基于智能卡数据的乘客路径选择估计：一种考虑跨日面板效应的潜在类别方法

Baichuan Mo[a]、ZhenLiang Ma[b,*]、Haris N. Koutsopoulos[c]、Jinhua Zhao[d]

[a]麻省理工学院土木与环境工程系，马萨诸塞州剑桥市，*MA02139*[b]瑞典皇家理工学院土木与建筑工程系，斯德哥尔摩，*10044*[c]东北大学土木与环境工程系，波士顿，*MA 02115*[d]麻省理工学院城市研究与规划系，马萨诸塞州剑桥市，*MA20139*

arXiv:2301.03808v1 [stat.AP] 10 Jan 2023

摘要

了解乘客在城市轨道交通系统中的路径选择行为，是有效运营与规划的先决条件。该领域在两个广泛但相互独立的方向上取得了积极进展：交通领域利用“小规模”调查数据进行行为建模，以及计算机科学领域利用“大数据”研究交通与出行模式。本文旨在弥合这两个方向之间的差距，提出一种概率方法，利用大规模智能卡数据推断乘客在城市轨道交通系统中的路径选择行为。该模型利用潜在类别和面板效应，捕捉乘客隐含的行为异质性与纵向相关性——这是大数据驱动的行为研究中的关键研究空白。我们根据乘客的路径选择行为，构建个体到达目的地的概率分布，并将相应路径选择模型参数的估计转化为最大似然估计问题。由于计算复杂度呈指数级增长，原始似然函数难以处理。我们推导出一个可处理的似然函数，并提出一种数值积分方法，以高效估计模型。此外，我们还提出一种方法，根据数值估计得到的海森矩阵和克拉美–罗界（系数方差的下界），计算估计所得选择参数的 t 统计量。基于合成数据的案例研究验证了模型性能及其对参数初始化和输入误差的稳健性，并突显了在路径选择估计中纳入拥挤影响的重要性。对香港港铁实际数据的应用发现了两类潜在乘客群体：时间敏感型（TS）和舒适导向型（CA）。时间敏感型乘客更倾向于选择出行时间较短的路径。他们大多是出行频率高、时间安排灵活性较低的固定通勤者。舒适导向型乘客更关注出行舒适体验，会选择步行与候车时间较短的路径。所提出的方法由数据驱动，且具有通用性，能够适用于其他离散选择结构。它可提供与传统选择建模相同的输出，并以经济、及时的方式深入理解乘客的选择行为，据此可制定、评估和监测更具依据的规划与管理策略。

关键词：路径选择；城市轨道系统；智能卡数据；潜在乘客群体，面板效应

1. 引言

许多大型城市轨道交通系统的客运量增速已超过运能增长，包括香港港铁（MTR）、伦敦地铁和纽约地铁系统（Zhu

*通讯作者

等人，2017年b，a）。车站和列车上的拥挤问题令人担忧，因为它会影响安全、服务质量和运营效率。多项研究测算了乘客为减少拥挤而愿意支付的费用（Li和Hensher，2011），并建议在投资评估中纳入拥挤负效用（Haywood和Koning，2015）。鉴于人们日益关注如何应对拥挤相关问题，了解拥挤情况下乘客的路径选择行为，对运营管理和规划实践都很重要。然而，估计路径选择比例或个体选择行为并非易事。由于无法直接观测城市轨道交通系统中乘客的路径选择，既有研究大多依赖显示偏好和陈述偏好调查数据（Raveau等人，2011，2014；Jin等人，2017；Zhang等人，2017）。调查是开展行为分析的有力工具，但受制于高昂的成本、报告准确性和调查覆盖范围。

自动收费系统（AFC）和自动车辆定位系统（AVL）数据为出行行为、需求建模、公共交通运营规划等领域的分析提供了机会（Pelletier 等，2011；Bagchi 和 White，2005；Koutsopoulos 等，2019）。除了乘客出行时间和地点的汇总趋势外，自动收费系统数据还提供了有关个体和／或特定群体出行模式的详细信息（Goulet-Langlois 等，2016；Briand 等，2017）。表 1 总结了地铁系统中利用自动收费系统数据和／或自动车辆定位数据开展的现有路径选择研究。已有多项研究利用自动收费系统数据估计乘客的路径选择概率（Sun 和 Xu，2012；Zhao 等，2016；Sun 和 Schonfeld，2016；Zhou 等，2015；Zhu 等（2021）；Mo 等，2022）。这些研究为现有条件下的汇总选择行为（即各路径所占比例）提供了有益见解。例如，Zhu 等（2021）提出了一种数据驱动方法，用于推断城市重型轨道交通系统中的乘客行程路线，并可利用自动收费系统数据和自动车辆定位数据估计各路径所占比例。然而，若不对个体路径选择行为进行建模，这些结果就无法用于运营规划应用，例如时刻表设计、网络扩建、运营策略和政策干预等。这是因为新的时刻表或网络扩建可能改变服务属性，进而导致个体选择行为发生变化。推断各路径所占比例无法捕捉这些变化的影响。

本研究重点在于利用自动收费系统和自动车辆定位数据，估计以备选路径属性为函数的路径选择模型。与此相关，Sun 等人(2015年提出了一种综合贝叶斯方法，利用新加坡大众捷运系统的自动收费系统数据，推断网络属性和乘客路径选择行为。Zhang 等人(2018年通过结合显示偏好（RP）调查数据和自动收费系统数据，构建了数据融合模型来估计个体的路径选择，并对乘客的风险态度进行建模。然而，这两项研究都对路段旅行时间作出了强假设（服从独立正态分布），忽略了在拥堵状况下，乘客可能因运能约束而滞留在主要车站这一事实。在高峰时段，乘客可能会在（通常）出行时间较短的路径上滞留。但上述模型无法区分较长的出行时间究竟是由于选择了较长的路径，还是由于在较短路径上多次滞留所致（Zhu 等人，2017a，2021；Mo 等人，2020b）。

为纳入滞留现象，Zhu 等人(2017年b) 提出了乘客列车分配模型（PTAM），将出行时间分解为接入、等待、车内、出站步行时间，并明确考虑乘客在起始站滞留的动态过程。该模型用于估计容量约束下非换乘行程在关键车站的滞留情况，并使用合成数据和实际数据进行了验证。Hörcher 等人(2017年) 将 PTAM 扩展到包含换乘的情形，并提出了一种离散选择模型（DCM），用于估计城市轨道交通系统中的拥挤用户成本。然而，该模型依据预先设定的概率阈值识别乘客实际采用的“路径”，这

表1：使用自动收费系统数据进行城市轨道交通路径选择推断与建模的文献综述

参考资料	Data		行为		特征			方法	
	AFC	AVL	汇总数据	个体数据	拥挤	异质性	面板效应	优化	概率
Sun 和 Xu (2012)	✓	✓	✓		✓				✓
Zhao 等人 (2016)	✓	✓	✓		✓				✓
Sun 和 Schonfeld (2016)	✓	✓	✓		✓				✓
Zhou 等人 (2015)	✓	✓	✓		✓				✓
Zhu 等人(2021)	✓	✓	✓		✓				✓
Mo 等人 (2022)	✓	✓	✓		✓			✓	
Sun 等(2015年)	✓			✓					✓
Zhang 等(2018)	✓			✓					✓
Hörcher 等(2017年)	✓	✓		✓	✓				✓
本研究	✓	✓		✓	✓	✓	✓		✓

最终可能会影响选择模型的估计质量。

这些数据驱动的行为研究为弥合不同领域中“小样本”（调查）与“大数据”（自动收费系统）研究之间的差距提供了良好的尝试，尽管它们探讨的是同一个问题：拥挤情况下乘客的路径选择（Chen 等，2016年）。然而，现有基于 AFC/AVL 数据的路径选择估计研究基本上都是为了在离散选择模型研究中校准标准离散选择模型的参数而设计的。与基于调查数据的离散选择模型研究相比，它们未能系统地考虑行为选择问题本身的共性与独特特征，也未能在大数据背景下相应地对这些特征进行建模，具体包括：

- **选择异质性**。众所周知，乘客对服务表现（即出行时间、等待时间）可能有不同的感知，并会采取不同的出行选择策略。忽略群体异质性可能导致估计偏差。在离散选择模型文献中，选择异质性通常通过潜在类别或混合模型来捕捉（Hess 等，2009；Mo 等，2021）。然而，据作者所知，此前没有基于自动收费系统数据的研究在个体路径选择行为建模中考虑乘客异质性。

- **面板效应（跨时间的选择相关性）**。自动收费系统数据记录了乘客多日的出行，因此面临的一项独特挑战是：应使用哪些日期的数据来估计乘客的常规行为。考量个体在多个日期的出行情况，对于稳健地估计选择行为至关重要，因为乘客有时可能会在某些日期偏离其惯常行为。然而，这也带来了如何对个体在不同时间和日期的路径选择进行建模以反映其时间相关性（即面板效应）的挑战，而以往研究并未考虑甚至讨论这一问题。

- **选择模型系数显著性（t 统计量）**。在离散选择模型研究中，模型系数的显著性水平对于得出行为与政策见解十分重要。然而，目前尚未发现有基于自动收费系统/自动车辆定位数据的研究报告经校准模型系数的 t 统计量，因此限制了此类研究开展全面行为解读的能力。

为填补研究空白，本文提出一种带有面板效应的潜在类别方法，利用自动收费系统（进出站刷卡）和自动车辆定位数据估计个体路径选择行为。该框架明确考虑了容量约束（拥挤）、选择异质性和面板效应。我们根据乘客的路径选择行为，对其到达目的地车站的到达时间概率进行建模，并将相应路径选择参数的估计表述为最大似然估计（MLE）问题。

原始似然函数难以处理，因为其中包含数量呈指数级增长的大量求和项，以及对一个服从正态分布的随机变量进行积分。我们推导出一种新的基于条件概率的公式，以消除大量求和项，并采用数值积分方法处理该正态随机变量，从而得到一个易于处理的似然函数，并实现高效的模型估计。由于解析海森矩阵难以推导，估计参数的 t 值根据数值估计的海森矩阵和克拉美–罗界计算。基于合成数据的案例研究验证了模型性能，并凸显了在路径选择估计中纳入拥挤影响的重要性。利用来自香港港铁（MTR）的实际数据进行的应用分析，揭示了该系统中的两个乘客潜在群体。本研究的主要贡献如下：

- 利用智能卡数据，对封闭式公共交通系统中的乘客路径选择问题进行建模，并考虑系统拥挤、选择异质性以及不同时段的面板效应。
- 构建基于最大似然估计的潜类别路径选择估计问题，纳入面板效应，并推导出便于计算的似然函数，以高效估计模型系数。
- 提出一种数值方法，基于数值估计的海森矩阵和克拉美–罗界（系数方差的下界），计算估计所得选择系数的 t 统计量。
- 利用合成数据和香港的真实世界数据验证模型性能，并识别出对出行时间和舒适度具有异质性偏好的乘客潜在群体。

本文其余部分安排如下：第 2 节阐述路线选择问题并提出最大似然估计方法。第 3 节利用合成数据和真实世界数据验证所提出的方法。最后一节总结主要发现并探讨未来研究方向。

2. 方法论

2.1. 问题描述

考虑一个封闭的自动收费系统，该系统可获取乘客随时间产生的进站刷卡和出站刷卡记录，同时也可从自动车辆定位系统获取车站的列车到达和列车出发信息。定义乘客i 具有一系列观测到的自动收费系统记录 $v_i = \{(o_i,1, d_i,1, t_i^{\text{in}},1, t_{i\,1}^{\text{out}},) ..., (o_{i,N_i}, d_{i,N_i}, t_i^{\text{in}},_{N_i}, t_i^{\text{out}},_{N_i})\}$，其中$o_{i,n}$,$d_{i,n}$,$t_i^{\text{in}}$,$_n$, t_i^{out},$_n$ 分别表示第 n次行程的出发地、目的地、进站刷卡时间和出站刷卡时间。所有乘客的集合定义为 $\mathcal{P}$ （即 $i \in \mathcal{P}$）。

为刻画乘客的选择异质性，我们假设总体中存在 K 个潜在群体，且同一群体中的乘客具有相同的选择偏好。令 g_i 为一个随机变量，用于表示乘客 i 所属的群体。乘客 i 属于潜在群体 G_k 的概率由以下多项逻辑回归模型表示：

$$\Pr(g_i = G_k; \theta) = \frac{\exp(\theta_k \cdot x_i)}{\sum_{k'=1}^{K} \exp(\theta_{k'} \cdot x_i)} \tag{1}$$

其中， x_i 是乘客 i的特征向量，包括从智能卡数据中提取的变量，如出行频率、卡类型、出行规律性等。 $\mathcal{G} = \{G_k \mid k = 1, 2, ..., K\}$ 是待估计的潜在群体集合（K 需要预先指定）。 $\theta = (\theta_k)_{k=1,...,K}$ 是待估计的参数向量，与个体特征相关。

根据随机效用最大化（RUM）假设（Ben-Akiva 和 Lerman，2018），乘客 i 在第 m 次行程中选择路径 n，且乘客 i 属于第 G_k组时，其效用可表示为：

$$U_{i,n,m}^{k} = \beta^{k} \cdot z_{n,m} + \alpha_{i}^{k} + \varepsilon_{i,n,m}^{k} \tag{2}$$

其中 β^k 是待估计的未知参数，与路径属性相关联。$z_{n,\ m}$：$= [y_{n,m}$, 对数 $PS_m]$和$y_{n,\ m}$ 是路径属性向量，包括车内时间、车外时间、滞留等待时间等变量。 PS_m 是“路径规模”因子，用于捕捉路径重叠导致的误差项相关性（Hoogendoorn-Lanser 和 Bovy，2007年）。包含路径规模因子的公式称为“路径规模Logit模型”（Prato，2009）。 PS_m 定义如下

$$PS_{m} = \frac{1}{L_k} \sum_{a \in A_m} \frac{l_a}{\sum_{m' \in \mathcal{R}_{i,n}} \delta_{a,m'}} \quad \forall\, m \in \mathcal{R}_{i,n}, n = 1, ..., N_i, i \in \mathcal{P} \tag{3}$$

其中 A_m 是路径 m的所有链路的集合。$\delta_{a,\ m'} = 1$ 如果链路 a 属于路径 m'，否则$\delta_{a,\ m'} = 0$。 L_k 是路径 m 的长度，l_a 是链路 a的长度。$\mathcal{R}_{i,\ n}$ 是乘客 i第 n次行程的所有可用路径的集合。

为捕捉个体行为随时间变化的相关性（即面板效应），效用函数（公式 2）还包含一个个体特定不可观测因子 α_i^k（一个随机变量）。假设面板效应随时间保持不变（即没有下标 n）（Ben-Akiva 和 Lerman，2018）。假设对于组 k 中的所有乘客，α_i^k均独立同分布（i.i.d.），且服从均值为零、方差为 $\mathcal{N}(0,\ (\sigma^k)^2)$ 的正态分布（均值为零是因为均值可作为备选项特定常数的一部分进行估计），其中 σ^k 是待估计的标准差。给定 α_i^k，假设不可观测误差项 $\varepsilon_{i,\ n,\ m}^k$ 在所有 i、n和 m之间独立同分布且服从Gumbel分布。

设 $\pi_{i,\ n,\ m}^k[\alpha_i^k]$ 分别表示乘客 i 在第 n次行程中选择路径 m 的概率，给定乘客 i属于组 G_k。根据随机效用最大化理论，

$$\pi_{i,n,m}^{k}[\alpha_i^k] = \frac{\exp(\beta^k \cdot z_{n,m} + \alpha_i^k)}{\sum_{m' \in \mathcal{R}_{i,n}} \exp(\beta^k \cdot z_{n,m'} + \alpha_i^k)} \tag{4}$$

由于乘客 N_i 共有 i 条行程记录，因此我们可以将序列选择概率表示为（Arellano 和 Honoré，2001）：

$$\Pr(r_{i,1} = m_1, ..., r_{i,N_i} = m_{N_i}) = \sum_{k=1}^{K} \Pr(g_i = G_k) \cdot \int_{\alpha_i^k} \left[\prod_{n=1}^{N_i} \pi_{i,n,m_n}^{k}[\alpha_i^k] \right] \cdot f(\alpha_i^k)\, \mathrm{d}\alpha_i^k$$
$$\forall m_1 \in \mathcal{R}_{i,1}, ..., m_{N_i} \in \mathcal{R}_{i,N_i} \tag{5}$$

其中，$r_{i,\ n}$ 是一个随机变量，表示乘客 i 在第 n 次行程中使用的路径。 $f(\alpha_i^k)$ 是 α_i^k的概率密度函数。

本研究旨在开发一种方法，同时估计 $\beta = (\beta^k)_{k=1,...,K}$、 θ、$\sigma = (\sigma^k)_{k=1,...,K}$，这些参数分别描述乘客的路径选择行为、选择异质性和面板效应。我们将在以下章节中构建最大似然估计问题，以估计这些参数。

方法的结构如图1所示。

本文使用的符号见表2。

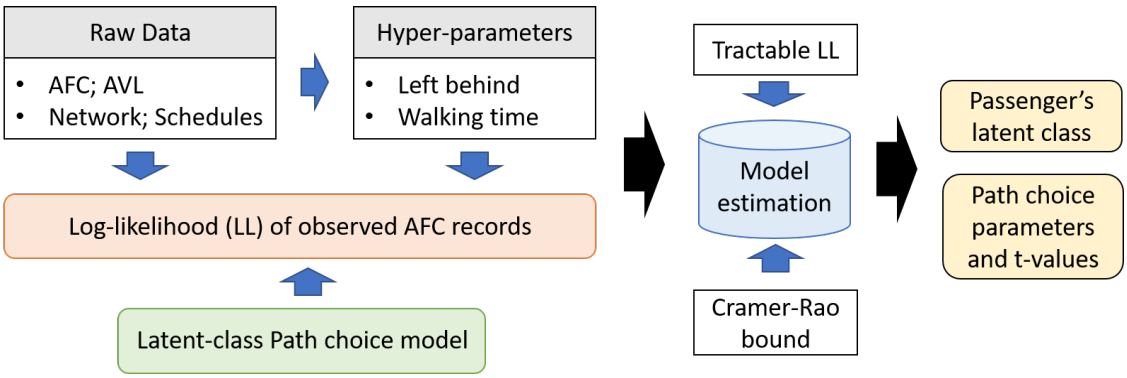


图1：方法论框架

表2：符号汇总

符号	描述
模型参数	
v_i	一系列乘客自动收费系统数据记录 i
$(o_{i,n}, d_{i,n}, t^{\text{in}}_{i,n}\ t^{\text{out}}_{i,n})$	乘客第 n 次行程的出发地、目的地、进站刷卡时间和出站刷卡时间，分别为分别
$\mathcal{P}$	所有乘客的集合
x_i	乘客 i的特征向量
N_i	乘客 i的出行次数总计
G_k	第 k 个潜在群体
$\mathcal{G}$	所有潜在群体的集合
K	潜在群体的数量
$z_{n,m}$	路径 m 和行程 n的路径属性向量
α_i^k	用于捕捉潜在群体 k中个体数据 i 的面板效应的随机变量
$U^k_{i,n,m}$	在乘客 i 属于某个潜在群体的情况下，乘客 i 在第 n次行程中选择路径 m 的效用组 G_k
PS_m	路径 m的路径规模因子
A_m	路径 m的所有链路集合
$\mathcal{R}_{i,n}$	乘客 i第 n 次行程的所有可用路径集合
τ	每个时间索引所代表的时间长度
$\mathcal{R}^{u,v}$	OD 对（u， v）的可行路径集合
$\pi^k_{i,n,m}[\alpha_i^k]$	乘客 i 在第 n次行程中选择路径 m 的概率，条件是该乘客 i 属于组 G_k
$r_{i,n}$	表示乘客 i 在第 n次行程中所使用路径的随机变量
$\mathcal{LL}(\theta,\ \beta,\ \sigma)$	所有观测值的对数似然函数
$f(\alpha_i^k)$	α_i^k的概率密度函数
$\Lambda^j_{i,n,m}$	与乘客 i的行程 n中路径 m 的第 j个路径区段相关的所有列车集合
$J_{i,n,m}$	乘客 i的第 n次行程中路径 m 的路径区段数
$\Omega_{i,n,m}$	乘客 i的第 n次行程中路径 m 的所有可能行程路线集合

$T_d(\cdot)$	一个返回列车在上车（或下车）对应路段车站的出发时间的函数
$T_a(\cdot)$	一个返回列车在上车（或下车）车站的到达时间的函数对应路段
$f_m^{\text{Eg}}(\cdot)$	路径 m的出站步行时间概率密度函数（PDF）
$f_m^{\text{Ac}}(\cdot)$	路径 m的进站步行时间概率密度函数
$B_{i,n,m}^{I_1}$	乘客 i 在行程 n中被滞留、无法搭乘列车 I_1 的最大次数对于路径 m
$\eta_{i,n,m}^{j,k}$	在第 j个路段的上车站被滞留 k 次的概率
$\mathcal{I}^{u,v,r}$	乘客 i第 n 次行程的路径 m OD 对 (u,v) 的路径 r 的区段集合
$E^{i,n,I_j,k}$	“乘客 i 在第 n 次行程中到达路段 j 的上车站”这一事件在列车 I_j-k 发车与 I_j-k-1 之间，路径 m 的乘客滞留 k 乘坐列车 I_j的次数”
$t_{i,n,m}^{j}$	从列车 I_{j-1} 下车处到下一个站台的换乘步行时间，适用于乘客 i使用路径 m的第 n次行程
$\hat{H}_{k,k}^{-1}$	对数似然函数的逆海森矩阵的第k个对角元素

待估参数	
β^k	与潜在组 G_k中的路径属性相关的参数
σ^k	α^k的标准差
θ^k	潜在组 G_k中与个体特征相关的参数向量

2.2. 模型构建

给定以下集合 乘客 i 的AFC数据记录 v_i，观测到 v_i 的概率可表示为

$$\begin{aligned}\Pr(v_i) &= \sum_{r_{i,1},\ldots,r_{i,N_i}} \Pr(v_i \mid r_{i,1},\ldots,r_{i,N_i})\Pr(r_{i,1},\ldots,r_{i,N_i}) \\ &= \sum_{r_{i,1},\ldots,r_{i,N_i}} \left\{ \left[\prod_{n=1}^{N_i} \Pr(o_{i,n}, d_{i,n}, t_{i,n}^{\text{in}}, t_{i,n}^{\text{out}} \mid r_{i,n}) \right] \times \Pr(r_{i,1},\ldots,r_{i,N_i}) \right\} \end{aligned} \tag{6}$$

其中，第二个等式根据贝叶斯定理得出。

由于给定路径 m 时出发地和目的地是已知的，因此 $\Pr(o_{i,n},d_{i,n},t_{i,n}^{\text{in}},t_{i,n}^{\text{out}} \mid r_{i,n}=m)$ 等价于乘客 i 在时间 $t_{i,n}^{\text{in}}$ 进入出发地并在时间$t_{i,n}^{\text{out}}$ 离开目的地的概率（记为 $\Pr(t_{i,n}^{\text{in}},t_{i,n}^{\text{out}} \mid r_{i,n}=m)$ 。请注意，

$$\Pr(t_{i,n}^{\text{in}}, t_{i,n}^{\text{out}} \mid r_{i,n}=m) = \Pr(t_{i,n}^{\text{out}} \mid t_{i,n}^{\text{in}}, r_{i,n}=m)\cdot \Pr(t_{i,n}^{\text{in}} \mid r_{i,n}) \propto \Pr(t_{i,n}^{\text{out}} \mid t_{i,n}^{\text{in}}, r_{i,n}=m) \tag{7}$$

“正比于”是因为我们没有对进站刷卡时间选择进行建模。因此，似然函数变为

$$\mathcal{L}(\theta,\beta,\sigma) = \prod_{i\in\mathcal{P}} \Pr(v_i) = \prod_{i\in\mathcal{P}} \left[\sum_{r_{i,1},\ldots,r_{i,N_i}} \prod_{n=1}^{N_i} \left[\Pr(t_{i,n}^{\text{out}} \mid t_{i,n}^{\text{in}}, r_{i,n})\right] \cdot \Pr(r_{i,1},\ldots,r_{i,N_i}) \right] \tag{8}$$

似然函数（方程 8）中唯一未知的部分是 $\Pr(t_{i,n}^{\text{out}} \mid t_{i,n}^{\text{in}}, r_{i,n}=m)$，它表示乘客 i 在时间 $t_{i,n}^{\text{out}}$ 在使用路径 $m \in \mathcal{R}_{i,n}$ 并于时间 $t_{i,n}^{\text{in}}$ 进站刷卡的条件下，于其目的地出站刷卡的概率。该概率可通过对不同的行程情景进行积分得到，每种情景都对应特定的步行、上车和滞留可能性（Zhu 等人，2021）。

为说明 $\Pr(t_{i,n}^{\text{out}} \mid t_{i,n}^{\text{in}}, r_{i,n}=m)$ 的推导，我们考虑一个涉及一次换乘的行程。图 2 以时空图的形式展示了乘客从1号线起点站刷卡进站，并在2号线目的地车站刷卡出站的所有可能移动方式。沿特定线路的移动称为“路径区段”。路径区段由线路、上车站和换乘站/下车站确定。每个路径区段都对应一组列车，其运行编号表示发车时间顺序。例如，图 2 中的第一个路径区段包含按时间顺序编号的列车1、2、3和4。设与乘客 i 的行程 n 中路径 m 的第 j 个路段相关联的所有列车集合为 $\Lambda_{i,n,m}^{j}$。例如，对于图 1 中的第一个路径区段，我们有 $\Lambda_{i,n,m}^{j}$ = {1号线1号列车、1号线2号列车、1号线3号列车、1号线 4} 号列车。略微滥用一下记号，对于列车 $I \in \Lambda_{i,n,m}^{j}$，列车 $I+k$ 表示同一线路上编号为 $+k$ ($k \in \mathbb{Z}$) 的列车。例如，如果列车 I 是1号线1号列车，那么列车 $I+1$ 就是 1 号线 2 号列车。

乘客刷卡进站后，直接步行前往起点站的站台。从进站闸机到起点站站台的步行时间称为“进站步行时间”。根据可用运力（即乘客可能滞留的情况），该乘客可能乘坐 1 号线路上的列车 2 或列车 3，走完第一个路径区段。请注意，乘客无法乘坐列车 1，因为他/她到达站台时，列车 1 已经发车。在换乘站下车后，乘客步行前往 2 号线路上下一路径区段的上车站台。从下车站台到下一个上车站台的步行时间称为“换乘时间”。同样，根据可用运力，乘客可能乘坐 2 号线路上的列车 2 或列车 3（乘坐列车 4 不可行，因为若乘坐列车 4，乘客无法在当前出站刷卡时间离开）。在目的地车站的站台下车后，乘客直接步行前往出站闸机并刷卡出站。从下车站台到出站闸机的步行时间称为“出站步行时间”。

一般而言，考虑乘客 $i \in \mathcal{P}$ 在其路径 $m \in \mathcal{R}_{i,n}$ 的 n 次行程。设 $J_{i,n,m}$ 分别为该行程中路径 m 的路径区段数。行程方案 $\mathcal{H} = \{I_1, I_2, ..., I_{J_{i,n,m}}\}$ 定义为“一系列列车编号”（每个列车编号对应一个路径区段），表示乘客在系统中的一种可能移动方式，其中 $I_j \in \Lambda_{i,n,m}^{j}$ 表示第 I_j 个 j 路径区段上的列车编号。例如，在图2中，一个可行行程为 $\mathcal{H}$ = {*1*号线*2*号列车、*2*号线列车 *3*}，表示乘客先搭乘1号线的2号列车，然后搭乘2号线的3号列车。

值得注意的是，对于特定乘客 i，根据其进站刷卡和出站刷卡时间以及换乘时间的可行性，可行行程的数量有限。例如，在图2中，任何包含发车时间早于*1*号线*2*号列车的1号线列车的行程方案均不可行，因为根据乘客的进站刷卡时间，他们无法搭乘这些列车。设 $\Omega_{i,n,m}$ 为乘客 i 第 n 次行程中路径 m 的可行行程集合。我们有

$$\Omega_{i,n,m} = \{\{I_1, ..., I_{J_{i,n,m}}\},\ \forall\, I_j \in \Lambda_{i,n,m}^{j},\ \mid T_d(I_1) \geq t_{i,n}^{\text{in}}, T_a(I_{J_{i,n,m}}) \leq t_{i,n}^{\text{out}}, T_d(I_j) \geq T_a(I_{j+1}),\ \forall\, j = 1, ..., J_{i,n,m}\} \quad (9)$$

其中，$T_d(\cdot)$（相应地，$T_a(\cdot)$）是一个函数，用于返回相应路段的上车（相应地，下车）车站的列车发车（相应地，到达）时间。此信息可从自动车辆定位数据中获取。公式

9 表示可行的行程方案需要满足以下条件：1）列车 I_1 在$t_{i,n}^{\text{in}}$ 之后发车，使乘客能够上车（即， $T_d(I_1) \geq t_{i,n}^{\text{in}}$，假设最短进站步行时间为零）。2）末班列车（即列车 $I_{J_{i,n,m}}$）在$t_{i,n}^{\text{out}}$ 之前到达，使乘客能够在$t_{i,n}^{\text{out}}$ 出站刷卡（即， $T_a(I_{J_{i,n,m}}) \leq t_{i,n}^{\text{out}}$，假设最短出站步行时间为零）。3）列车 I_{j+1} 的发车时间晚于列车I_j 的到达时间，使乘客能够成功换乘（即， $T_d(I_j) \geq T_a(I_{j+1})$，假设最短换乘时间为零）。

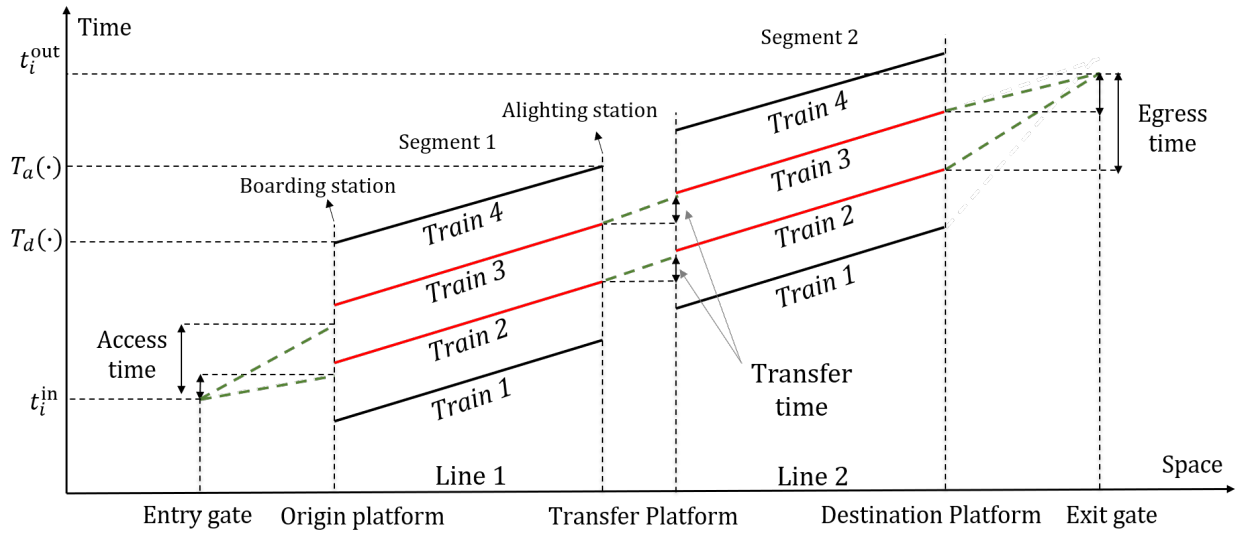


图 2：涉及一次换乘的行程时空图（改编自Zhu 等人(2017年b)）。红线表示可行行程路线

给定可行行程集合 $\Omega_{i,\ n,\ m}$，$\Pr(t_{i,\ n}^{\text{out}} \mid t_{i,\ n}^{\text{in}},\ r_{i,\ n} = m)$ 可改写为：

$$\Pr(t_{i,n}^{\text{out}} \mid t_{i,n}^{\text{in}}, r_{i,n} = m) = \sum_{\mathcal{H} \in \Omega_{i,n,m}} \Pr(t_{i,n}^{\text{out}} \mid \mathcal{H}, t_{i,n}^{\text{in}}, r_{i,n} = m) \cdot \Pr(\mathcal{H} \mid t_{i,n}^{\text{in}}, r_{i,n} = m). \tag{10}$$

我们首先推导 $\Pr(t_{i,n}^{\text{out}} \mid \mathcal{H}, t_{i,n}^{\text{in}}, r_{i,n} = m)$，即给定行程方案 $\mathcal{H}$、路径 m 和进站刷卡时间时，在时间 $t_{i,n}^{\text{out}}$出站刷卡的概率。由于行程方案包含末班列车的到达时间信息 $T_a(I_{J_{i,n,m}})$，该概率等价于出站步行时间等于 $t_{i,\ n}^{\text{out}} - T_a(I_{J_{i,n,m}})$ 的概率。设路径 m 的出站步行时间概率密度函数（PDF）为 $f_m^{\text{Eg}}(\cdot)$。那么，$\Pr(t_{i,n}^{\text{out}} \mid \mathcal{H}, t_{i,n}^{\text{in}}, r_{i,n} = m)$ 可表示为

$$\Pr(t_{i,n}^{\text{out}} \mid \mathcal{H}, t_{i,n}^{\text{in}}, r_{i,n} = m) = f_m^{\text{Eg}}(t_{i,n}^{\text{out}} - T_a(I_{J_{i,n,m}})). \tag{11}$$

请注意，方程11 使用概率密度表示概率，这不会影响最大似然估计中的参数估计。

现在我们来推导 $\Pr(\mathcal{H} \mid t_{i,\ n}^{\text{in}},\ r_{i,\ n} = m)$，即给定路径 m 和进站刷卡时间时，选择行程方案 $\mathcal{H}$的概率。由于路段 j 上乘坐的列车仅取决于

路段 $j-1$上乘坐的列车，而不取决于所有 $k>1$的 $j-k$，因此可以利用马尔可夫性质扩展这一概率：

$$\begin{aligned}\Pr(\mathcal{H} \mid t_{i,n}^{\text{in}}, r_{i,n}=m) &= \Pr(I_1, I_2, ..., I_{J_{i,n,m}} \mid t_{i,n}^{\text{in}}, r_{i,n}=m) \\ &= \Pr(I_1 \mid t_{i,n}^{\text{in}}, r_{i,n}=m) \cdot \prod_{j=2}^{J_{i,n,m}} \Pr(I_j \mid I_{j-1}, t_{i,n}^{\text{in}}, r_{i,n}=m) \end{aligned} \tag{12}$$

在以下内容中，我们分别推导 $\Pr(I_1 \mid t_{i,\,n,\,r_{i,\,n}}^{\text{in}}=m)$ 和 $\Pr(I_j \mid I_{j-1},\ t_{i,\,n,\,r_{i,\,n}}^{\text{in}}=m)$。

请注意，$\Pr(I_1 \mid t_{i,n,r_{i,n}}^{\text{in}}=m)$ 表示在路径m的第一个路段搭乘列车 I_1 的概率。此事件有两种不同的发生情形：1) [未发生滞留] 乘客在列车 I_1 和 I_1-1 的出发时间之间到达站台，并且搭乘列车 I_1，未发生滞留。2) [发生滞留] 乘客在列车 I_1-k 和I_1-k-1 出发之间到达站台，并在滞留 k 次后成功搭乘列车 I_1。给定可行行程集合 $\Omega_{i,n,m}$，乘客搭乘列车 I_1 前可能滞留的次数存在一个最大值。将 k 的上界记为 $B_{i,n,m}^{I_1}$，其中 $B_{i,n,m}^{I_1}=\arg\max_k\{k\in\mathbb{N} \mid \exists\, \mathcal{H}'\in\Omega_{i,n,m}$ s.t. $I_1'=I_1-k, I_1'\in\mathcal{H}'\}$, $\mathbb{N}$ 是自然数集合（包括零）。而列车 $I_1-B_{i,n,m}^{I_1}$ 表示乘客 i 在第一个路段能够搭乘的最早一班列车。

令 $t_{i,\,n,\,m}^{j}$ 表示乘客 i在路径 m 中从路径区段 $j-1$ 的下车站台到路径区段 j 的上车站台的步行时间，$t_{i,\,n,\,m}^{0}$ 表示进站步行时间。于是，$t_{i,\,n}^{\text{in}}+t_{i,\,n,\,m}^{0}$ 表示乘客到达其起点站站台的时间。因此，在列车 I_1-k 和 I_1-k-1 发车之间到达站台的概率可表示为

$$\Pr(T_d(I_1-k-1)\le t_{i,n}^{\text{in}}+t_{i,n,m}^{0}\le T_d(I_1-k) \mid t_{i,n}^{\text{in}}, r_{i,n}=m)=\int_{T_d(I_1-k-1)-t_{i,n}^{\text{in}}}^{T_d(I_1-k)-t_{i,n}^{\text{in}}} f_m^{\text{Ac}}(t)dt := \rho_{i,n,m}^{I_1,k} \quad \forall k=0,1,..,B_{i,n,m}^{I_1} \tag{13}$$

其中 $f_{\text{Ac}m}(\cdot)$表示路径 m的进站步行时间概率密度函数。公式 13可预先计算，因为 $f_m^{\text{Ac}}(\cdot)$是定积分。

令 $\eta_{i\,n\,,\,,\,m,k}^{j,k}$ 表示乘客 i的行程 n中，在路径 m 第 j个路段的上车站滞留 k 次的概率。令 $E_{i\,n\,,\,,\,m}^{I_j,k}$ 表示事件：“乘客 i 在第 n次行程中，于列车 I_j-k 和 I_j-k-1 发车之间到达路径 m 的路段 j 上车站，并滞留 k 次后搭乘列车 I_j”。我们有：

$$\Pr(E_{i,n,m}^{I_1,k} \mid t_{i,n}^{\text{in}}, r_{i,n}=m)=\rho_{i,n,m}^{I_1,k}\cdot\eta_{i,n,m}^{1,k} \quad \forall k=0,1,..,B_{i,n,m}^{I_1}. \tag{14}$$

那么，$\Pr(I_1 \mid t_{i,\,n,\,r_{i,\,n}}^{\text{in}}=m)$ 可改写为

$$\Pr(I_1 \mid t_{i,n}^{\text{in}}, r_{i,n}=m)=\sum_{k=0}^{B_{i,n,m}^{I_1}}\Pr(E_{i,n,m}^{I_1,k} \mid t_{i,n}^{\text{in}}, r_{i,n}=m)=\sum_{k=0}^{B_{i,n,m}^{I_1}}\rho_{i,n,m}^{I_1,k}\cdot\eta_{i,n,m}^{1,k}. \tag{15}$$

至此，$\Pr(I_1 \mid t_{i,\,n,\,r_{i,\,n}}^{\text{in}}=m)$ 的推导完成。 $\eta_{i\,n\,,\,,\,m}^{1,\,k}$ 可根据自动收费系统和自动车辆定位数据进行估计

using 一种高斯混合模型（Ma等人，2019），具体将在附录中说明 A.

现在，我们推导公式 12 中的 $\Pr(I_j \mid I_{j-1}, t^{\text{in}}_{i,n}, r_{i,n} = m)$，即在乘客已在路径 m 的第（$j-1$）个路段乘坐列车 I_{j-1} 的条件下，乘坐列车 I_j 的概率。其推导方式与 $\Pr(I_1 \mid t^{\text{in}}_{i,n}, r_{i,n} = m)$ 类似。乘客 i 可能在列车 $I_j - k$ 和 $I_j - k - 1$ 的发车时间之间到达路段 j 的上车站，并因滞留 k 次而乘坐列车 I_j（注意，$k = 0$ 表示未发生滞留）。在于 T_a 下车（I_{j-1}）的条件下，到达站台的时间处于列车 $I_j - k$ 和 $I_j - k - 1$ 的发车时间之间的概率可表示为

$$\begin{aligned}&\Pr(T_d(I_j - k - 1) \le T_a(I_{j-1}) + t^j_{i,n,m} \le T_d(I_j - k) \mid I_{j-1}, t^{\text{in}}_{i,n}, r_{i,n} = m)\\ &= \int_{T_d(I_j-k-1)-T_a(I_{j-1})}^{T_d(I_j-k)-T_a(I_{j-1})} f^{\text{Tr}}_{m,j}(t)dt = \tilde{\rho}^{I_j,k}_{i,n,m} \qquad \forall k = 0, 1, .., B^{I_j}_{i,n,m}.\end{aligned} \tag{16}$$

其中，$B^{I_j}_{i,n,m}$ 是在满足可行行程约束的条件下乘坐列车 I_j 时可能的最大滞留次数，定义为 $\arg\max_k\{k \in \mathbb{N} \mid \exists \mathcal{H}' \in \Omega_{i,n,m} \text{ s.t. } I'_j = I_j - k, I'_j \in \mathcal{H}'\}$。$t^j_{i,n,m}$ 是从列车 I_{j-1} 下车处到下一站台的换乘步行时间。$f^{\text{Tr}}_{m,j}(\cdot)$ 是 $t^j_{i,n,m}$ 的概率密度函数。为简化表达，定义 $\tilde{\rho}^{I_j,k}_{i,n,m}$。根据 $E^{I_j,k}_{i,n,m}$ 的定义，可得

$$\Pr(E^{I_j,k}_{i,n,m} \mid I_{j-1}, t^{\text{in}}_{i,n}, r_{i,n} = m) = \tilde{\rho}^{I_j,k}_{i,n,m} \cdot \eta^{j,k}_{i,n,m} \quad \forall k = 0, 1, .., B^{I_j}_{i,n,m}. \tag{17}$$

那么，$\Pr(I_j \mid I_{j-1}, t^{\text{in}}_{i,n}, r_{i,n} = m)$ 可以改写为

$$\Pr(I_j \mid I_{j-1}, t^{\text{in}}_{i,n}, r_{i,n} = m) = \sum_{k=0}^{B^{I_j}_{i,n,m}} \Pr(E^{I_j,k}_{i,n,m} \mid I_{j-1}, t^{\text{in}}_{i,n}, r_{i,n} = m) = \sum_{k=0}^{B^{I_j}_{i,n,m}} \tilde{\rho}^{I_j,k}_{i,n,m} \cdot \eta^{j,k}_{i,n,m}. \tag{18}$$

公式 8 中 $\mathcal{L}(\theta, \beta, \sigma)$ 的所有部分推导完成后，最大似然估计问题仍有两个挑战。首先，计算 $\Pr(t^{\text{out}}_{i,n} \mid t^{\text{in}}_{i,n}, r_{i,n})$ 需要滞留概率 $\eta^{j,k}_{i,n,m}$，以及三个概率密度函数 $f^{\text{Ac}}_m(\cdot)$、$f^{\text{Eg}}_m(\cdot)$ 和 $f^{\text{Tr}}_{m,j}(\cdot)$ 作为输入。概率密度函数可以通过实地实验获得，但滞留概率并不容易得到。本研究采用 Ma 等人（2019）提出的模型，利用自动收费系统和自动车辆定位数据估计 $\eta^{j,k}_{i,n,m}$。详情见附录 A。第二个挑战是，计算 $\Pr(v_i)$（公式 6）涉及对数量呈指数级增长的不同路径求和，并且需要对服从正态分布的随机变量进行积分，因此在数值上难以求解。下一节将推导一种新的基于条件概率的公式，以消除大量求和，并采用数值积分方法处理正态随机变量，从而得到一个可处理的似然函数，并实现高效的模型估计。

2.3. 可处理的对数似然函数

为消除公式 6中对不同路径求和项数呈指数级增长的问题，我们注意到，给定 α^k_i和 g_i后，乘客 i在每次行程中的路径选择彼此独立。数学上，

$$\Pr(v_i \mid \alpha^k_i, g_i = G_k) = \prod_{n=1}^{N_i} \Pr(t^{\text{out}}_{i,n}, t^{\text{in}}_{i,n} \mid \alpha^k_i, g_i = G_k) = \prod_{n=1}^{N_i} \sum_{m_n \in \mathcal{R}_{i,n}} \Pr(t^{\text{out}}_{i,n} \mid t^{\text{in}}_{i,n}, r_{i,n} = m_n) \cdot \pi^k_{i,n,m_n}[\alpha^k_i] \tag{19}$$

请注意，公式19 总共只有 $N_i \times |\mathcal{R}_{i,\ n}|$ 个求和项，而公式6 中的项数为 $|\mathcal{R}_{i,\ n}|^{N_i}$。根据公式 19，可以分别对 α_i^k 和 g_i进行积分和求和，从而得到 $\Pr(v_i)$。由于 α_i^k 是正态随机变量，我们采用近似数值积分方法，得到一个易于处理的公式。请注意，数值积分有多种求积公式（Davisand Rabinowitz，2007年）。本文采用最简单的中点法进行插值，因为这并非本研究的重点。令 α^{U} 和 α^{L} 分别为 α_i^k的上界和下界。我们将 $[\alpha^{\mathrm{L}}, \alpha^{\mathrm{U}}]$ 划分为长度相等的离散区间，区间长度为 Δ。令 $\mathcal{S}$ 为每个区间中点的集合。具体而言， $\mathcal{S} = \{\alpha^{\mathrm{L}} + \frac{k \cdot \Delta}{2} \mid \forall k$ =1、3、5、..... $|\mathcal{S}|$、 $|\mathcal{S}| = \frac{2(\alpha^{\mathrm{U}} - \alpha^{\mathrm{L}})}{\Delta} - 1\}$。因此，$\Pr(v_i)$可改写为

$$\Pr(v_i) \approx \sum_{k=1}^{K} \Pr(g_i = G_k) \cdot \sum_{\alpha_i^k \in \mathcal{S}} \Pr(v_i \mid \alpha_i^k, g_i = G_k) \cdot f(\alpha_i^k) \cdot \Delta \tag{20}$$

Δ 是决定近似精度与计算效率之间权衡的参数，其中较小的 Δ 表示积分划分更精细，但计算成本更高。

根据 $\Pr(v_i)$ 的新表达式，我们可以使用 $\mathcal{L}(\theta,\ \beta,\ \sigma) = \prod_{i \in \mathcal{P}} \Pr(v_i)$，以可处理的形式计算似然函数。

2.4. 模型估计

新的对数似然函数可表示为

$$\mathcal{LL}(\theta, \beta, \sigma) = \sum_{i \in \mathcal{P}} \log \Pr(v_i) = \sum_{i \in \mathcal{P}} \log \left[\sum_{k=1}^{K} \sum_{\alpha_i^k \in \mathcal{S}} \Pr(g_i = G_k) \cdot \Pr(v_i \mid \alpha_i^k, g_i = G_k) \cdot f(\alpha_i^k) \cdot \Delta \right] \tag{21}$$

由于 $\mathcal{LL}(\theta,\ \beta,\ \sigma)$ 是初等函数的组合，因此连续且可微。因此，可以使用任意一阶或二阶优化方法求解最大似然估计。本研究采用 BFGS算法（Nocedal and Wright，2006年）。BFGS 是一种拟牛顿法，仅使用一阶导数，并已在许多优化问题中表现出良好的性能。然而，由于该函数包含多个非线性项的乘积，其凸性未知。可能不是凹函数，因此在不同初始值下，BFGS算法可能收敛到不同的局部极小值。因此，我们在第3.1 节针对不同初始值进行了敏感性分析，并表明模型估计结果稳定。此外，数值结果显示，在合理范围的路径属性值内， $\mathcal{LL}$ 是凹函数。

获得最优参数 θ^*、 β^*和 σ^*后，我们根据数值估计的海森矩阵和克拉美–罗界，计算估计参数的 t 值。注意，由于 $\mathcal{LL}$二阶可微，因此也可以推导出解析海森矩阵。由于函数形式复杂，为简化计算，我们采用数值海森矩阵。本研究采用均匀网格间距下的四阶近似公式计算二阶导数（Fornberg，1988年）。具体公式见附录B（也可以使用其他近似公式）。利用二阶导数公式，我们可以计算 $\mathcal{LL}$（θ，β，σ）在点$\hat{}$ $\hat{}$（θ^*，β^*，σ^*）处的数值海森矩阵。将海森矩阵记作 H。注意，根据二阶最优性条件， H 为负半定，这在代数上等价于对数似然函数的局部凹性（Bierlaire，2020年）。

令 $\Theta = (\theta, \beta, \sigma)$ 为包含所有参数的向量。根据克拉美–罗界（Cramér，2016；Rao，1992），估计参数 Θ_k 的方差为

$$\mathrm{Var}[\hat{\Theta}_k] = -\hat{H}^{-1}_{k,k} \tag{22}$$

其中，$\hat{H}^{-1}$ 是海森矩阵的逆矩阵，$\hat{H}^{-1}_{k,k}$ 是其第 k 个对角元素。然后，相应的t值计算如下：

$$\text{t-value}[\hat{\Theta}_k] = \frac{\hat{\Theta}_k}{\sqrt{\mathrm{Var}[\hat{\Theta}_k]}} \tag{23}$$

3. 案例研究

3.1. 模型验证与敏感性测试

在现实中，很难收集乘客实际的路径选择数据。为验证所提出的方法，我们使用通过模拟乘客的路径选择、列车运营及其相互作用而生成的合成数据（Mo 等人，2022；Zhu 等人，2021；Mo 等人，2020a）。

图3 展示了合成城市轨道交通网络的配置，其中包含7个车站（A~G）和3条线路（红线、绿线和蓝线）。每条连接线上的数字表示车内出行时间。该网络取自香港港铁地铁系统，也代表了典型地铁网络在线路和换乘方面的结构。假设红线上行方向C站的站台十分拥挤，导致大量乘客滞留。其他所有站台均假设没有乘客滞留。

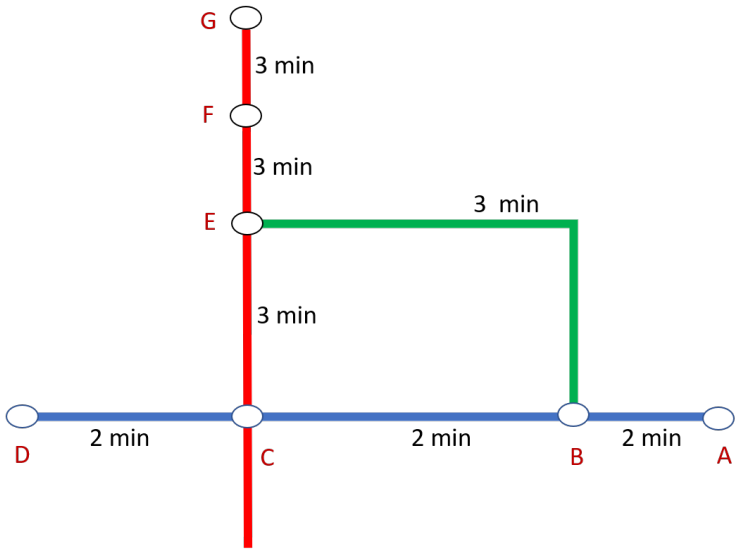


图3：合成城市轨道交通网络

为生成合成数据，我们假设乘客的路径选择行为取决于四个路径属性：1）车内时间（即路径上的列车运行时间）；2）车外时间（即接入、出站、换乘和未因滞留而产生的等待时间之和）；3）换乘次数（即路径上的换乘次数）；以及 4）拒绝等待时间（即在拥挤的站台因滞留而产生的等待时间

）。我们还假设，乘客的潜在群体可由两个社会人口统计变量 $x^{(1)}$和 $x^{(2)}$表征，其中 $x^{(1)}$取自 $\mathcal{U}[-4,4]$， $x^{(2)}$取自 $\mathcal{U}[-2,2]$。假设合成乘客分为两个潜在群体：“时间敏感型”（TS）和“舒适导向型”（CA）。时间敏感型乘客在选择路径时倾向于尽量缩短总出行时间，这意味着车内时间、车外时间和拒绝等待时间对乘客路径选择效用的影响相近。舒适导向型乘客偏好步行或等待时间较短的路径，即使车内时间可能更长。换言之，相较于车内时间，车外时间和拒绝等待时间对这类乘客的路径选择效用影响更大。

表3展示了用于生成合成数据的潜类别路径选择模型参数。这些参数根据Jin等人（2017年）中的调查结果选取。根据调查建模结果，我们将时间敏感型和舒适度敏感型乘客的车内时间及路径规模因子参数设为相同。此外，对于时间敏感型组（舒适度敏感型组），我们将选择参数设为调查结果中参数的 $0.7\times$ （$1.5\times$）。潜类别模型的参数分别设为1.5和0.6，分别对应 $x^{(1)}$ 和 $x^{(2)}$。

表3：合成数据生成的路径选择参数

变量	TS	CA	Jin等人（2017年）
路径选择参数			
车内时间	-0.2676	-0.2676	-0.2676
车外时间	-0.2980	-0.6386	-0.4257
换乘次数	-1.3068	-3.1737	-1.8669
拒绝等待时间	-0.3222	-0.7825	-0.4603
对数 PS_m	0.5815	0.5815	0.5815
σ^k	1	1	N/A
潜在类别参数			
$x^{(1)}$		1.5	
$x^{(2)}$		0.6	

表4 汇总了网络、列车运营和乘客的参数。我们通过根据进站刷卡时间模拟出站刷卡时间，为 9 个 OD 对（起点站 A 、 B 、 D，终点站 E 、 F 、 G ）生成合成数据。所有 OD 对均有 2 条路径。例如，OD 对 (A, E) 的可选路径为 A-B-E 和 A-B-C-E。在不失一般性的前提下，我们假设有 2,700 名乘客，每人进行 3 次行程（即 $N_i = 3$）。每次行程均关联一个随机选取的 OD 对。算法 1 详细说明了合成数据生成流程。

表 4：合成数据生成的系统设置

实体	设置
网络	步行距离 30–50 米（接入、出站、换乘）
列车运营	发车间隔 2+δ_H 分钟。 δ_H 从 $[-10, 10]$ 秒中均匀抽取 车内时间+δ_V （见图3） δ_V从 $[-20, 20]$ 秒中均匀抽取
乘客	步行速度分布服从对数正态分布 均值为 1.2m/s，标准差为 0.5m/s。 C站红线上行方向的滞留概率 分别为 20% 未发生滞留 , 50% 滞留一次 以及 30% 滞留两次

算法1合成数据生成

输入： 路径选择参数；乘客集合 $\mathcal{P}$。
输出： 合成自动收费数据（进站/出站站点/时间）

1：初始化样本实例数量 N。
2：**对** $i \in \mathcal{P}$ **中的每个元素执行**
3：　样本 $x_i^{(1)} \sim \mathcal{U}[-4, 4]$, $x_i^{(2)} \sim \mathcal{U}[-2, 2]$。
4：　抽样组 $g_i \sim \Pr(g_i; \theta)$。将该组记为 G_k。
5：　抽样 $\alpha_i^k \sim \mathcal{N}(0, \theta^k)$。
6：　**对于** $n = 1$到 N_i，**执行**
7：　　抽样 $t^{\text{in}}{}_{i,n} \sim \mathcal{U}$[7:00AM，10:00AM]。
8：　　为此次行程随机抽取一个 OD 对。
9：　　计算路径选择概率 $\pi^k{}_{i,n,m}[\alpha_i^k]$。
10：　　根据 π_i^k 抽取一条路径 $m \in \mathcal{R}_{i,\ n}$,$_{n,m}[\alpha_i^k]$。
11：　　根据路径 {v1} 抽取此次行程的实际出行信息（即列车运行时间、发车间隔、接入、出站、换乘和未能乘车的等待时间）。获取 t_i^{out} ,$_n$.
12：　保存 $t_i^{\text{in}}{}_m$ 和 $t_i^{\text{out}}{}_{,n}$ 和 OD 作为行程记录。
13：将所有行程记录合并为合成自动收费数据。

共生成了这 2,700 名乘客的 8,100 次行程。随后使用合成自动收费系统数据进行模型估计。由于我们已知选择参数的“真实”值（表3），因此可以验证模型的性能。我们使用 Python Fletcher，2013）`Scipy` 软件包中的 BFGS算法求解最大似然估计。$\alpha^{\text{L}} = -3$，$\alpha^{\text{U}} = 3$和 $\Delta = 1$用于数值积分。

表5 展示了路径选择参数的估计结果。括号中的百分比表示相对于“真实”参数的相对误差。请注意，模型估计采用了实际的步行速度分布和滞留概率。这些输入的敏感性分析见第3.1节。为便于比较，我们还估计了一个不含潜在类别的基准模型。潜类别模型对实际参数的估计平均百分比误差约为10%，其估计精度优于基准模型。TS组的车外时间参数误差最大（-33.2%），这可能是因为车外时间与换乘次数高度相关，增加了数值估计的难度。请注意，这些参数的绝对值相对较小，因此估计参数的绝对误差可以接受。

就拟合优度而言，零模型（所有参数均为零）的初始对数似然（记为 $\mathcal{LL}_0$）为 $-52,219.17$；最终潜类别模型的对数似然（记为 $\mathcal{LL}^*$）为 $-51,526.13$；最终基准模型的对数似然（记为 $\mathcal{LL}^{\text{B}}$）为 $-51,571.67$。我们进行对数似然比检验（Wilks，1938年），得到统计量 $\chi^2 = -2(\mathcal{LL}^{\text{B}} - \mathcal{LL}^*) = 91.08$，据此可知，在约束自由度为 5 时（即潜类别模型的参数数量减去基准模型的参数数量），p值为0。这表明，潜类别模型的设定显著优于基准模型。我们还计算了采用“真实”参数时的对数似然（记为 $\mathcal{LL}^{\text{True-para}}$），其值为 $-51,535.16$。该值小于 $\mathcal{LL}^*$，这意味着估计参数的拟合优度优于“真实”参数。这表明，估计误差可能主要来自数据生成过程中的随机误差，而非模型估计。

所有参数的绝对 t 值均大于 1.96，表明这些参数对乘客的路径选择具有显著影响。这是合理的，因为合成数据是根据这些参数生成的。我们还观察到，与其他成本参数相比，车内时间的显著性最高，这与调查结果一致（Jin等人，2017年）。

表5：合成数据的估计结果

变量	潜在类别 估计值（误差）	t值	基准 估计值（误差）	t值
选择模型				
车内时间	-0.2599（-2.9%）	-10.62	-0.2254 (-15.8%)	-10.03
时间敏感型：车外时间	-0.1991 (-33.2%)	-2.88	-0.3010 (1.0%)	-6.88
时间敏感型：换乘次数	-1.3327 (+1.9%)	-4.91	-1.8777 (+43.7%)	-16.77
时间敏感型：拒绝等待时间	-0.3789（+17.6%）	-6.18	-0.4901（+52.1%）	-18.80
舒适度敏感型：车外时间	-0.5419 (-15.14%)	-3.82	-0.3010（-52.9%）	-6.88
舒适度敏感型：换乘次数	-2.8071（-11.5%）	-4.98	-1.8777（-40.8%）	-16.77
舒适度敏感型：拒绝等待时间	-0.7298（-6.7%）	-6.35	-0.4901 (-37.4%)	-18.80
log PS_m	0.6758 (+16.2%)	12.72	0.6054 (+4.1%)	12.31
σ^k	0.9191 (-8.1%)	11.05	0.9122 (-8.8%)	14.40
潜在群体模型				
$x^{(1)}$	0.6213（+3.6%）	4.99	N/A	N/A
$x^{(2)}$	1.8637（+24.3%）	7.01	N/A	N/A
乘客数量：	2,700		8 100 。观测数量：，：-。；：-。；：-。；	
$\mathcal{LL}_0$ 52,219 17 $\mathcal{LL}^*$ 51,526 13 $\mathcal{LL}^{\text{B}}$ 51,571 67 $\mathcal{LL}^{\text{True-para}}$：-51,535.16$\chi^2 = -2$（$\mathcal{LL}^{\text{B}} - \mathcal{LL}^*$） $= 91.08$，似然比检验 p值：0				

为进一步验证模型性能，我们进行了敏感性分析，以探究参数初始化对模型性能的影响。此外，我们还评估了步行速度分布和滞留分布估计不准确是否会影响模型性能。

图4 展示了参数初始化的敏感性分析。总共进行了 20 次实验。每次实验中，所有参数的初始值均从 $\mathcal{U}[-5,5]$中均匀抽取。我们观察到，无论参数初始值如何，最终估计得到的参数都收敛至相同的数值，这表明模型估计对参数初始化具有稳健性。

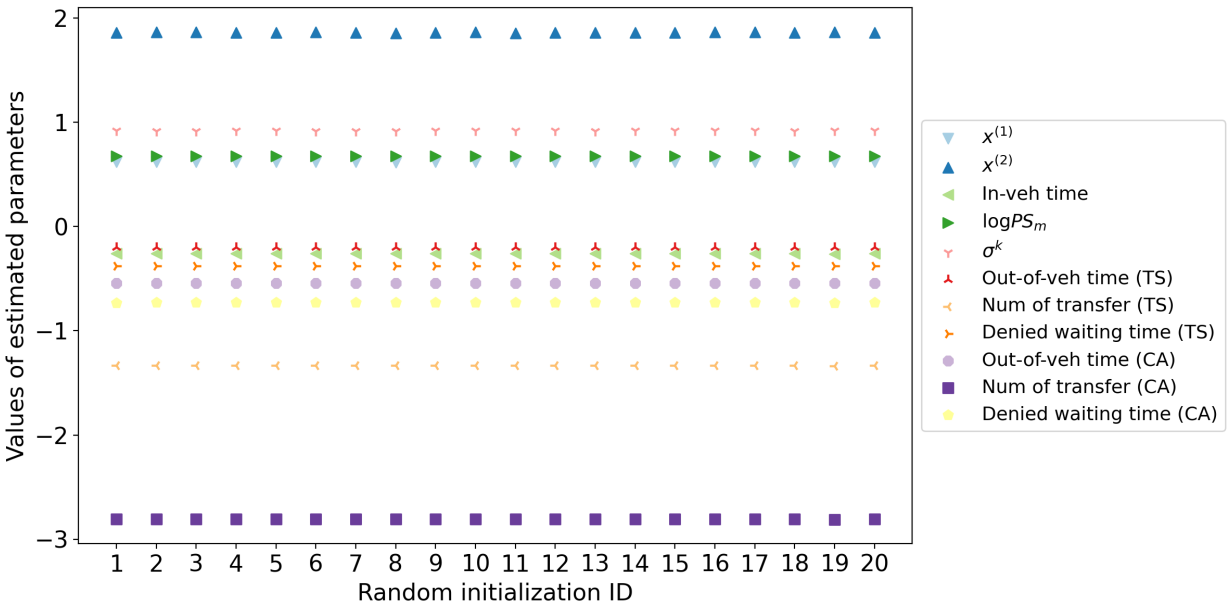
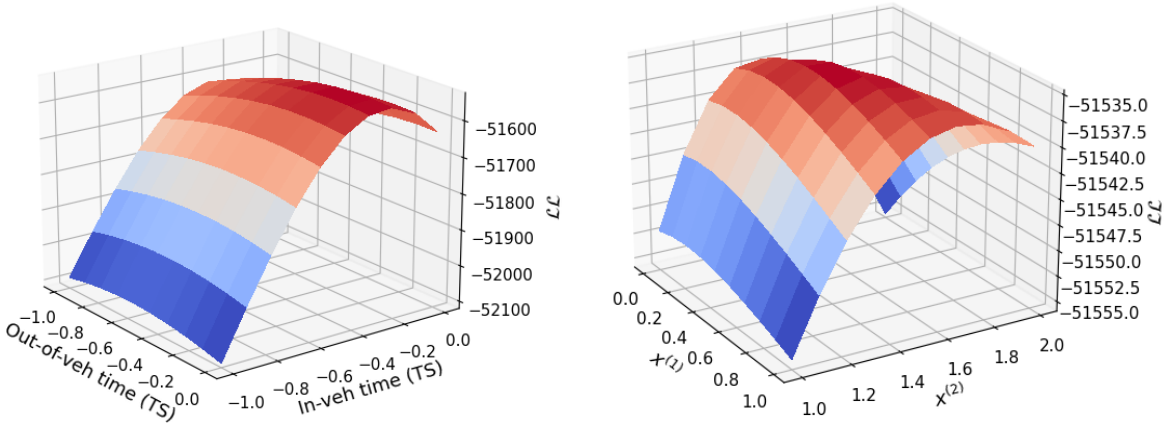


图4：不同初始值下的估计参数

设置

图5 展示了 $\mathcal{LL}$ 该值随变量值变化的情况。对数似然函数在最优值附近呈凹形[1], 这进一步表明估计结果具有稳健性。



(a) $\mathcal{LL}$ 与车内时间和车外时间的关系（时间敏感型）

(b) $\mathcal{LL}$与 $x^{(1)}$ 和 $x^{(2)}$的关系

图5：对数似然函数曲面

图6展示了输入不同步行速度时的模型估计结果。由于现实中估计乘客的步行速度可能并不准确，我们评估了模型对于步行速度分布误差的稳健性。设 μ^{WS} 和 σ^{WS} 分别为生成合成数据时实际步行速度的均值和标准差（即 $\mu^{\mathrm{WS}} = 1.2m/s$ 和 $\sigma^{\mathrm{WS}} = 0.5m/s$）。估计模型时，我们将速度分布参数设为 $(\Gamma_1 \cdot \mu^{\mathrm{WS}},\ \Gamma_2 \cdot \sigma^{\mathrm{WS}})$，其中

[1]由于篇幅有限，我们仅展示关于四个变量的函数曲线

Γ_1, $\Gamma_2 \in \{0.8, 1, 1.2\}$，表示速度参数输入中的不同扰动（$\Gamma_1 = \Gamma_2 = 1$表示没有误差）。图6 显示，步行速度分布的变化对模型性能影响不大。

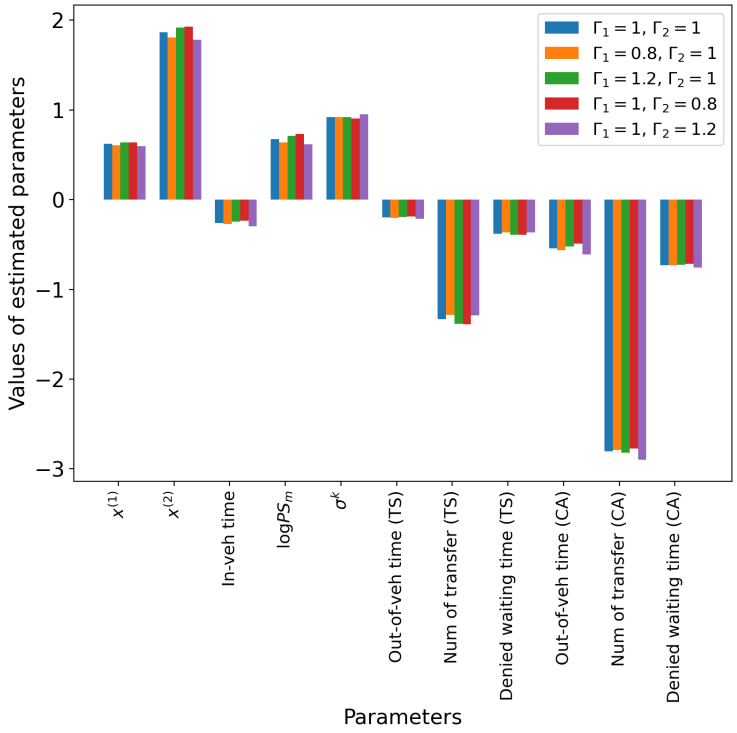


图6：步行速度输入的敏感性分析

图7 展示了红线 C 站上行方向在不同滞留概率输入下的估计结果，反映了滞留概率存在估计误差时模型的表现。同样，选择滞留概率进行敏感性分析，是因为这些数值由数据估计得出，可能存在误差。比较了三种情景：实际拥挤程度（20% 未滞留、50% 滞留一次、30% 滞留两次，即无误差）、拥挤程度较低（80% 未滞留、20% 滞留一次、0% 滞留两次）以及拥挤程度较高（10% 未滞留、20% 滞留一次、70% 滞留两次）。可以看出，车内时间参数 $x^{(1)}$和 $x^{(2)}$ 对滞留输入并不敏感。但其他参数（如换乘次数和车外时间）则受到较大影响。原因可能在于，滞留估计误差会影响模型对其他因素如何影响出行时间的评估。也就是说，额外 10 分钟的行程时间可能是由更多换乘、较长的车外时间或滞留造成的。如果滞留情况估计不准确，其他因素对总出行时间（以及乘客选择）的影响可能会产生偏差。因此，结果凸显了在路径选择估计中纳入拥挤和容量约束的重要性。

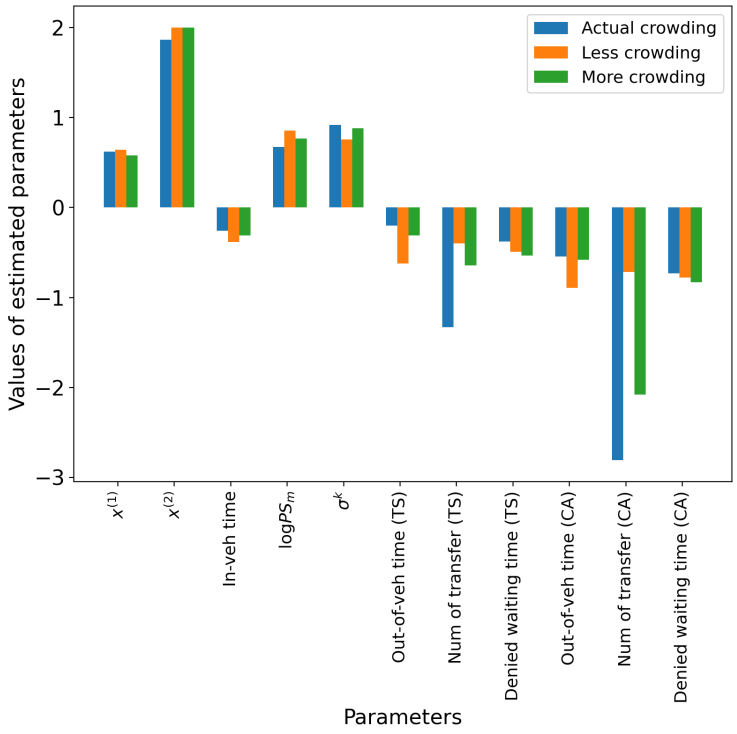


图 7：滞留输入的敏感性分析

3.2. 港铁实证案例研究

所提出的方法也应用于香港港铁网络的实际自动收费系统和自动车辆定位数据。图8 展示了港铁网络及选定的 OD 对区域（起点位于黑色虚线框内，目的地位于红色区域）。选择这些 OD 对的原因是：1）每对 OD 之间都有多条路径，适合应用路径选择建模；2）这些车站的 OD 客流量足够大，可以估计滞留概率分布（Ma等人，2019）。

我们随机选取了 3,425 名在这些 OD 对之间出行的乘客。我们考虑出发时间在晚高峰（下午 5:30 至 7:30）的行程。最终，从 2018年7月的自动收费系统数据中共收集到 6,425 次行程（即每位乘客平均有 1.88 次行程）。假设步行时间服从对数正态分布，其均值和方差根据港铁员工的数据进行校准。 $\alpha^{\mathrm{L}} = -9$、 $\alpha^{\mathrm{U}} = 9$和$\Delta = 1$用于数值积分。

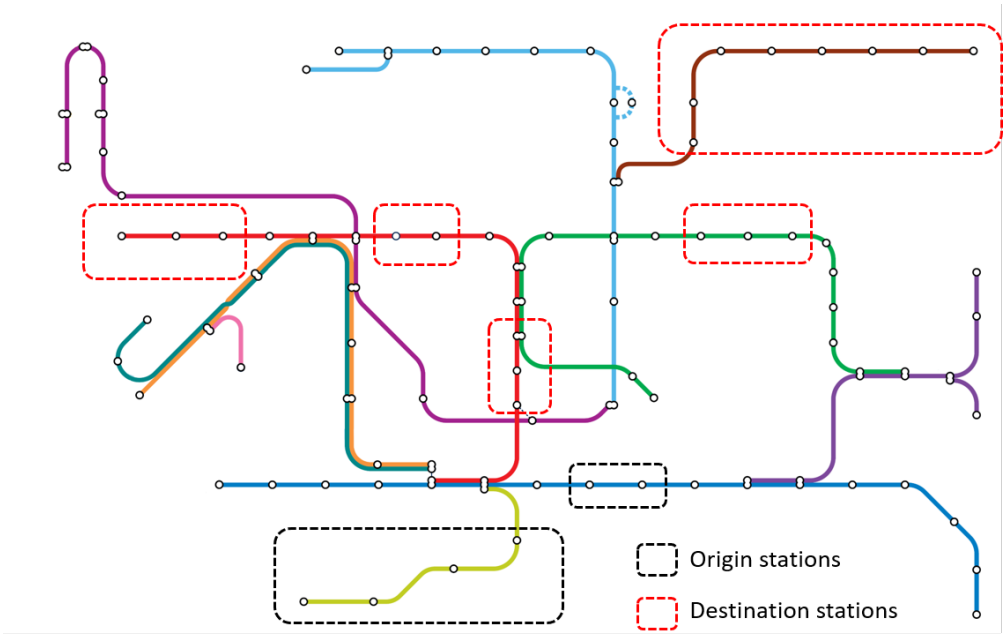


图8：香港地铁网络

表6：地铁数据的描述性统计

变量	Mean	标准差	Max	Min
个人特征				
每周平均出行天数	4.31	2.06	7.00	0.12
首次出行出发时间的标准差（小时）	0.68	0.51	3.54	0.01
覆盖 70% 行程的最少站点数	3.22	0.15	15	1
是否为学生（是 = 1）	0.10	0.30	1	0
路径属性				
车内时间（分钟）	34.0	15.5	85.8	3.50
车外时间（分钟）	9.20	1.11	24.6	0.50
拒绝等待时间（分钟）	1.03	2.48	16.3	0.0
乘客数量：3,425。出行次数：6,425				

我们假设乘客的潜在群体可由以下属性刻画，这些属性均可直接从自动收费系统数据中提取：1）出行频率，定义为每周平均出行天数；2）时间安排灵活性，以工作日首程出发时间的标准差衡量；3）出行空间集中度，定义为一个月内覆盖 70% 行程所需的最少车站数；4）持卡人是否为学生（虚拟变量）。所有这些属性均根据 2018年7月的自动收费系统数据计算。数据的描述性统计见表 6。本实验考虑两个潜在群体。之所以考虑两个而非更多群体，是因为两个潜在群体的估计结果更易于解释。除“换乘次数”外，路径属性与合成数据实验中的相同。由于“换乘次数”与“车外时间”高度相关，因此将其从模型中剔除。与合成数据实验类似，我们令两个群体的车内时间参数相同，以便比较车外时间与换乘次数的尺度。

表7：真实世界数据的估计结果

变量	组 1（舒适度敏感型）		组 2（时间敏感型）	
	估计值	t值	估计值	t值
选择模型				
车内时间	-0.2785	-18.13	与组 1 相同	
车外时间	-1.1320	-2.09	-0.7457	-9.04
拒绝等待时间	-3.0450	-1.95	-0.4517	-9.90
log PS_m	1.2611	4.38	与组 1 相同	
σ^k	2.7294	11.99	与组 1 相同	
潜在群体模型（组 *2* 设为基准组）				
ASC[1]	-0.9644	-6.83	0（固定）	
每周平均出行天数	-0.0987	-3.91	0（固定）	
首次出行出发时间的标准差	0.8667	4.19	0（固定）	
覆盖 70% 行程的最少站点数	0.1865	0.40	0（固定）	
如果是学生（是 = 1）	1.0192	2.14	0（固定）	

[1]: 特定替代项常数：备选项特定常数 乘客数量：3,425。观测数量：6,425$\mathcal{LL}_0$: -32,157.84；最终 $\mathcal{LL}^*$: -30,867.48$\chi^2 = -2(\mathcal{LL}_0 - \mathcal{LL}^*) = 2$，580.72，似然比检验 p值：0

估计结果见表 7。两个潜在群体分别称为组 1 和组 2。群体分配多项Logit模型（公式1）中，将组 2 设为基准备选项。结果显示，所有参数的符号和尺度均合理。所有与时间相关的参数均显著为负。车内时间参数的值为 -0.2785，与调查结果（-0.2676）相近（Jin等人，2017年）。对于组 1 和组 2，车外时间和拒绝等待时间参数的绝对值均大于车内出行时间参数的绝对值，反映出乘客对步行与候车时间比对在车内乘坐时间更加敏感。这些结果也与调查结果一致。

比较组 1 和组 2 的结果可以发现，车外时间和未能乘车的等待时间对组 1 乘客路径选择效用的影响更大，这表明组 1 可能属于舒适导向型（CA），而组 2 属于时间敏感型（TS）。决定潜在群体归属的参数显示，舒适导向型乘客的出行频率较低（即每周平均出行天数的影响为负），时间安排灵活性较高（即首次出行出发时间标准差的影响为正），且学生更有可能属于这一群体。这些结果表明，舒适导向型乘客很可能是非规律用户，他们使用港铁系统主要是为了非工作活动（如娱乐）。因此，他们更注重出行舒适度，并偏好步行与候车时间较短的路径。相比之下，时间敏感型乘客出行频率较高，时间安排灵活性较低，他们使用地铁系统可能主要是为了规律通勤出行。此外，他们更注重节省总出行时间，以便准时抵达目的地。空间集中度变量（即覆盖 70% 行程的最少站点数）对归入舒适导向型群体虽有正向影响，但在统计上并不显著（t值为 0.4）。σ^k对两个群体都具有显著影响，这意味着不同群体的面板效应各不相同（即有些人的出行模式更稳定，而有些人则不然）。

4. 结论

了解乘客路径选择对于城市轨道交通系统的运营管理至关重要，尤其是在拥挤情况下。本文提出一种概率方法，利用自动收费系统（进出站刷卡）和自动车辆定位数据，在考虑列车运能约束的情况下估计路径选择模型。通过带面板效应的潜类别模型，捕捉选择异质性和纵向行为相关性。本文采用乘客列车分配模型来描述乘客的移动，并对进出站、因拥挤而滞留以及换乘过程进行显式建模。为便于模型估计，本文推导出一个易于处理的似然函数。基于数值估计的海森矩阵和克拉美–罗界，计算估计参数的 t值。该方法由数据驱动，能够灵活适配不同的选择模型，并且易于使用非线性优化器求解。

使用合成数据验证模型性能，并据此估计个体选择参数。敏感性分析证实，该模型对参数初始化以及输入中的小误差（步行速度和滞留分布）具有稳健性。分析还表明，在拥挤条件下，忽略容量约束（滞留）可能导致路径选择参数的估计值产生偏差。研究还使用来自香港港铁地铁系统的真实世界数据应用了该模型。结果揭示了两类不同的乘客：时间敏感型（TS）和舒适导向型（CA）。TS 乘客通常是出行频率较高、时间安排灵活性较低的固定通勤者。他们更倾向于选择行程出行时间较短的路径。CA 乘客更注重出行舒适度，偏好步行与候车时间较短的路径。

案例研究可以得出一些政策启示。由于这两类乘客对车内与车外时间的重视程度不同，公共交通机构可以利用这一发现，设计乘客接受度更高的定制化路线推荐系统。此外，路线推荐还可以通过建议 TS 和 CA 乘客在高峰时段选择不同路线，帮助缓解拥堵。值得关注的未来研究方向包括：1）探究网络干预（如网络扩展和需求管理政策）下，选择偏好和学习行为随时间的演变。2）开发下游模型，利用潜在乘客群体信息，以改进路线推荐或票价政策。

5. 作者贡献

Baichuan Mo：概念化、方法学、软件、正式分析、数据整理、撰写初稿、撰写——审阅与编辑、可视化。**ZhenliangMa**：概念化、方法学、监督、正式分析、数据整理、撰写初稿、撰写——审阅与编辑。**HarisN.Koutsopoulos**：概念化、监督、正式分析。**Jinhua Zhao:** 概念化、监督、项目管理、资金获取。

6. 致谢

作者感谢芝加哥交通管理局（CTA）对本研究的支持，以及为本研究提供数据。

附录

附录 A. 滞留概率校准

滞留概率可以通过 Ma等人提出的高斯混合模型进行估计Ma等人（2019）。其核心思想是，因滞留次数不同而滞留的乘客，其出行时间也会不同。令t_i^{Jn}为表示乘客 i 出行时间的随机变量（即出站刷卡时间减去进站刷卡时间）。图A.9展示了特定 OD 对的行程时间分布示例。可以观察到三个簇，分别表示乘客在起点站滞留了 0、1 和 2 次。因此，我们可以将行程时间分布建模为高斯混合模型：

$$\Pr(t_i^{\text{Jn}};\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{w}) = \sum_{c=0}^{C} w_c \cdot \Phi(\mu_c,\sigma_c) \tag{A.1}$$

其中，w_c 是簇 c, $w_c > 0$ 和 $\sum_{c=0}^{C} w_c = 1$中的乘客比例（未知）。 C 是簇的总数（即起点站滞留时间的最大数量）。 $\Phi(\mu_c,\sigma_c)$是高斯分布 $\mathcal{N}(\mu_c,\sigma_c)$的概率密度函数。

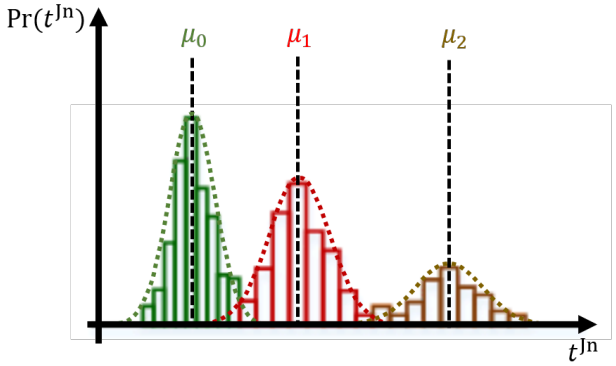


图 A.9：行程时间分布示例

可通过求解以下问题来估计高斯混合模型：

$$\max_{\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{w}} \quad \sum_{i\in\mathcal{P}'} \log(\Pr(t_i^{\text{Jn}};\boldsymbol{\mu},\boldsymbol{\sigma},\boldsymbol{w})) \tag{A.2a}$$

$$\text{s.t.} \quad \text{Auxiliary Constraints} \tag{A.2b}$$

$$\sum_{c\in C} w_c = 1 \tag{A.2c}$$

$$w_c \geq 0 \quad \forall c = 0,1,...,C \tag{A.2d}$$

其中 $\mathcal{P}'$ 是用于估计滞留概率的乘客集合。辅助约束用于确保模型稳定性。这些约束包含关于行程时间分布的先验经验知识，例如， μ_c 与 μ_{c+1} 之间的差值应接近一个发车间隔；未发生滞留时的平均行程时间（即 μ_0）应接近接入时间、出站时间与车内时间之和。关于该模型的更多信息，请参见Ma等人(2019)。

该模型针对特定车站和时间，因此可以通过调整 $\mathcal{P}'$，校准系统中各车站在不同时间间隔内的滞留概率。估计得到的 w_c （即集群 c中的乘客比例）是相应车站和时间段内发生 c 次滞留的概率。

附录B。海森矩阵的数值计算

根据 Fornberg (1988年)，给定一般函数 $f(x,\ y)$，可按四阶精度计算二阶导数：

$$\frac{\partial^2 f(x,y)}{\partial x^2}|_{x_0,y_0} = \frac{1}{h_x^2}\left[\frac{-1}{12}f(x_{-2},y_0) + \frac{4}{3}f(x_{-1},y_0) + \frac{-5}{2}f(x_0,y_0) + \frac{4}{3}f(x_{+1},y_0) + \frac{-1}{12}f(x_{+2},y_0)\right] + \mathcal{O}(h_x^4) \tag{B.1}$$

and

$$\frac{\partial^2 f(x,y)}{\partial x \partial y}|_{x_0,y_0} = \frac{1}{h_x h_y}\left[\frac{-1}{48}f(x_{-2},y_{-2}) + \frac{1}{3}f(x_{-1},y_{-1}) + \frac{-1}{3}f(x_{-1},y_{+1}) + \frac{-1}{3}f(x_{+1},y_{-1}) + \frac{1}{3}f(x_{+1},y_{+1}) + \frac{1}{48}f(x_{+2},y_{-2}) + \frac{1}{48}f(x_{-2},y_{+2}) + \frac{-1}{48}f(x_{+2},y_{+2})\right] + \mathcal{O}(h_x^2 h_y^2) \tag{B.2}$$

其中 h_x 和 h_y 分别是 x 和 y的小扰动。 x_k (y_k)表示 $x_0 + kh_x$ ($y_0 + kh_y$) 。推导基于均匀网格间距下的泰勒级数展开（Fornberg，1988年）。

参考文献

Arellano, M., Honoré, B., 2001. 面板数据模型：若干近期进展，载于：《计量经济学手册》。爱思唯尔。第5卷，第3229–3296页。Bagchi, M., White, P.R., 2005. 公共交通智能卡数据的潜力。《Transport Policy》12，464–474。Ben-Akiva, M., Lerman, S.R., 2018. 离散选择分析：理论及其在出行需求中的应用。Transportation Studies。Bierlaire, M., 2020. PandasBiogeme 简介。PandasBiogeme 简介。Briand, A.S., Côme, E., Trépanier, M., Oukhellou, L., 2017. 利用智能卡数据分析公共交通乘客行为的逐年变化。《交通研究C辑：新兴技术》79，274–289。Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M., 2016. 大数据与小数据在出行行为（即人类移动性）分析中的作用。《交通研究C辑：新兴技术》68，285–299。Cramér, H., 2016. 《统计学的数学方法》（PMS-9），第9卷。普林斯顿大学出版社。Davis, P.J., Rabinowitz, P., 2007. 数值积分方法。Courier Corporation。

Fletcher, R., 2013. 《实用优化方法》。John Wiley & Sons。Fornberg, B., 1988. 任意间距网格上的有限差分公式生成。计算数学 51，699–706。Goulet-Langlois, G., Koutsopoulos, H.N., Zhao, J., 2016. 推断公共交通用户多周活动序列中的模式。《交通研究C辑：新兴技术》64，1–16。Haywood, L., Koning, M., 2015. 公共交通拥挤成本的分布：来自巴黎的新证据。《交通研究A辑：政策与实践》77，182–201。Hess, S., Ben-Akiva, M., Gopinath, D., Walker, J., 2009. 潜在类别选择模型中的偏好异质性、相关性与弹性，载于：第88届交通研究委员会年会，Citeseer。第09–2428页。Hoogendoorn-Lanser, S., Bovy, P., 2007. 通过纳入行程分段特定路径规模因子，对多模式路径选择中的重叠进行建模。《交通研究纪要》2003，74–83。Hörcher, D., Graham, D.J., Anderson, R.J., 2017. 利用大规模智能卡和车辆位置数据估计拥挤成本。《交通研究B辑：方法论》95，105–125。Jin, F., Yao, E., Zhang, Y., Liu, S., 2017. 考虑感知换乘阈值的地铁乘客路径选择模型及其应用。PLOS ONE 12，e0185349。Koutsopoulos, H.N., Ma, Z., Noursalehi, P., Zhu, Y., 2019. 用于规划、监测、控制和信息服务的公共交通数据分析，载于：出行模式、大数据与交通分析。爱思唯尔，第229–261页。Li, Z., Hensher, D.A., 2011. 拥挤与公共交通：支付意愿证据综述及其在项目评估中的相关性。《Transport Policy》18，880–887。Ma, Z., Koutsopoulos, H.N., Chen, Y., Wilson, N.H., 2019. 城市轨道交通系统拒载估计：替代形式与比较分析。《交通研究纪要》2673，771–778。Mo, B., Li, R., Dai, J., 2020a. 动态起讫点需求估计：利用车牌识别数据的混合框架。《计算机辅助土木与基础设施工程》35，734–752。Mo, B., Ma, Z., Koutsopoulos, H.N., Zhao, J., 2020b. 城市轨道交通系统的容量约束网络性能模型。《交通研究纪要》2674，59–69。Mo, B., Ma, Z., Koutsopoulos, H.N., Zhao, J., 2022. 利用智能卡数据对城市轨道交通系统进行事后路径选择估计：聚合时空超网络方法。《交通科学》。Mo, B., Wang, Q.Y., Moody, J., Shen, Y., Zhao, J., 2021. 现有出行方式的主观评价和惯性对采用自动驾驶按需出行的影响。《交通研究C辑：新兴技术》130，103281。Nocedal, J., Wright, S., 2006. 数值优化。施普林格科学与商业媒体。Pelletier, M.P., Trépanier, M., Morency, C., 2011. 公共交通中的智能卡数据应用：文献综述。《交通研究C辑：新兴技术》19，557–568。Prato, C.G., 2009. 路径选择建模：过去、现在与未来的研究方向。《选择建模杂志》2，65–100。Rao, C.R., 1992. 统计参数估计中信息及其可达到的精度，载于：《统计学突破》。施普林格，第235–247页。Raveau, S., Guo, Z., Muñoz, J.C., Wilson, N.H., 2014. 地铁网络路径选择的行为比较：时间、换乘、拥挤、拓扑结构与社会人口特征。《交通研究A辑：政策与实践》66，185–195。Raveau, S., Muñoz, J.C., De Grange, L., 2011. 地铁拓扑路径选择模型。《Transportation

《研究A辑：政策与实践》45，138–147。Sun, L., Lu, Y., Jin, J.G., Lee, D.H., Axhausen, K.W., 2015. 一种用于地铁网络客流分配的集成贝叶斯方法。《交通研究C辑：新兴技术》52，116–131。Sun, Y., Schonfeld, P.M., 2016. 利用自动收费系统数据估计基于时刻表的轨道交通路径选择。《交通工程学报》142，04015037。Sun, Y., Xu, R., 2012. 轨道交通旅行时间可靠性与乘客路径选择行为估计：基于自动收费系统数据的分析。《交通研究纪要》2275，58–67。Wilks, S.S., 1938年。用于检验复合假设的似然比大样本分布。《数学统计年刊》9，60–62。Zhang, Y., Yao, E., Wei, H., Zheng, K., 2017年。地铁网络中的受约束多项式概率单位路径选择模型：构建、估计与应用。PLOS ONE 12，e0178789。Zhang, Y., Yao, E., Zhang, J., Zheng, K., 2018. 结合自我报告的显示偏好与智能卡数据估计地铁乘客路径选择。《交通研究C辑：新兴技术》92，76–89。Zhao, J., Zhang, F., Tu, L., Xu, C., Shen, D., Tian, C., Li, X.Y., Li, Z., 2016. 利用智能卡数据估计复杂地铁系统中的乘客路径选择模式。《IEEE智能交通系统汇刊》18，790–801。Zhou, F., Shi, J.g., Xu, R.h., 2015. 基于自动收费系统数据的城市轨道交通路径选择比例估计方法。《工程数学问题》2015。Zhu, Y., Koutsopoulos, H.N., Wilson, N.H., 2017a. 利用自动化数据推断拥挤地铁系统中的滞留乘客。《交通研究程序汇刊》23，362–379。Zhu, Y., Koutsopoulos, H.N., Wilson, N.H., 2017b. 基于自动化数据的概率乘客列车分配模型。《交通研究B辑：方法论》104，522–542。Zhu, Y., Koutsopoulos, H.N., Wilson, N.H., 2021. 拥挤的城市轨道交通网络乘客行程推断模型。《交通研究C辑：新兴技术》123，102896。