# 公共交通中的大数据：来源和方法综述

**辛西娅·维迪塔，苏西洛·蒂莫西**

**所属单位：美国城市与区域规划学院**

**摘要：**大数据的收集作为传统资源密集型人工数据收集方法的替代方案，在过去十年中变得更加可行。此类数据的可用性，加上更复杂的预测统计技术，促使人们更加关注这些数据的应用，尤其是在交通分析方面。在交通文献中，人们越来越重视将通常收集的公共交通数据来源开发成更强大的分析工具。人们普遍认为，将大数据应用于交通问题将产生以前通过传统交通数据集无法获得的新见解。然而，关于大数据的构成、大数据收集和应用的伦理含义以及如何最好地利用新兴数据集，存在许多歧义。探索大数据的现有文献没有提供清晰一致的定义。虽然大数据的收集量不断增加，其在研究和实践中的应用也在不断扩大，但应用于此类数据的分析方法之间存在显着差异。本文总结了最近关于大数据来源的文献及其在解决公共交通问题时常用的方法。我们评估主要的大数据源、最常研究的主题和采用的方法。文献表明智能卡和自动化数据是研究人员最常用于进行公共交通分析的两个大数据源。审查的研究表明，大数据已在很大程度上用于了解公交用户的出行行为和评估公共交通服务质量。文献中报道的技术在很大程度上反映了那些用于较小数据集的技术。通常与大数据相关的更高级统计方法的应用仅限于少数研究。为了充分发挥大数据的价值，需要采用新的分析方法。

**关键字：**大数据、公共交通、交通分析、公交规划、规划方法、统计

## Big data in public transportation: a review of sources and methods

**Cynthia Widita, Susilo Timothy**

**(Affiliation: School of City and Regional Planning, USA)**

**Abstract：**The collection of big data, as an alternative to traditional resourceintensive manual data collection approaches, has become significantly more feasible over the past decade. The availability of such data, coupled with more sophisticated predictive statistical techniques, has contributed to an increase in attention towards the application of these data, particularly for transportation analysis. Within the transportation literature, there is a growing emphasis on developing sources of commonly collected public transportation data into more powerful analytical tools. A commonly held belief is that application of big data to transportation problems will yield new insights previously unattainable through traditional transportation data sets. However, there exist many ambiguities related to what constitutes big data, the ethical implications of big data collection and application, and how to best utilize the emerging data sets. The existing literature exploring big data provides no clear and consistent definition. While the collection of big data has grown and its application in both research and practice continues to expand, there is a significant disparity between methods of analysis applied to such data. This paper summarizes the recent literature on sources of big data and commonly applied methods used in its application to public transportation problems. We assess predominant big data sources, most frequently studied topics, and methodologies employed. The literature suggests smart card and automated data are the two big data sources most frequently used by researchers to conduct public transit analyses. The studies reviewed indicate that big data has largely been used to understand transit users' travel behavior and to assess public transit service quality. The techniques reported in the literature largely mirror those used with smaller data sets. The application of more advanced statistical methods, commonly associated with big data, has been limited to a small number of studies. In order to fully capture the value of big data, new approaches to analysis will be necessary.

**Keywords**：Big data, public transportation, transport analysis, transit planning, planning methods, statistics

## 引言

对人类时空运动的研究至少已经存在了 50 多年。由于需要预测未来的出行需求以更好地指导通常是大型交通项目的投资，交通研究人员长期以来一直在寻求开发模型来预测人们在时间和空间中的出行方式，并寻求了解影响出行相关的因素选择。最近，全球变暖和空气污染等重大挑战都可以追溯到对汽车的过度依赖，进一步激励交通研究人员和从业者制定有效的战略，以转向更可持续的交通方式（例如，公共交通和步行和骑自行车）。几十年来，交通研究人员主要使用主动征集的数据，例如，包括要求受试者通过纸质、网络或电话采访自我报告其活动和旅行的旅行调查。

各种形式的数据对于几乎所有研究领域的决策过程都很重要。数据的可用性决定了可以进行的研究类型，从而决定了可以从数据中收集的信息。准确的数据收集和适当的分析方法对于揭示交通系统面临的当前问题和未来挑战至关重要。几十年来，交通规划分析一直依赖于主要通过主动征集获得的人工收集数据，特别是为了了解交通用户的行为，例如家庭出行调查。这种特殊类型的数据通常很少收集（即每 5-20 年一次），往往是有意和有意地开发以进行交通规划、评估特定交通政策和解决相关研究目的。陈等人将此类通过主动征集获得的数据称为小数据。但是，此类数据的大小差异很大，可为交通界提供重要的见解。这与新兴的大数据概念形成鲜明对比，大数据通常是自动收集的，并且不太特定于应用程序。因此，我们将参考 Chen 等人的数据类型。在过去十年中，技术的快速发展导致数据的可用性急剧增加。有意收集的数据不再是唯一的形式：技术使被动收集数据成为可能，即大数据。大数据这个术语的出现是为了描述许多领域中可用的海量数据。在过去的十年里，这个词变得越来越普遍。作为一个包罗万象的术语，它“涉及实时收集、管理和分析大量增加的数据（现在通常是拍字节、艾字节和泽字节）的新方法和技术。随着研究人员致力于将这些海量数据整合到一系列研究中，人们对大数据的兴趣迅速增长。大数据主题的谷歌趋势显示自 2012 年初以来人们的兴趣显着增加。2017 年，大数据一词的搜索热度达到了最高，反映在垂直尺度上的值为 100 或接近 100。对现有交通相关大数据文献的考察表明，交通大数据应用主要集中在两个子领域：道路使用者行为和公共交通运营。在本文中，我们旨在通过强调用于交通分析的大数据的典型来源、对研究主题的范围进行分类以及讨论大数据在公共交通中应用的潜力和局限性来填补这一研究空白。

## 公共交通大数据

使用大数据的公共交通研究在 2013 年左右开始出现。这些研究中使用的数据来源主要是 GPS 点和轨迹、智能卡数据、自动乘客计数（APC）、自动票价控制（AFC）、自动车辆位置（AVL）、传感器数据、手机数据、网络数据和社交媒体数据。这种类型的数据每天可以产生大量数据，可用于研究从单个乘客的行为到大型公共交通系统的运作的任何事物。我们将本研究中检查的 81 篇论文分为六类：服务/绩效（20）、旅行需求（15）、旅行行为（26）、管理（12）、复原力和健康/安全（7）。

服务/性能

服务/绩效类别的研究涵盖了大数据的使用，可以帮助机构评估他们的服务并确定潜在的改进。 2015 年的一项研究使用马尔可夫链对多模式运输数据进行建模。该模型可用于改善公共交通系统并提高效率。 Strategway 是一个拟议的“公共交通路线规划网络解决方案组”，它使用大数据来识别城市居民的交通需求，并根据这些需求构建最佳路线网络，同时最大限度地降低拥有成本。 BusViz 是一种基于网络的应用程序，旨在帮助公交服务运营商和监管机构有效地使用大量现场数据来监控和可视化其公交车队的性能。另一种工具是基于 Impala（ESTRI）的高效时空数据检索方法，由研究人员开发，用于提高共享大量数据的效率。研究人员使用该工具检索大量时空轨迹数据，绘制太原公交分布图，并指出该映射在智能公共交通系统中用于交通调度、规划和行为管理等活动的有用性。

公交用户行为

了解乘客行为可以改进运输机构和其他决策者的决策。在公共交通中，大数据已被用于了解习惯行为、模式、极端出行行为以及影响这些行为的因素。 2017 年的一项研究使用智能卡数据检查了公共汽车乘客的个人和总体旅行行为。表现出习惯性行为的个人和公交路线是通过使用粘性指数来识别的，高粘性归因于总是在同一条路线上旅行的个人，而低粘性归因于那些以更多样化的方式旅行的人。作者指出了这些信息在设计和安排公共交通系统时的有用性。一项研究利用加拿大加蒂诺 5 年多的智能卡数据，使用具有相似旅行时间的乘客的聚类方法来评估乘客行为如何随时间变化，以确定长期旅行模式。

一项关于伊斯坦布尔 BRT 线路的研究进行了分析，以评估通勤者的时空出行行为。另一项关于 BRT 的研究使用智能卡数据库来研究 BRT 和其他公交车的时空模式，并确定 BRT 出行是否具有与其他公交车出行不同的空间和时间模式。一项关于中国高铁的研究使用从铁路票务网站自动收集的大数据来研究这些铁路沿线旅行行为的时空模式。 2016 年一项研究的重点是那些 (1) 比平均时间早，(2) 比平均时间晚，(3) 距离比平均距离更长，以及 (4) 每天比平均时间多的人将智能卡数据与传统家庭调查结合使用。一项类似的研究提出了一系列针对智能卡数据的数据挖掘方法，并用它们来研究北京的时空通勤模式。由于研究人员通过人工收集的交通智能卡持有人调查验证了挖掘的交通智能卡数据，因此所提出的数据挖掘方法的通勤者识别准确率高达 94.1%。

出行需求

与出行行为密切相关的是出行需求的计算。此类研究主要集中在使用大数据技术计算起点-终点 (OD) 矩阵。2015 年的一项研究提出了一种计算机软件系统，该系统利用大数据执行分析、创建起点-目的地矩阵，并使用呼叫详细记录 (CDR) 以及开放和众包的地理空间数据、调查和人口普查记录。 2017 年的一项研究使用分布式计算技术来分析门票销售和公交车位置数据，以估计需求和起点-目的地矩阵。另一项研究创建并展示了 QZTool，这是一个通过使用浮动电话数据自动生成起点-终点矩阵的系统。通过将 QZTool 创建的矩阵与从交通需求模型创建的矩阵进行比较来评估该工具，发现这些矩阵“高度一致”。2016 年的一项研究创建了公共交通的微观模拟模型，并使用智能卡和通用交通提要规范(GTFS) 数据进行动态行程分配。本研究中的程序旨在提高公共交通规划的准确性。

管理

管理类别的研究提出了机构维护其基础设施、管理大量数据或来自不同来源的数据以及通过大数据分析支持其整体系统和决策的方法。开发了一种基于多代理的模拟，以使用智能卡数据作为中国北京的主要输入来提供有关站点重要性程度的信息，相关机构可以使用这些信息来分配资源。 Maktoubian 提出了一个基于流式大数据分析 (SBDA) 的状态维护 (CBM) 新平台。该平台中使用的数据可以来自各种来源，例如传感器、图像、技术文档和网页。2016 年的一项研究提出了一个三层管理系统，并在巴西福塔莱萨进行了测试。该系统使用大数据向交通机构提供有用信息，包括计算的旅行时间和需求、网络分析和决策支持。处理来自不同来源（例如传感器、GPS 信号、视频、软件等）的大量数据可能会阻碍机构有效地使用其数据。Zhang 创建了一种数据处理方法来处理来自不同来源的大量数据，重点是中国的高速列车控制系统。同样与中国的高铁 (HSR) 相关，通过从高铁售票系统中挖掘开放大数据，研究人员能够进行一系列空间分析，并说明高铁系统的全国空间配置。

另一项在德国汉堡使用公交车的研究开发了一个大数据动态驾驶周期，可供城市地区使用以确定其公交车的驾驶周期。 Nuzzollo 和 Comi 讨论了交通运营控制和旅客信息工具的开发，并提出了两种工具来预测车上占用率并为旅客提供个性化的出行前和途中信息。同样，针对智利圣地亚哥的公共交通系统 Transantiago 讨论了从交通数据中收集信息的方法，以及服务质量、速度概况和时间使用模式的活动方法。大数据也被证明是一个具有巨大潜力的领域，可以为发展中国家的过境决策提供信息，特别是基于网络和位置的数据。

弹性和健康/安全

一些研究考察了公共交通系统的弹性以及这些系统对公共健康/安全主题的影响等主题。使用爬行工具收集

每日通过步行、公共交通、私家车和自行车从社区到健康食品商店的旅行时间，检查交通变化区域的健康食品获取情况。沿着某种类似的思路，使用从中国开封的应用程序编程接口（API）派生的大数据估算了医疗设施的空间可达性。另一项研究考察了 WMATA 使用大数据评估火车站拥挤情况和所需的安全标准。匹兹堡的一项研究从众包数据中提取了交通系统的感知安全性。结合来自智能卡的犯罪记录数据和乘客数据，作者说明了潜在的不安全交通换乘，可以通知交通机构定位交通站点。

## 结论

随着越来越多的机构和研究人员看到新见解的潜力，大数据在公共交通分析中的使用正在增加。然而，到目前为止，此类数据的实用性受到最初为有意收集的数据采用的有限技术和方法的应用而受到限制。在本研究审查的 81 篇论文中，只有八篇（不到 10%）使用了机器学习技术等先进方法。其余文献应用了为较小数据集开发的方法。虽然这种方法在技术上是合理的，但它可能会限制从新兴大数据中获得的洞察力。

我们还注意到所审查的研究中常见的一些局限性和挑战。首先，许多研究人员建议，需要使用传统的旅行调查来补充或验证从新兴大数据中得出的分析。这种观点强调了这两种形式的数据的互补性概念，可以得出可靠的信息和基于证据的推论，最终有利于交通机构、交通用户和公众。其次，虽然大数据被用作分析的主要来源，但研究人员仍然在很大程度上依赖有意收集的数据来得出感兴趣的探索性变量。例如，社会人口统计和建筑环境措施是从主动征集数据形式发展而来的。第三，虽然大数据可以提供大量信息，但所提供的信息通常只包含所讨论的单一交通系统的单一模式。鉴于公共交通中的多式联运概念，其中用户的决策可能会受到服务区域中多种模式的可用性的影响，开发可扩展的模型可以捕获不同的兴趣系统以及对用户行为的相关影响是至关重要，应作为未来的研究途径加以追求。第四，审查的研究还强调需要将分析扩展到长期，最好是数年。

## 参考文献

1. Briand, A.-S., Côme, E., Trépanier, M., & Oukhellou, L. (2017). Analyzing year-to-year changes in public transport passenger behaviour using smart card data. Transportation Research Part C: Emerging Technologies, 79, 274–289. doi:10.1016/j.trc.2017.03.021

2. Cai, H., & Xu, M. (2013). Greenhouse gas implications of fleet electrification based on big datainformed individual travel patterns. Environmental Science & Technology, 47(16), 9035–9043. doi:10.1021/es401008f

3. Faroqi, H., Mesbah, M., Kim, J., & Tavassoli, A. (2018). A model for measuring activity similarity between public transit passengers using smart card data. Travel Behaviour and Society, 13, 11– 25. doi:10.1016/j.tbs.2018.05.004

4. Ferreira, J. C., Monteiro, V., Afonso, J. A., & Afonso, J. L. (2016). Methodology for knowledge extraction from mobility big data. In Advances in intelligent systems and computing. Distributed computing and artificial intelligence, 13th international conference (pp. 97–105). doi:10.1007/978-3-319-40162- 1_11

5. Günther, R., Wenzel, T., Wegner, M., & Rettig, R. (2017). Big data driven dynamic driving cycle development for busses in urban public transportation. Transportation Research Part D: Transport and Environment, 51, 276–289. doi:10.1016/j.trd.2017.01.009

6. Hanft, J., Iyer, S., Levine, B., & Reddy, A. (2016). Transforming bus service planning using integrated electronic data sources at NYC transit. Journal of Public Transportation, 19(2), 89–108. doi:10.5038/ 2375-0901.19.2.6

7. He, L., Agard, B., & Trépanier, M. (2018). A classification of public transit users with smart card data based on time series distance metrics and a hierarchical clustering method. Transportmetrica A: Transport Science, 1–20. doi:10.1080/23249935.2018.1479722

8. Kumar, P., Khani, A., & He, Q. (2018). A robust method for estimating transit passenger trajectories using automated data. Transportation Research Part C: Emerging Technologies, 95, 731–747. doi:10.1016/j.trc.2018.08.006

9. Lantz, K., Khan, S., Ngo, L. B., Chowdhury, M., Donaher, S., & Apon, A. (2015). Potentials of online media and location-based big data for urban transit networks in developing countries. Transportation Research Record: Journal of the Transportation Research Board, 2537, 52–61. doi:10.3141/2537-06

10. Li, R., Kido, A., & Wang, S. (2015). Evaluation index development for intelligent transportation system in smart community based on big data. Advances in Mechanical Engineering, 7(2), 541651. doi:10. 1155/2014/541651