，这些传统方法使收集到的出行轨迹信息能响应交通需求模型。未来的研究应该侧重将概率模型和机器学习技术应用于数据科学。设计这些数据挖掘方法是为了处理由手机移动追踪数据衍生的零散和掺杂偏差的数据的不确定性。此外，这些方法还适用于将不同的相关数据组整合到一个数据融合方案中，以便用出行日志信息丰富大数据。总之，建模知识已经在交通运输领域发展成熟，因此强烈建议在交通规划方面应用数据驱动方法时应建立相应领域专业知识的基础。这些新的挑战呼吁交通模型师和数据科学家之间的多学科协作。

关键词：大数据；交通规划；出行需求建模；基于个体仿真；智能公交卡；手机网络数据

Transport Modelling in the Age of Big Data

Written by Cuauhtemoc Anda[1], Alexander Erath[1], Pieter Jacobus Fourie[1], Translated by Zong Jing[2]
(1.ETH Zurich, Future Cities Laboratory, Singapore-ETH Centre, Singapore 138602, Singapore; 2.China Academy of Urban Planning & Design, Beijing 100037, China)

Abstract: New Big Data sources such as mobile phone call data records, smart card data and geo-coded social media records allow to observe and understand mobility behaviour on an unprecedented level of detail. Despite the availability of such new Big Data sources, transport demand models used in planning practice still, almost exclusively, are based on conventional data such as travel diary surveys and population census. This literature review brings together recent advances in harnessing Big Data sources to understand travel behaviour and inform travel demand models that allow transport planners to compute what-if scenarios. From trip identification to activity inference, we review and analyse the existing data-mining methods that enable these opportunistically collected mobility traces inform transport demand models. We identify that future research should tap on the potential of probabilistic models and machine learning techniques as commonly used in data science. Those data-mining approaches are designed to handle the uncertainty of sparse and noisy data as it is the case for mobility traces derived from mobile phone data. In addition, they are suitable to integrate different related data sets in a data fusion scheme so as to enrich Big Data with information from travel diaries. In any case, we also acknowledge that sophisticated modelling knowledge has developed in the domain of transport planning and therefore we strongly advise that still, domain expert knowledge should build the fundament when applying data-driven approaches in transport planning. These new challenges call for a multidisciplinary collaboration between transport modellers and data scientists.

Keywords: Big Data; transport planning; travel demand modelling; agent-based simulation; public transport smart card; mobile phone network data

收稿日期：2017-12-01

作者简介：Cuauhtemoc Anda，男，博士，数据科学家，城市计算工程师，主要研究方向：基于大数据的交通规划仿真。E-mail: anda@arch.ethz.ch
译者简介：宗晶(1985—)，女，河南洛阳人，硕士，工程师，主要研究方向：交通规划。
E-mail: 13199702960859l6@163.com

文章来源：International Journal of Urban Sciences，2017年第21卷第1期增刊，第19-42页，Taylor & Francis Ltd.(www.tandfonline.com)文章链接：http://dx.doi.org/10.1080/12265934.2017.1281150

0 引言

在人们通过移动电话、公共交通智能卡或者全球定位系统支持的设备获取出行数据之前，构建交通需求模型生成大规模甚至全民样本困难且昂贵。这些模型的目的不仅是为了复制相关精度的实际交通流，还包括应用假设场景来评估不同基础设施开发决策的影响。

尽管有新的大数据来源，但在规划实践中使用的交通需求模型，几乎完全是基于交通调查和人口普查等传统数据。在过去几十年里，随着计算能力的成倍增长，所应用的统计模型变得更加复杂，最重要的变化是从基于出行的模型到基于活动的模型的进化：基本模型构架保持不变，用覆盖一个小样本人群的出行调查推算实际人口的出行情况。

通过新的大数据源，如手机通话记录、智能卡数据和社交媒体记录的地理编码，观察和理解前所未有的交通行为的细节。但是简单的观察对于规划目标没有特别的帮助。考虑到是在假设情景中进行预测，我们需要将大数据中包含的信息置于假设背景中理解，以使大数据信息能适合于交通需求建模框架，并预测交通需求模型。

本文回顾了近年来利用大数据分析交通行为领域的最新进展①，并介绍了可以预测假设情况的交通需求模型。为此，首先介绍最新的交通需求模型的基础，包括最新的基于个体的方法。

本文聚焦于利用相关大数据的研究，重点研究与交通需求模型直接相关的方法和工具包，因此不包括那些从大数据源获得实时分析的方法。本文的目的是从方法论角度全面梳理大数据如何提高对出行的理解以及如何应用于交通需求模型。基于此，确定了各种方法的优点和缺点及其在交通预测模型中的适用性。本文得出的结论不仅包括对应用大数据建模的描述，还提出了弥补研究不足的技术要求。

1 交通需求模型和出行数据

1.1 交通需求模型

交通需求模型通过预测不同交通和土地利用方案的影响情况[1]为决策制定提供支持。有两种交通需求预测方法：1)集计模型，将交通需求定义为小区间的集计交通流；2)基于个体的模型(agent-based model)，在整个模型中保留个体层面的出行需求。

1.1.1 经典的四阶段法

20世纪60年代引入四阶段法需求模型[2]。最初被描述为基于出行的模型，其目的是预测不同交通方式、交通线路在任意两个OD小区之间的出行次数。第一步出行生成是每个小区产生和吸引的出行估计次数。第二步出行分布是连接起点和讫点之间的出行量，以及它们被吸引到哪里。第三步方式划分决定了每次出行的交通方式，如汽车或公共交通。第四步是预测每次出行会选择的路线，并模拟拥堵引起的交通延误。由于这种延误不仅影响交通方式和路径的选择，还影响区域选择行为，通常包括步骤二、三和四的反馈与循环。四阶段法的建模数据要求包括家庭出行调查信息、人口普查信息和交通网络信息。

1.1.2 基于活动的模型

20世纪90年代初以来，基于活动的模型被认为是优于四阶段法的选择，避免了四阶段法的固有局限性。为了解基于活动的模型的重要性，文献[3]强调四阶段法在本质上是集计的，即测量单位不是个体，而是来自任何特定小区的出行。此外，四阶段法在如何使用不同子模型的行为参数上缺乏稳定性和一致性。而且，当涉及交通需求管理政策，如出行定价政策的评估时，四阶段法的独立性假设经常被视为致命缺点。

基于活动的模型的基本原理是理解出行需求源于参与活动的必要性。基于活动的模型的目标是预测每个个体在受到时间和资源制约的前提下，一定时间周期内的活动次数、顺序和类型。然而，基于活动的模型允许在空间上对交通需求进行分解描述，通过路径选择和交通仿真，这种交通需求通常会再次集计成所谓的OD矩阵，以描述在任意两个起讫点之间的出行次数。这种限制起初是由于模型缺乏模拟交通的相关空间范围，即全天时段整个城市或区域，但是如今仍旧适用于基于交通仿真的计算要求。

除了与四阶段法相同的数据要求外，基于活动的模型还需要一种额外的输入数据，即在单个家庭和个人层面的“虚拟人口”以及代表地区利益的实际人口数据。这种虚拟人口包括一系列社会人口属性，可被用于交通需求建模过程中。此外，对于虚拟人口中的每一个个体，都有一个完全描述性的日常

活动计划，包括工作或教育等日常活动的地点。

用于交通战略规划的基于个体的模型通常从基于活动的建模方法中获得交通需求，基于受到交通网络及其属性约束的系统[4]，采用微观和完全动态的交通仿真模拟个体的个性化需求。

最初，TRANSIMS[5]发展成为第一个大规模用于基于个体的交通模型的建模工具，专注于取代集计的交通分配方法，之后基于个体模型的实现和最新的发展，如MARSim[6]，SimMobility[7]，SimAGENT[8]按照出行方式、时间、目的地和活动调度进行不同程度的整合，形成了一致性的建模框架。这一综合框架使得在整个建模过程中可以非集计的形式模拟交通需求。除了增强行为一致性外，还允许对现代交通需求的管理工具进行建模和分析，如基于时间或需求定价，以及共享汽车和自动驾驶等新的交通形式。

多元个体建模(multi-agent-based modelling)建立在大规模独立个体的基础上，他们执行自己的决策，并与其他个体、环境相互作用。对于个体，一个初始的日常活动计划需要用活动的位置、时间、开始和结束时间以及两个活动间行程，包括交通方式和交通线路精确描述。

在正在发展的几个基于个体的交通模型中，MATSim以一个特殊的作用被认为是目前应用最为广泛的模型。MATSim可以在一个协同进化的学习循环中集成广泛的决策维度，但是受限于模块化框架，它也只能用于交通仿真，并与其他基于活动的出行需求模型结合使用。

1.2 用大数据描述出行

随着移动设备和定位传感技术的普及，精确的地理位置数据代表着巨大且不断增长的大数据集。以交通规划为目的，基于非集计的活动模型，本文仅局限于从个体获取出行数据的相关研究。对个体数据轨迹感兴趣是因为它可以提供关于交通方式更准确、更有趣的视角。除此之外还包括由基础设施检测器记录的出行信息，这些检测器记录了某些交叉口的交通量(如线圈检测器、视频车辆检测系统和ERP系统)。

智能卡自动收费(Smart Card Automated Fare Collection, SC-AFC)系统和移动电话网络在城市中的设置覆盖面广，是本文研究的重点。两者可归类于大规模随机出行检测器，能以前所未有的规模和详细程度提供对城市动态和人们活动的观察。此外，两者还拥有一个优势，即无须额外的基础设施收集出行信息，因为其本身就是为了收集公共交通费用并允许移动通信网络使用。

其他的数据集可作为补充数据，如GPS数据、特征点(Points of Interest, POI)、土地利用、人口普查和交通调查数据。正如文献[9]提到的，补充数据集有三个目的：1)验证基于大规模出行检测器数据的分析结果；2)明确缩放因子，将结果扩样至总体样本；3)增加城市空间信息以获取更深层次的结果。

2 智能卡数据

SC-AFC系统应用于世界各地的许多公共交通系统中，并持续被公共交通运营商使用。公共交通系统引入智能卡的主要目的是利用其灵活性和安全性进行收费。任何(时空)转换产生的信息很快就成为交通和城市规划的丰富数据源。从公共交通客流分析到OD矩阵创建，智能卡数据(为城市动态和出行方式)提供了城市公共交通的洞察视角。下文将介绍从重建个体出行到OD矩阵预测，如何利用智能卡数据及使用基于个体的建模方法进行交通规划。

2.1 个体出行的重建

SC-AFC系统的实施取决于城市及其票价政策。阿姆斯特丹、悉尼和新加坡等城市根据公共交通出行的总里程收取车费，而不管是使用公共汽车还是火车。这就要求乘客上车、下车或者换乘时刷卡。然而，伦敦、旧金山等城市则实行非阶梯票价，即无论在哪里上下车，全线票价相同，因此乘客上下车只需要刷一次卡。在任何情况下，为进一步分析人们的出行活动，挖掘智能卡数据的主要挑战在于重建个体出行。

2.1.1 预测下车站

由于SC-AFC系统只要求验证上车站，因此第一步是预测下车站。一般来说，可以基于两个明确假设使用出行链(Trip-Chaining)算法推断下车站[10]。第一个假设是在出行结束后，出行者将回到之前下车站；第二个假设是在一天结束时，出行者将返回当天第一次出行的上车站。

针对文献[10]提出的初始概念，一些研

究对其进行了改进。文献[11]将这一概念扩展至轨道交通和公共汽车的换乘线路中。文献[12]尝试整合第二天甚至一周的出行方式，以补充魁北克市(Quebec)加蒂诺(Gatineau)公共交通系统的信息缺失。文献[13]提出了一种利用时间约束而非距离约束的多方式公共交通的预测方法。在这些研究中，个体出行重建的成功率从66%提升至80%。

此外，文献[14]提出了基于概率无向图模型(undirected graphical probabilistic model)通过智能卡数据重建个体出行的方法。该文献提出了一种集成学习方法，将费用、地理空间和时间空间(geospatial and temporal spaces)结合起来，从而推断出一系列关键领域特定的约束因子。通过使用在这些约束条件下的半监督随机算法，可推断出确切的上下车站，即使存在未知信息的交通记录。只有10%的出行有明确上下车站数据，超过78%的出行存在上下车站信息缺失的情况。这项工作的实用性不仅仅是重建仅有出行起点的出行过程，而且是一个通过智能卡刷卡记录恢复个体出行历史的系统方式。这个预处理阶段可以有效地支撑后期交通需求模型的构建和分析。

3.1.2 阶段、行程和OD

确定下车站后，个体出行重建的第二步是推断这个下车站是否是最终目的地(即行程结束)，或只是一个多阶段行程的一个阶段(即换乘)。常见的识别方法是利用时间法则。例如，文献[13]使用30 min的时间法则。如果一个人在某一个特定地点停留超过30 min，即可认为该地点是目的地。在伦敦的案例中，时间阈值取决于交通方式，即地铁换乘公共汽车为20 min，公共汽车换乘地铁为35 min，公共汽车间换乘为45 min[15]。

只有智能卡数据才能获取时空维度上的个体活动，这就限制了识别个体活动的渠道，因为一天的行程不全都是使用公共交通。文献[16]描述了公共交通出行一致性概念的局限性，一致性意味着同一个人通过公共交通到达活动地点，那么就必须通过公共交通结束此次行程。然而智能卡数据不能记录公共交通以外的其他交通方式，通过分析最后一段行程的下车点和接下来一段行程的上车点能明确识别是否为统一的交通方式。这就可以确定在两段行程之间是否还采用其他交通方式，如出租汽车、小汽车或者步行。

以新加坡的一个典型工作日为例，文献[16]发现在智能卡数据中记录的不只有一次行程的人群，90%的出行开始于上一次下车点1 km范围内。这说明：1)大多数公共交通出行者在多次公共交通出行之间并不会使用其他交通方式，因此他们的出行链较连贯；2)有可能一个区域只存在特定种类的活动。

一旦个体出行被重新构建到已知的起讫点上，应用程序就可能把这次行程加入公共交通OD矩阵。针对那些无法重建的行程，建立扩展因子是典型的解决方案。文献[13]显示了如何在没有目的地的前提下为智能卡数据构建扩展因子，以及推测没有起点或者刷卡记录的数据分布规律。对于前者，假定行程的分布与其他相同起点的行程一样，而对于后一种情况，假定行程的分布只与他们的时间分配有关。

2.1.3 初级活动鉴定

通过进一步研究公共交通稳定出行可以增强对可能的活动地点的解释。文献[17]提出了一种基于规则的直接分类方法，包括卡片类型信息和行程的时间属性。工作目的对应成人卡，指活动时间超过2 h且活动前的出行不是当天的最后一项行程。上学目的对应学生或者未成年人的卡片，指活动时间超过5 h且该活动也不是当天最后一项活动。最后，回家目的指活动结束后的出行是当天的最后一项行程，其他的行程将被分配到其他活动目的。

文献[18]是最新的基于规则的研究。该研究包括一项空间规则，通过预先识别用户家庭所在车站判断基于家的出行频率和出行距离。基于此，研究扩展了文献[10]的假设：1)一天中，最后一段行程的终点站通常与第一段行程的起点站一致；2)第一段行程的起点站通常与前一日最后一段行程的终点站相同；3)对于大多数乘客来说，第一段行程的开始和最后一段行程的结束都在家附近。通过这些假设，研究构建了一个运行平均算法，称之为基于中心点的检测算法(center-point based detection algorithm)。该算法的主要优点是操作简单且方法稳定，从某种意义上说，它可以识别一天出行一次的用户的家庭所在车站(例如不稳定出行)。

尽管被认为是一个简单的操作，但是当试图扩展约束条件时，基于规则的活动计算效率变得低下，更不用说在详细规则中需要手动操作时的效率。此外，结果的准确性可能会受影响，特别是在识别其他如工作和次

要活动等更为灵活的活动模式的情况下。通过引入概率(选择)模型，可以改进这种严格分类的缺点。

文献[16]提出了一个以活动持续时间、活动开始时间和土地利用作为效用变量的多因子Logit模型，以匹配离散选择空间，包括工作活动、家庭活动和其他活动等目标。分段线性函数是构建模型的实用工具。对于活动持续时间和启动时间，利用当地交通调查信息对效用函数进行校准，而对于土地利用，校准信息依靠来自城市规划部门的总体规划。

文献[19]提出了另一种概率模型方法，建立一个连续空间模型来确定家庭和工作地点。研究引入了一个得分函数，通过对一组受过训练的使用者进行逻辑回归和标定得出。与文献[16]类似，家庭和工作地点标签主要是由与事件相关的时间因素确定。然而，两种概率模型方法之间的主要区别不在于他们是否选择离散或连续空间，而是标定过程中迁移学习方案(transfer learning scheme)[16]使用多源数据(居民出行调查)，传统学习方案使用单一来源被标记的数据子集[19]。

最后，文献[20]呈现了概率模型在无监督模式下接受训练(即没有标记的例子)的情况，以识别智能卡记录的活动模式。文献通过提出一个连续的隐藏马尔科夫模型(Hidden Markov Model, HMM)，发现8个集群被按照家庭活动和家以外活动描述为不同的模式，其内部结构的释放概率是一个混合高斯模型。这个模型的优点在于不仅能找到新的观测对象在集群中的成员关系，还能生成活动链来构建虚拟人口。虽然该模型展示了在出行数据中发现活动模式的方法，但是如果只想获得基本活动，那么就不清楚其基于规则方法的区别(如文献[8]的实际优势)。

对活动预测结果进行完全验证几乎无法做到，因为在智能卡记录总量中，这需要个体拥有完整的行程信息。由于这个原因，用部分验证来确定模型的准确性。例如，一种常见的方法是将识别的热点区域数据得到的结果，直观对比家庭出行调查和人口普查[14]。

2.2 基于个体的交通模型与仿真

智能卡数据的非集计特点体现为基于多元个体的交通模型的适当输入。假设每个独特的智能卡信息代表一个个体，交通需求可以直接从智能卡数据中获取。

文献[21]在阿姆斯特丹和鹿特丹第一次尝试实施基于个体的公共交通微观仿真。仅仅基于智能卡数据，工作的主要挑战是个体活动计划的生成。研究聚焦于同一个通勤者连续几天基于家的出行模式。工作和家庭所在车站被认为是工作日期间使用最多的两个车站，周末期间家庭所在车站客流量最大。智能卡身份信息并不与这一模式完全吻合，但通过在出行中间站引入虚拟活动来重建某个特殊日的活动链，以描述当天的交通需求。最后，对于高度不规则的交通模式，每一次出行都会单独生成。

生成虚拟人口的过程受到各种制约，主要是建模过程中的各种假设。未来研究的机遇在于通过更准确、更有效的实际交通需求来确定出行目的和社会人口特征。为此，可将对智能卡数据的长期观察看作是应用现代数据挖掘技术来推断额外信息的机会。沿着这个思路，文献[22]探索了如何将特征行为(eigenbehaviours)的概念[23]应用于推导时空模式。

使用智能卡数据进行仿真的另一个挑战是将公共交通工具与其他交通方式(如小汽车)之间潜在的相互作用进行建模。最近，文献[24]的一项研究为新加坡公共交通开发了一种简化的基于个体的交通仿真。不同于文献[21]，在连续两个车站间，通过一个随机公共汽车速度模型(stochastic bus speed model)取代MATSim队列模型来解释与私人小汽车的相互影响。该模型根据一个多项式回归模型拟合，假设车站到车站的运行速度遵循正态分布[25]。正如文献[26]指出的，在交通网络中决定小汽车速度的各项参数不仅与(从智能卡数据中获取的)需求有关，还与网络描述中的地理信息有关。为说明仿真框架中存在的停留时间的易变性，他们考虑了文献[27]研究的模型。

以简化的交通仿真方案为例，说明机器学习如何替代MATSim模型。智能卡记录的统计数据是用来训练模型的，而不是从多元个体仿真中获得公共汽车出行时间。结果不仅大大提高了仿真时间，而且使仿真系统网络的重新设计成为可能。尽管如此，仍有一些限制因素需要解决，例如轨道交通轨迹的重建，对步行、等待和换乘活动更好的表达，这些活动并不能直接从智能卡数据中获取。

3 手机数据

无论GSM，CDMA还是LTE，移动网络需要手机和蜂窝网络之间进行定期和频繁的交互信息(例如脉冲信号)。为了给用户提供服务，移动网络需要频繁的对手机进行定位，即使手机处于待机状态。通过附近的基站计算用户的位置，这一结果的精度相当于在市区几百米范围内的基站覆盖的大小。通过网络触发和事件触发更新手机定位信息。

网络触发定位更新发生在：

1）手机连接到蜂窝网络；

2）在两个不同区域之间进行呼叫和移动(例如切换)；

3）待机并移动到属于新位置区域(Location Area, LA)的网格；

4）当相关计时器已经结束，则网络进行调查(例如定期位置更新，通常每2 h更新一次)。

时间触发定位更新发生在下列情况：

1）拨打或接听电话时；

2）使用短信服务(发送和接收)；

3）用户连接到互联网(如浏览网页或发送电子邮件)。

由此，从移动网络中获取的位置更新数据构成了日常活动和交通模型的潜在信息来源。与家庭调查相比，手机数据提供了大样本量和长时间的观察周期，而成本可以忽略不计。然而，人们必须克服处理移动电话轨迹以应对出行重建的挑战，因为这类数据流中包含的信息的空间分辨率和时间分辨率都很低。具体而言，位置估计值的精度取决于给定区域内的基站的分布，而位置更新的频率则取决于用户的使用情况。因此，普遍的挑战是如何从稀疏和杂乱的监测数据中提取人们出行的丰富语义(例如出行目的)[28]。

3.1 手机数据挖掘通道

文献中出现的第一个方法是根据话单数据(Call Detail Records, CDRs)生成基于出行流的OD矩阵[29-31]。由于OD矩阵是通过捕捉来自不同交通分析小区的突发流产生的，而不是个体出行重建过程，这些方法不符合个人活动的需求。此外，文献[32]讨论如果手机数据的空间分辨率低，前面的方法会存在偏差。另外，它们并非用于处理移动电话原始记录的偏差，如所谓的超音速跳跃(supersonic jumps)或信号跳跃(signal jumps)(即离群值)。这些事件都是短时间内突发的事件。虽然这种跳跃通常是系统固有的数据偏差，但一些跳跃可能是由外部机制触发的，目的是保护用户的隐私[33]。

由于上述原因，需要一条数据挖掘管道，从移动电话位置更新中提取确切的个人行程。首先，需要一个预处理阶段去处理偏差测量和基站间信号跳跃。其次，个人行程提取阶段，可以分割停留位置(即活动片段)，由此估计行程的开始和结束时间。第三，活动或出行目的地推测阶段，用于估算家庭、工作、学校等主要活动地点以及餐饮、购物等次要活动地点。

3.2 预处理技术

对于第一个目标，文献[33]对三种不同类型的滤波器进行评估，以检测移动电话轨迹数据的异常值：递归原生滤波器(Recursive Naive Filter)、递归超前滤波器(Recursive Look-Ahead Filter)和卡尔曼滤波器(Kalman Filter)。一方面，前两种主要表现为低通滤波器[28, 34]。它们通过引入出行速度的上限约束来消除较大的定位误差。因此，可以通过每一对连续的点(递归原生滤波器)或者每一个三合点(递归超前滤波器)计算速度，并与特定阈值相比较。另一方面，卡尔曼滤波器是重建轨迹的概率方法。结果表明，在排除异常点的情况下，递归超前滤波器的效果更好，并保持了轨迹的准确性。虽然卡尔曼滤波器也消除了异常点，但轨迹失去了准确性。然而，文献[35]通过使用高斯混合模型来扩展现实挖掘数据库[36]的空间分辨率，考虑到话单数据的低分辨率，需要更复杂的概率滤波器来替代原生滤波器。

文献[36]提出了另一种专为处理手机数据偏差开发的预处理技术。首先利用基于密度空间维度的聚类方法解决基站间跳跃的问题，以确定可能的停留点，包括来自基站间跳跃数据的虚构停留点。然后，通过几乎相同的时间戳识别出波动图。最后，通过选取个人花费更多时间的集群，过滤掉震荡点(例如虚构的集群)。这种方法可作为移动通信数据的时间解决方案。

3.3 停留点提取

基于时间规则(temporal-based rules)的研究层面：文献[37]研究德国西南部一个地区的位置区域更新情况。该算法提出的原则

是，如果用户在位置区域停留的时间比直接穿过该区域所需的时间更长，那么用户在该位置区域可能会开始或结束一段行程。为此，研究提出了60 min原则，如果第一次登入信息和最后一次登出信息的时间间隔大于60 min，则认为该位置区域是一个停留点。当然，由于提取的行程信息在一个大的位置区域层面，而不是在基站区域层面，故该方法受到一些限制。

基于距离聚类(distance-based clustering)的研究层面：文献[34]提出一种基于从电话、短信和互联网使用中生成的话单数据来识别基站塔层面的出行的方法。在预处理阶段，应用一个低通滤波器，以10 min一次的采样率来解释信号的跳跃；应用一个低级别的距离聚类技术，识别一个共同位置附近的小波动，并理顺移动电话追踪轨迹。为了提取停留点，对1 km范围内的融合点进行基于距离的聚类分析。集群的质心被定义为一个虚拟位置，在最后一步中，通过将标识的虚拟位置连接起来重建个人路径。然而，由于一个虚拟位置可在一个临时事件中创建，因此该方法缺乏对事件的可靠过滤。

基于频率聚类(frequency-based clustering)的研究层面：文献[19]提出从时间分布稀疏、空间低分辨率分布的话单数据中识别停留位置的方法，认为被访问最多的基站是一个人生活中的重要场所。文献没有使用时间或空间聚类算法来获取这些位置，而是使用手机基站访问数据。该方法包括应用集群引导算法(cluster leader algorithm)，根据联系手机基站的总天数对其进行排序。这种方法适用于低分辨率的跟踪和长时间的观测。然而，只有主要活动和一些次要活动地点可以被识别。

时空聚类(spatio-temporal clustering)层面：文献[32, 38−40]利用时间和距离聚类技术过滤经过基站的数据。首先，通过测量两个相邻点之间的距离，并与距离阈值进行比较(例如漫游300 m)，从而在空间上进行分组。其次，如果第一次和最后一次观察之间的时间间隔大于时间阈值(例如10 min)，则认为可能存在停留。然后，潜在的停留点被设置为集群中的质心。由于位置上的偏差，在不同的观测日和不同的地理坐标下可能会有多个潜在的相同位置。考虑到这一点，最后不考虑记录的时间顺序利用聚类算法确定停留区域。

同样，文献[41]使用了基于密度的聚类算法(即漫游距离)，其ε参数取值为100 m，时间阈值为5 min，以此过滤出通过点。与基于频率的聚类算法相比，只要基于密度聚类算法的时间分辨率不稀疏(例如数据集包含网络更新数据)，时空聚类算法就能检测到任何活动的位置。

行程验证(trip validation)层面：因为有更多手机用户在出行行为中没有系统差异，所以有必要对算法进行验证。例如，检测到的地点数量与手机使用之间不存在相关性。文献[32]根据手机使用频率将用户分为五组，检查各组每天的日常出行情况，包括出行次数、不同目的地的数量。通过比较上述数据的频率分布，得出这些数据有相似模式的结论。

活动开始时间和持续时间(activity start times and durations)层面：确定停留位置后，文献[28]接下来将预测到达时间，方法是计算到达活动记录的最早值(即到达时间的上限)与下限值的平均值，对上一个位置的最后记录时间以及上一个位置与当前位置之间的出行时间求和可预测时间下限。行程时间被确定为连续的中心点之间的距离除以假设的旅行速度。在预期的出发时间内执行相同的过程，活动持续时间通过减去估算时间计算得到。

文献[40]用另一种方法推断出活动的到达、离开时间。文献建议使用从全国家庭出行调查中得出的出行持续时间概率函数。为工作日和周末构建6 h出行分布和对应的出行目的：基于家的工作出行(home-based work, HBW)、基于家的其他出行(home-based other, HBO)和非基于家的出行(non-home-based, NHB)。然后，在观察的时间窗口中随机生成离开时间，得到对应的时间(工作日、周末)和出行目的(HBW、HBO和NHB)分布。

3.4 活动推测

在传统的调查数据中，活动目的由被调查者提供，而在手机数据中，活动类型是设定好的。此外，没有任何数据来源(交通调查或者手机数据)能准确地确定出行目的地的确切位置，但是这些精确的位置在一片区域内。一般来说，我们可以在文献中找到分布预测的两种不同的方法，即时间频率模型和概率模型。

3.4.1 基于时间-频率规则的活动推测

推断背景信息例如位置函数或访问目的，其直接方法之一是通过时间-频率规则来推断。文献[32, 39-40]改进了文献[30-31]在使用访问频率和时间数据识别工作、家庭和其他地点的总体思路。一个用户的家庭位置被定义为在工作日和周末20:00至次日7:00之间最常观察到的停留点。然而，工作地点被定义为在工作日7:00—20:00停留最多的地点。由于有些人不工作，如果一些位置每周访问不超过1次，或者地点离家不超过500 m(为了避免通过信号偏差识别出错误的工作位置)，工作地点就会留下空白。另一种变化[40]是工作地点被确定为用户从家庭移动的最大距离的停留点，以此来识别夜班工作。

3.4.2 基于概率模型的活动推理

用于推断活动(出行)目的的时间-频率规则是一种直接的方法，但是对某些群体可能不适用。此外，它们仅限于在主要活动位置的某些模式。通过概率模型推理是更可靠的方法。概率方法用于处理观察中的不确定性，并捕获模型解释变量之间的相互依赖关系。这使其他相关数据集在模型中集成，例如语义丰富的地理信息数据，以提高结果的准确性，并允许对更广泛的活动类别进行分类。

推导概率模型的一个有力工具是概率图模型(Probabilistic Graphical Models, PGM)。PGM是概率分布的图形表示，其中一个节点代表一个随机变量，而连接阶段的边缘显示它们之间的因果关系。通常以图中编码独立和条件独立假设描述因子形式的随机变量之间的联合概率。两个典型的概率图模型是贝叶斯网络(Bayesian Networks，即有向的非循环图)和马尔科夫随机场(Markov Random Field，即无向图)。前者将联合概率分解为条件概率分布，后者根据吉布斯分布(Gibbs distribution)和图中点集(the cliques in the graph)分解。在定义模型表达之后，下一步是找到模型参数。可以通过以下算法得到：最大似然估计(Maximum Likelihood Estimation, MLE)、最大后验概率(Maximum a Posteriori, MAP)或者贝叶斯推断(Bayesian Learning)。例如，期望最大化(Expectation-Maximization, EM)算法是一种迭代方法，当模型依赖于潜在变量(即未被观察变量)时，可以找到MLE或MAP。最后，在推理步骤中，我们试图查询完整的联合概率，例如根据观察所得的活动概率对新观测信息进行分类。推理算法可以分为精确推理算法(如置信传播、MAP推理)和近似推理算法(如变分法)。

1）生成模型(generative models)。文献[42]通过建立贝叶斯网络将出行分为五种不同活动类别：家庭、工作、休闲、购物和其他。模型中的解释变量包括：开始时间、持续时间、每个停留位置以及当前和上一次活动之间的转换概率。通过家庭出行调查对模型进行标定，并进行逐步分类。首先区分家庭、工作和其他出行；其次进一步将其他活动分为休闲、购物或其他。该方法分类成功率达到79.4%。

文献[41]采用输入-输出隐藏马尔科夫模型(Input-Output Hidden Markov Model, IO-HMM)，解释了话单数据的活动模式。IO-HMM不仅允许潜在变量(即不同的输出变量)中包含多个观察值，而且还允许潜在变量的识别不仅基于之前的活动还要基于一些环境信息变化(即不同的输入变量)。为达到这一目标，首先用3.4.1节中定义的一组相似的时间频率规则确定主要活动地点(家和工作)，然后用IO-HMM推断次要活动。模型输入的信息代表向一个新活动转移的起始点信息；因此，这些数据被定义为一天中的某一时间、一周中的某一天以及工作时间的累积变量。与此相反，模型的输出信息包括向新活动转移时未能获取的信息：与家的距离、与工作地点的距离、活动持续时间和该地点以往是否被访问过。与文献[42]相反，模型在无监督的情况下采用EM算法(例如不用标签的案例)进行调试。确定8个不同的活动集群：家庭、远距离出行、中等距离出行、娱乐、买咖啡或等车(coffee/transport)、个人事务、就餐或购物以及工作。

这两种方法[41-42]可以进一步被归类为生成模型，因为它们用随机变量建立联合概率模型。生成模型的一个重要好处是，它们不仅可以用来对新的观察进行分类，还可以生成样本和创建虚拟人口，从而进一步作为基于活动的模型的需求输入信息。

2）判别模型(discriminative models)。判别模型是无方向图，而不是模拟联合概率，直接将 $p(Y/X)$ 的条件概率建模。当我们只关注观察到的特征的目标变量(例如活动)，则有适用的模型。由于判别模型并不对特征之间的关系进行模拟，这些模型允许包含更

多重叠特征来完善分类任务。文献[28]提出马尔科夫逻辑网络(Relational Markov Network)，揭示手机数据中的活动时空结构。MRN是马尔科夫随机场的扩展，它是为关系数据库中的集体分类而设计的。值得注意的是，文献[28]根据土地利用类型、活动持续时间、开始时间的分布概率进行模拟，求得这些活动之前是否被访问，活动是否有一个特定位置，以及在检测位置只显现出一个活动情况下的离散变量。该模型采用无监督的方法进行测试，采用EM和拒绝抽样(Rejection Sampling)方法进行推理，计算土地利用和活动类型的后验分布。

文献[28]的结论是，由此产生的集群反映了与传统调查数据吻合的出行链和活动调度模式。此外，对比研究城市(波士顿和维也纳)显示集群具有相似性。尽管如此，还是有一些改进建议。首先，研究传统调查中发现的活动集群与传统活动类型之间的关系。其次，引入POI数据库进一步验证结果。第三，将模型(例如基于个体的模型)预测的交通量与实际交通量进行对比作为验证步骤。

3.5 方式推演

从无处不在的计算设备推演交通方式是不同研究面临的共同挑战。然而，多数建议的方法都是基于手机的传感器，如GPS、加速度计和陀螺仪，因为这些传感器可以进行细微取样。不过，更广泛的分类只能基于话单数据(细节调用记录)。这些方法通过预测移动电话的速度并将其与交通方式相关联来推断出行方式。例如，文献[43]使用出行起讫点信息和旅行时间，将出行方式分成三组：小汽车、公共交通和步行。首先，研究过滤了数据集，只保留超过3 km的出行和更新位置频率超过1次·h^{-1}的用户。然后，按照起讫点进行分组，再通过k均值算法聚类来划分出行方式。最后，用谷歌地图的出行时间信息对结果进行验证。

虽然学术界对话单数据的关注主要集中于活动(出行)目的估计过程，但是，随着智能手机普及率的增长和更多细节信息可供使用(即上网使用痕迹)，将出现能通过话单数据找到特定出行方式或可以融合智能卡刷卡数据等其他数据集的更可靠的算法。这种算法将有助于理解影响方式选择的行为参数。

3.6 虚拟人口和基于个体的仿真

使用手机数据满足基于活动的模型的数据需求是交通规划中的一大希望。然而，目前存在的挑战之一是发现充分利用移动数据的真正益处，以更好的数据挖掘方法获取手机数据和利用机器学习算法开发大数据驱动的基于个体的仿真。文献[44-45]展示了一项初步研究，该研究仅基于手机数据仿真MATSim模型得到虚拟人口。然而，这一虚拟方法存在缺陷，即研究中使用的话单数据不足以代表真实的话单数据。

最新的智慧港湾(SmartBay)项目，尝试基于个体的模型开展交通规划[46]。利用去除隐私的话单数据构建旧金山湾区MATSim模型。包括直接从话单数据派生出需求模型，以及在个体模型人群中赋予特定的社交结构从而模拟不同的出行目的地与方式选择。类似于文献[19]提出的方法，以基于活动频率的插补法来确定主要位置。基于人口调查数据可估算调节过程中的修正系数，其中涉及综合区点插值方法[47]和一种优化的迭代比例拟合结果。与原有湾区都市区交通需求模型比较发现，城市的发展变化十分明显，尤其是硅谷IT部门的快速成长导致城市就业分配的巨大变化。

智慧港湾项目目前正在推进，未来计划包括文献[41]提到的为推演次要活动设计的生成模型，结合机器学习工具对同一次活动的目的地选择建立扩展模型，并在方式选择中引入社会影响。

4 讨论

4.1 大数据驱动下基于个体的交通规划建模

传统的交通预测数据来源于家庭出行调查，该调查具有不可否认的价值。它们不仅包括个人和家庭成员出行模式的详细数据，还包括出行方式和出行目的等相关信息。然而，它们不能完全反映基于个体的交通建模的优势。这里存在两个主要的限制：1)家庭出行调查仅代表了一小部分人群(通常约1%)；2)家庭出行调查通常每5~10年更新一次②。

便携式移动传感器克服了这些弊端，并成为继续开发基于个体的交通规划模型的有效途径。其弊端是这种广泛收集的随机信息是未经处理的原始数据，需要进行额外的分

析工作才能确定出行和出行目的，以便在基于个体的仿真中进行整合。因此，关键的挑战是开发鲁棒性算法和设计一种数据挖掘方法，从稀疏的出行跟踪数据中提取个人每日行程安排。

4.2 从GPS到话单数据模型的可转移性

当使用稀疏的话单数据来提取活动时，其中一个方向是采用最初为GPS数据开发的方法。例如，文献[28]将基于话单数据的活动推理用在文献[48]提出的马尔科夫逻辑网络中，最初用于GPS追踪；而文献[14]和文献[41]应用随机场条件模型(Conditional Random Fields)[49]处理智能卡数据，应用隐藏马尔科夫模型[50]处理话单数据。其中一个原因是，在不考虑活动识别的前提下，GPS轨迹已经成为众多研究中的主要研究对象[51-53]。因此，一个重要的研究问题是，这些模型多大程度上适用于低分辨率的出行轨迹，例如手机话单和智能卡提供的数据。此外，除了出行轨迹在粒度级别上的差异外，基于GPS的研究通常有一个带有活动标签的受控样本；因此，通常情况下模型以监督的形式接受训练。对于话单数据，这样的训练样本不易获得。

因此，这些模型应该依赖于无监督学习和半监督学习方法。最后，另一个需要注意的重要问题是，基于GPS的活动推理模型通常在小样本范围内被训练和验证(例如文献[49]中的4个人)。这无疑加重了对模型表现的质疑，当扩展到城市尺度时，我们不禁会想将这些模型扩展到大规模低分辨率出行轨迹的可能性。

4.3 概率机器学习和交通建模

为GPS开发的活动推理模型由概率机器学习衍生而来，是人工智能(AI)的一个分支。人工智能和机器学习是大数据时代交通建模的高相关性学科。为了解它们的重要性以及适应交通运输工程的方式，我们来看一个简单的例子。想象一下自己如何理解什么是“猫”，我们会回想起一些图片以及在幼儿园里被教会“猫”的概念。尽管一开始可能无法区分猫和老虎，但在观察了几个猫的实际例子以后，我们对于什么是“猫”变得更加清晰。一般来说，得到的数据越多，我们的观念就越坚定，不确定性也越少。

在人工智能中，概率被用作计算人们对这些观念的确定程度。在城市大数据背景下，我们对一种现象及其周围环境拥有大量的观察结果。例如，线圈检测器数据、出租汽车GPS数据、公共交通智能卡数据和手机数据。所有这些信息都可以代表交通运输网络的现状。基于这些观察，通过概率机器学习来计算和提高我们对交通网络的认识。

另一个重要的问题是人们如何使用实用的机器学习和概率模型。通常，人们试图将感兴趣的问题映射到一个标准的算法上，例如线性回归。模型本身限制了我们考虑非相关的解释变量(例如条件独立)。然而，感兴趣的问题可能会更好建模，包括更丰富的解释变量和其他类型的假设。因此，我们更希望有一个框架可以构建最能代表问题的模型。概率图模型即是通过基于模型的机器学习研发的一款面向开发人员的模型框架，目前已提供摘要版[54]。

4.4 解锁不同数据集的知识

在大规模人类移动传感器(如手机话单数据、智能卡刷卡数据)中，低时空分辨率可以得到较长的观察周期或额外的数据集补偿。此外，在大数据时代，人们的愿望是从多个不同但存在潜在联系的数据集中获取知识[55]。例如，从稀疏的话单数据中推断出行目的，其中一个直观的方法是通过包括POIs数据集的模型来丰富空间特征，它可以提供有关某一区域发生的活动类型的信息。该模型支持来自概率图模型框架的跨区域数据融合[55]。

另一个重要的方面是在城市出行环境中应用机器学习的独特挑战。在计算机视觉、自然语言处理等机器学习的领域中，训练集和测试集通常来自相同的集合。例如，一个识别手写数字的模型采用具备相同特征空间的图片进行训练和测试。然而，在城市出行数据的例子中，用不同来源、不同类型的观察来解释相同的现象，我们所需要的能力就是利用所有这些信息生成模型。因此，特别有趣的方法包括转移学习法(从相关领域中提取有趣的知识以帮助学习目标领域)、多视图学习法(通过多个不同的特征集学习)、半监督学习法(使用标记和伪标记的数据来训练模型)。

4.5 数据隐私和市民参与

由于智能卡和手机数据在记录个人出行

模式方面的普遍化和细节化，数据的隐私性越来越受到关注。例如，尽管话单数据去除了隐私数据，文献[56]指出即使只有4个时空点，通过手机天线获取的空间分辨率足以识别95%的个体。

人们在位置混淆不能够重新识别用户身份时，采取了一些措施以便能提取有用的出行模式。这些保护隐私的算法目前由新兴的差分隐私(Differential Privacy, DP)主导。DP是一种数学保障，通过在序列中引入受控的偏差[57]隐藏数据库中的参与用户。预算参数(ε)表示隐私程度和精度之间的权衡。文献[58]扩展了DP位置数据保护的概念。虽然已经证明DP关于某些基于位置和集聚位置信息的服务是有效的[57-58]，但当应用于个人出行轨迹时，DP看起来是对隐私和精度之间的一种折中，且未能达到最先进的技术水准[59-60]。

对于特定的大尺度、多个体交通规划仿真实例，在构建仿真过程的不同阶段都可能出现保护隐私的机制。然而，最终不应期望通过追踪任意个体来仿真还原真实个体的情况。出于这一原因，首要的原则是不能使用真实的总体数据和日程信息，因此需要在集计层面设计行为模式类似真实情况的虚拟人口。来自概率图的生成模型(如贝叶斯网络、隐藏马尔科夫模型)是必不可少的基本工作，因为可以从联合概率分布中提取出样本，从而使创建虚拟人口成为可能。

最后，公众参与对进一步发展智慧规划解决方案至关重要。一方面，随着技术越来越普及，人们需要加强对自身数据价值的认识。另一方面，研究组织应继续改进安全和隐私保护机制，以维护数据挖掘生态系统。这种生态系统应该通过数据共享协议和参与感鼓励公众积极参与进来。作为回报，应开发更好的数据驱动应用程序以体现使用匿名数据的社会效益。我们希望用一种令人信服的方式解决这些问题，这对于数据驱动、基于个体的交通规划模型的开发和实际应用至关重要。

5 结论和研究成果

5.1 总结

引言部分对交通需求建模的最新进展进行了介绍。我们认识到基于出行的模型和基于个体的模型与记录人们移动的大数据源密切相关，因为这两者都直接源于个体出行模式的概念，而不是集计交通流的概念。为了充分利用基于个体的模型能力，不仅使用传统的数据输入(例如交通调查、人口普查)，还包括公共交通智能卡和手机数据随机收集的出行轨迹，这些数据记录了前所未有的规模和精细水平的交通行为。然而，为了识别出行活动和出行目的，必须进行额外的分析工作，以便将其整合到基于活动的交通需求框架中。

第一章对大数据源中提取出行行为所需要的方法论进行文献综述。从出行识别到活动推演，及文献在交通需求模型中的应用，对公共交通智能卡和手机数据逐步进行了述评。

最后，本文讨论了文献回顾的结果，并针对概率机器学习和交通模型明确了未来的挑战。

5.2 未来研究方向

本文将大数据与机器学习(例如概率图模型)相结合将成为继续发展交通模型的最大潜力，具体来说，是为了改进基于个体的交通规划模型。为此，未来的研究方向包括：

1）改进更具代表性的虚拟人口生成模型的设计。为此，需要确定给定的特定数据集，这些数据是最优的特征工程(feature-engineering)策略和随机变量之间的最佳关联。此外，创建虚拟人口的过程(社会经济方面)和分配活动计划的过程可以与更健康的生成模型设计联系起来。

2）从学习的角度看，由于不同的数据源能够解释城市出行现象的某些部分，最有前景和挑战性的方法将从迁移学习、多视图学习和半监督学习的模式中产生。

3）对于活动推理的具体工作，本文回顾了生成模型[41]和判别模型[28]。然后将两种模型结合起来，通过一组更丰富的特征集(判别模型)在活动推理中获得更好的结果，并从联合分布(生成模型)中取样。

4）通过寻找基站的信号特定模式和智能卡刷卡等额外数据源，重新审视交通方式推理。

5）基于个体仿真行为参数的超参数优化。例如，通过贝叶斯函数优化。

6）在基于个体的仿真选择模型中考虑社会效应。

7）进一步探索预处理阶段的概率滤波器。

8）针对面向大型数据驱动的基于个体

仿真的交通规划，探讨隐私指标的具体定义。

总之，我们认识到复杂的建模知识已经在交通规划领域发展起来，因此强烈建议在交通规划中应用数据驱动的方法时，需建立相应领域专业知识的基础。这些新的挑战需要交通模型专家和数据处理专家之间进行跨学科的合作。

注释：

Notes:

① 主要工作从2010年至2016年第二季度。

② 一些权威机构已经开始使用智能手机进行连续调查，以降低相应负担并提高数据质量，特别是在捕捉短时间活动方面。

致谢：

Acknowledgement:

感谢Seungjae Lee在首尔大学组织举办的2016年首尔大城市论坛，本文初稿发表于该会议。

公开声明：

Disclosure Statement:

本文作者不存在潜在的利益冲突。

基金：

Funding:

本研究成果隶属于由苏黎世ETH和新加坡国家研究基金会(FI370074016)联合成立的新加坡ETH中心未来城市实验室，得到“研究人才和科技企业”项目(Campus for Research Excellence and Technological Enterprise)的资助。

参考文献：

References:

[1] Castiglione J, Bradley M, Gliebe J. Activity-Based Travel Demand Models: Aprimer[R]. Washington DC: Transportation Research Board, 2015.

[2] de Dios Ortuzar J, Willumsen L G. Modelling Transport[M]. New Jersey: John Wiley & Sons, 2011.

[3] Rasouli S, Timmermans H. Activity-Based Models of Travel Demand: Promises, Progress and Prospects[J]. International Journal of Urban Sciences, 2014, 18(1): 31−60.

[4] Balmer M, Axhausen K, Nagel K. Agent-Based Demand-Modeling Framework for Large-Scale Microsimulations[J]. Transportation Research Record: Journal of the Transportation Research Board, 2006(1985): 125−134.

[5] Smith L, Beckman R J, Anson D, et al. TRANSIMS: Transportation Analysis and Simulation System[C]//The 5th TRB National Transportation Planning Methods Applications Conference, Seattle, April 17−21, 1995.

[6] Horni A, Nagel K, Axhausen K W. The Multi-Agent Transport Simulation MATSim[M]. London: Ubiquity Press, 2016.

[7] Adnan M, Pereira F C, Lima Azevedo C M, et al. SimMobility: A Multi-Scale Integrated Agent-Based Simulation Platform[C]//The 95th Annual Meeting of the Transportation Research Board, Walter E Washington Convention Center, Washington DC, January 10−14, 2016.

[8] Goulias K G, Bhat C R, Pendyala R M, et al. Simulator of Activities, Greenhouse Emissions, Networks, and Travel (SimAGENT) in Southern California[C]//The 91st Annual Meeting of the Transportation Research Board, Washington Marriott Wardman Park, Omni Shoreham, and Washington Hilton hotels, Washington DC, January 22−26, 2012.

[9] Calabrese F, Ferrari L, Blondel V D. Urban Sensing Using Mobile Phone Network Data: A Survey of Research[J]. ACM Computing Surveys, 2014, 47(2): 1−20.

[10] Barry J, Newhouser R, Rahbee A, et al. Origin and Destination Estimation in New York City with Automated Fare System Data[J]. Transportation Research Record: Journal of the Transportation Research Board, 2002 (1817): 183−187.

[11] Zhao J, Rahbee A, Wilson N H M. Estimating a Rail Passenger Trip Origin- Destination Matrix Using Automatic Data Collection Systems[J]. Computer-Aided Civil and Infrastructure Engineering, 2010, 22(5): 376−387.

[12] Trépanier M, Tranchant N, Chapleau R. Individual Trip Destination Estimation in a Transit Smart Card Automated Fare Collection System[J]. Journal of Intelligent Transportation Systems, 2007, 11(1): 1−14.

[13] Munizaga M A, Palma C. Estimation of a Disaggregate Multimodal Public Transport Origin− Destination Matrix from Passive Smartcard Data from Santiago, Chile[J]. Transportation Research Part C: Emerging Technologies, 2012, 24(9): 9−18.

[14] Yuan N J, Wang Y, Zhang F, et al. Reconstructing Individual Mobility from Smart Card Transactions: A Space Alignment Approach[C]//The 13th IEEE International Conference on Data Mining, Dallas, December 7−10, 2013.

[15] Seaborn C, Attanucci J, Wilson N. Analyzing Multimodal Public Transport Journeys in London with Smart Card Fare Payment Data [J]. Transportation Research Record: Journal of the Transportation Research Board, 2009 (2121): 55−62.

[16] Chakirov A, Erath A. Activity Identification and Primary Location Modelling Based on Smart Card Payment Data for Public Transport[C]//The 13th International Conference on Travel Behaviour Research, Toronto, July 15−20, 2012.

[17] Devillaine F, Munizaga M, Trépanier M. Activities of Public Transport Users by Analyzing Smart Card Data[J]. Transportation Research Record: Journal of the Transportation Research Board, 2012(2761): 48−55.

[18] Zou Qingru, Yao Xiangming, Zhao Peng, et al. Detecting Home Location and Trip Purposes for Cardholders by Mining Smart Card Transaction Data in Beijing Subway[J]. Transportation, 2018, 45(3): 919−944.

[19] Isaacman S, Becker R, Cáceres R, et al. Identifying Important Places in People's Lives from Cellular Network Data[C]//Lyons K, Hightower J, Huang E M. Pervasive Computing, No. 6696 in Lecture Notes in Computer Science. Berlin: Springer, 2011: 133−151.

[20] Han G, Sohn K. Activity Imputation for Trip-chains Elicited from Smart-Card Data Using a Continuous Hidden Markov Model[J]. Transportation Research Part B: Methodological, 2016, 83: 121−135.

[21] Bouman P. Recognizing Demand Patterns from Smart Card Data for Agent-Based Microsimulation of Public Transport[D]. Rotterdam: Erasmus University Rotterdam, 2012.

[22] Bouman P, Van der Hurk E, Kroon L, et al. Detecting Activity Patterns from Smart Card Data[C]//Delft University of Technology. BNAIC 2013: Proceedings of the 25th Benelux Conference on Artificial Intelligence. Delft: Delft University of Technology, 2013.

[23] Eagle N, Pentland A S. Eigenbehaviors: Identifying Structure in Routine[J]. Behavioral Ecology and Sociobiology, 2009, 63(7): 1057−1066.

[24] Fourie P J, Erath A L, Ordóñez Medina S A, et al. Using Smartcard Data for Agent-Based Transport Simulation[C]//Schmöcker JD, Kurauchi F. Public Transport Planning with Smart Card Data. Boca Raton: CRC Press, 2016: 133−160.

[25] Fourie P J. Reconstructing Bus Vehicle Trajectories from Transit Smart-Card Data[R/OL]. 2014[2016−06−01]. http://transp-or.epfl.ch/heart/2014/abstracts/257.pdf.

[26] Sarlas G, Axhausen K W. Localized Speed Prediction with the Use of Spatial Simultaneous Autoregressive Models[C]//The 94th Annual Meeting of the Transportation Research Board, Walter E Washington Convention Center, Washington DC, January 11−15, 2015.

[27] Sun Lijun, Tirachini A, Axhausen K W, et al. Models of Bus Boarding and Alighting Dynamics[J]. Transportation Research Part A: Policy and Practice, 2014, 69: 447−460.

[28] Widhalm P, Yang Y, Ulm M, et al. Discovering Urban Activity Patterns in Cell Phone Data[J]. Transportation, 2015, 42(4): 597−623.

[29] Caceres N, Wideberg J P, Benitez F G. Deriving Origin−Destination Data from a Mobile Phone Network[J]. Intelligent Transport Systems Iet, 2007, 1(1): 15−26.

[30] Iqbal M S, Choudhury C F, Wang P, et al. Development of Origin−Destination Matrices Using Mobile Phone Call Data[J]. Transportation Research Part C: Emerging Technologies, 2014, 40(1): 63−74.

[31] Wang P, Hunter T, Bayen A M, et al. Understanding Road Usage Patterns in Urban Areas [J]. Scientific Reports, 2012(2): 1−6.

[32] Jiang S, Ferreira Jr J, González M C. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore[C]//ACM SIGKDD 4th International Workshop on Urban Computing, Sydney, May 20, 2015.

[33] Horn C, Klampfl S, Cik M, et al. Detecting Outliers in Cell Phone Data[J]. Transportation Research Record: Journal of the Transportation Research Board, 2014, 2405: 49−56.

[34] Calabrese F, Lorenzo G D, Liu L, et al. Esti-

mating Origin-Destination Flows Using Mobile Phone Location Data[J]. IEEE Pervasive Computing, 2011, 10(4): 36−44.

[35] Ficek M, Kencl L. Inter-Call Mobility Model: A Spatio-Temporal Refinement of Call Data Records Using a Gaussian Mixture Model[J]. INFOCOM, 2012, 131(5): 469−477.

[36] Eagle N, Pentland A. Reality Mining: Sensing Complex Social Systems[J]. Personal and Ubiquitous Computing, 2006, 10(4): 255−268.

[37] Schlaich J, Otterstätter T, Friedrich M. Generating Trajectories from Mobile Phone Data [C]//The 89th Annual Meeting Transportation Research Board, Washington DC, January 10−14, 2010.

[38] Jiang Shan, Fiore G A, Yang Y, et al. A Review of Urban Computing for Mobile Phone Traces: Current Methods, Challenges and Opportunities[C]//The 2nd ACM SIGKDD International Workshop on Urban Computing, Chicago, 2013.

[39] Toole J L, Colak S, Sturt B, et al. The Path Most Traveled: Travel Demand Estimation Using Big Data Resources[J]. Transportation Research Part C: Emerging Technologies, 2015, 58: 162−177.

[40] Alexander L, Jiang S, Murga M, et al. Origin−Destination Trips by Purpose and Time of Day Inferred from Mobile Phone Data[J]. Transportation Research Part C: Emerging Technologies, 2015, 58: 240−250.

[41] Yin M, Sheehan M, Feygin S, et al. A Generative Model of Urban Activities from Cellular Data[J]. IEEE Transactions on Intelligent Transportation Systems, 2018, 19(6): 1682−1696.

[42] Yang Y, Widhalm P, Athavale S, et al. Mobility Sequence Extraction and Labeling Using Sparse Cell Phone Data[C]//The 30th AAAI Conference on Artificial Intelligence, Phoenix, February 12−17, 2016.

[43] Wang Huayong, Calabrese F, Lorenzo G D, et al. Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records[C]//The 13th International IEEE Conference on Intelligent Transportation Systems, 2010: 318−323.

[44] Zilske M, Nagel K. Studying the Accuracy of Demand Generation from Mobile Phone Trajectories with Synthetic Data[J]. Procedia Computer Science, 2014, 32: 802−807.

[45] Zilske M, Nagel K. A Simulation-Based Approach for Constructing All-Day Travel Chains from Mobile Phone Data[J]. Procedia Computer Science, 2015, 52: 468−475.

[46] Pozdnoukhov A, Campbell A, Feygin S, et al. San Francisco Bay Area: The Smartbay Project−Connected Mobility[C]//Horni A, Nagel K, Axhausen K W. The Multi-Agent Transport Simulation MATSim, London: Ubiquity Press, 2016: 485−490.

[47] Kaiser C, Pozdnoukhov A. Enabling Real-Time City Sensing with Kernel Stream Oracles and MapReduce[J]. Pervasive and Mobile Computing, 2013, 9(5): 708−721.

[48] Liao Lin, Fox D, Kautz H. Location-Based Activity Recognition Using Relational Markov Networks[J]. International Joint Conference on Artificial Intelligence, 2005, 26(2): 773−778.

[49] Liao Lin, Fox D, Kautz H. Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields[J]. International Journal of Robotics Research, 2007, 26(1): 119−134.

[50] Duong T V, Bui H H, Phung D Q, et al. Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model[J]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, 1(4): 838−845.

[51] Baratchi M, Meratnia N, Havinga P J M, et al. A Hierarchical Hidden Semi-Markov Model for Modeling Mobility Data[C]//2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Motif Seattle Hotel, Seattle, September 13−17, 2014.

[52] Liao Lin, Patterson D J, Fox D, et al. Learning and Inferring Transportation Routines[J]. National Conference of Artificial Intelligence, 2007, 171(5/6): 348−353.

[53] Yan Z, Chakraborty D, Parent C, et al. SeMiTri: A Framework for Semantic Annotation of Heterogeneous Trajectories[C]//The 14th International Conference on Extending Database Technology, Uppsala, Sweden, March 21−24, 2011.

(下转第 74 页)