多源交通数据融合下的城市出行需求动态预测

刘红波

广东利通科技投资有限公司 广东省 广州市 510530

【摘　要】：面向大规模通勤与天气活动扰动下的分钟级出行需求研判，提出一套多源交通数据融合的动态预测方法。以公交刷卡、地铁闸机、共享单车订单与出租车轨迹为基础，构建覆盖统一时间窗与交通小区的时空对齐流程，提取时间、空间与跨方式交互特征，并设计注入路网先验的注意力 LSTM 模型。以上海中心城区一周数据为例，5 分钟粒度测试显示，所提模型在微平均指标上取得 MAE 0.85、RMSE 1.72、MAPE 12.6%，较 ARIMA 在 MAE 上降幅 31%，较单一 LSTM 降幅 13%，同时在早晚高峰与突发窗口保持稳定优势，核心站点改善更显著。在 5、10 与 15 分钟预测步长下均保持相对优势，短步长改善更为显著。结果表明，多源互补与注意力加权可有效提升分钟级预测的准确性与泛化能力，为运力调度与应急处置提供更可靠的先验支撑。

【关键词】：多源交通数据融合；分钟级出行需求预测；注意力 LSTM；时空对齐；路网先验

0 引言

大规模通勤峰谷切换、突发天气与活动冲击使城市客流在分钟级呈现显著波动，运力调度与应急处置亟需在更细粒度上把握需求演化。依赖单一数据源难以兼顾覆盖面与解释力：公交刷卡缺失下车与非公交出行，地铁闸机难描绘换乘链路，出租车与网约车轨迹存在空驶偏差与过拟合风险，移动信令则受采样与定位误差影响。多源融合由此成为提升稳定性与泛化能力的关键路径，通过将轨道与公交构成的出行基底与出租车和共享单车的弹性供给结合，并引入气象、节假日与 POI 等外部信息，可更完整地表征出行强度、结构与外生扰动。现有研究在特征级与决策级融合、以及嵌入路网拓扑的图时空模型方面已取得进展，但跨平台时序对齐、隐私保护计算、跨城迁移与概念漂移自适应等问题仍未充分解决，流数据下的轻量部署与可解释性机制亦有待完善。基于此，本文面向分钟级预测提出统一的预处理与特征体系，并构建注入路网先验的注意力 LSTM 框架，以期在提升预测精度的同时兼顾时段稳定性与部署可用性。

1 多源交通数据融合在出行需求预测中的应用背景

鉴于大规模通勤、突发天气与活动导致的客流波动，城市交通管理部门需要把出行需求的变化在分钟级时空尺度上进行研判，以支撑运力调度与应急处置。单一数据源在覆盖面与解释力上存在局限：例如公交刷卡仅反映持卡乘客的上车事件，缺失下车与非公交出行，时间分辨率受车次制约；出租车与网约车 GPS 虽具高频轨迹，却易产生空驶偏差与过拟合；轨道闸机数据无法刻画换乘链路；移动信令数据作为蜂窝网络产生的人群轨迹信号，受归一化采样与基站漂移影响，难以精确定位。由此推导，把多源交通数据进行时空对齐与语义融合，能把出行强度、出行结构与外生扰动纳入建模框架，借助公交与轨道闸机刻画交通基底，叠加出租车与网约车订单捕捉弹性供给，引入移动信令、POI 与气象事件来补全群体分布与事件冲击，预测稳定性与泛化能力有望提升。现有研究已在特征级融合、决策级融合以及以路网拓扑为先验的图神经网络驱动的时空模型上取得进展，然而在跨平台数据的时序对齐、隐私保护计算、跨城迁移与概念漂移自适应方面仍存不足，流数据下的轻量部署与可解释性机制也亟待完善。

2 多源交通数据融合下的出行需求预测模型构建

2.1 多源交通数据的预处理方法

鉴于分钟尺度的需求预测对输入稳定性与时空一致性提出更高要求，本研究把公交刷卡、地铁闸机、共享单车订单以及出租车 GPS 轨迹作为数据来源，围绕采样频率差异、定位误差与事件语义不一致等问题开展预处理。公交刷卡记录体现乘客上车时刻与线路站点，地铁闸机提供进出站时间与站点编码，共享单车订单记录借还点与时间戳，出租车轨迹包含高频坐标、速度与载客状态。预处理目标在于把多源观测映射到统一时间粒度与空间网格，并把噪声与缺口控制在可控范围。

在数据清洗阶段，运用地理围栏与业务规则剔除闸机回退刷、刷卡重复刷以及出租车漂移点，并以速度阈值和停留时长对轨迹进行切分校核。针对缺失值，时序缺口借助相邻时间段插值与日型模板重建，空间缺损依靠邻近站点或网格扩散弥补，单车缺失还车以最近合法停放点回推，出租车载客状态缺标依据速度与停靠行为补全。时空对齐方面，把多源数据重采样到统一的 5 min 或 10 min 时间窗,空间上汇聚到既定交通小区，并把闸机进出站转化为区级出入强度，刷卡数据在线路站间转移概率约束下推断下车点，得到区级 OD 序列。标准化与编码环节统一时间戳和时区，投影把经纬度转换为统一坐标，运用 z 分数对计数与速度进行尺度对齐，把线路与站点等类别字段

编码为稀疏索引，开展脱敏，并把气象与节假日标签对齐到同一时间窗。

2.2 基于融合数据的特征提取策略

面向分钟级预测输入的稳定性，本研究把融合后的多源观测转化为时间、空间与跨方式交互三类特征。时间维度围绕统一时间窗刻画小时段、工作日与周末类型以及法定节假日标识，并引入近邻滑动窗口的累计量与环比变化，用于表达短时惯性与突发扰动对需求的影响。空间维度以交通小区为载体，统计区内公交上车、地铁进出、共享单车借还与出租车载客事件的强度，同时把一阶相邻小区的总量与差值纳入以反映空间溢出；对出发地至目的地对采用路网最短路径距离与历史平均行程时间构建阻抗特征，补充站点到小区的映射权重与轨道车站节点连通性指标，以呈现网络结构差异，并把方向性引入特征体系，分别记录出发端与到达端的强度与峰位偏移，结合目的地区域的功能混合度来增强吸引与辐射差异的表达。需重点关注的是，多源互补特征用于显式描绘方式间的交互关系：基于刷卡与闸机序列，衡量公交到站后一段时间内相邻地铁站进站量的提升幅度，从而构建公交与地铁的换乘关联强度；围绕轨道站点 300 m 缓冲区，统计共享单车还车与出租车下客在进站高峰前后各 15 min 的量比与时滞分布，用以刻画接驳供给的时空匹配；在中心城区的核心站点，还引入出租车空驶比与接单响应时长的分位数以表征弹性供给状态。为保证跨源量纲一致性，类别变量采用稀疏索引编码，节律性时间变量采用环形嵌入，计数特征在分方式与分区尺度上进行稳健标准化，从而把尺度差异对下游模型训练的干扰压低到可接受水平。

2.3 动态预测模型的设计与实现

鉴于分钟级需求在外生扰动与跨方式耦合下呈现高波动性，本研究构建嵌入注意力机制的 LSTM 融合模型，把 2.2 节形成的时间、空间以及跨方式交互特征统一到序列预测框架中。见图 1，模型自左向右由输入层、注意力层、时序编码层以及输出层组成：输入层接收长度为 L 的多源特征序列，并把小区通道与节律性时间嵌入、外生事件嵌入进行拼接；注意力层在时间维和空间通道维生成可学习权重，用以凸显换乘增强与接驳不均衡对应的关键片段；时序编码层采用双层 LSTM 并加入残差与层归一化，以稳定长滞后依赖的建模；输出层借助全连接映射得到目标时窗的区级出行需求。训练阶段由滑动窗口构造样本，选用 Adam 优化，学习率 1e-3 并配合余弦退火，批大小 128，隐藏单元 128，丢弃率 0.2，权重衰减 1e-5，同时设置早停以缓解概念漂移诱发的过拟合。为提高跨城迁移可用性，把参数共享约束在方式通道上，并把路网邻接先验以掩码形式注入注意力权重的归一化过程，从而在不增加推理负担的前提下提高对网络结构差异的适配性。

$$\hat{D}_{t+h}=\sigma\left(W_o\,\mathrm{LSTM}\left(\sum_{s\in\mathcal{S}}\sum_{\tau=t-L+1}^{t}\alpha_{s,\tau}\,x_{s,\tau}\right)+b_o\right)$$

其中，$\hat{D}_{t+h}$ 表示在预测步长 h 下目标小区在 $t+h$ 时刻的出行需求，$x_{s,\tau}$ 为小区 s 在时间 τ 的融合特征向量，$\alpha_{s,\tau}$ 为注意力层生成的权重，$\mathrm{LSTM}(\cdot)$ 为堆叠长短期记忆编码映射，W_o 与 b_o 为输出层参数，$\sigma(\cdot)$ 为非线性激活函数，$\mathcal{S}$ 为相邻小区集合，L 为历史窗口长度。

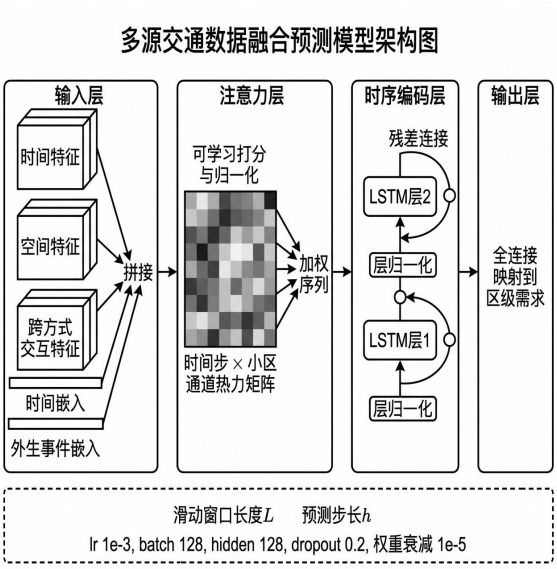

图1 多源交通数据融合预测模型架构图

3 出行需求动态预测模型的效果评估

3.1 实验数据与评估指标设置

结合分钟级运力调度的实际应用场景，本研究把上海市中心城区连续 1 周的多源交通数据构建为统一实验数据集。数据覆盖公交刷卡、地铁闸机、共享单车订单以及出租车 GPS 四类来源，经 2.1 节所述流程完成清洗、对齐与标准化后，按 5 min 时间窗汇聚至交通小区尺度，覆盖工作日 5 天与周末 2 天，形成约 100000 条样本。每条样本由长度为 L 的历史多源特征序列以及对应目标时窗的区级出行需求组成，为契合不同业务时效，把若干预测步长 h 纳入评测范围，并把训练样本借助滑动窗口以步长 1 个时间窗进行构造，使短期与中期依赖得以学习。为降低时间穿越带来的偏差，把样本按自然时间顺序切分为训练集 70%、验证集 20% 与测试集 10%，并把小区划分与路网先验在各子集上保持一致；特征缩放与嵌入等参数仅在训练集上拟合，验证与测试阶段固定使用训练统计量。评估环节把预测值与真实值在测试集上逐时窗逐小区进行对比，选用平均绝对误差、均方根误差以及平均绝对百分比误差作为综合指标，前者关注偏差的平均幅度，中者对大误差更为敏感，后者用于

反映相对误差分布；当目标值接近零时，对分母设置极小阈值并进行掩码处理，以减弱异常零需求对 MAPE 的扰动。为体现面向不同运行时段的适配性，把指标在全局层面采用微平均进行汇总，并在早高峰、晚高峰与平峰三个时段上进行分层统计以供后续分析使用；训练采用早停策略由验证集触发终止，测试阶段在固定参数下开展滚动预测与评估，从而把模型在未见样本上的泛化表现稳定呈现。

3.2 模型预测效果的对比分析

鉴于分钟级运力调度对误差收敛速度与跨时段稳定性的需求，本研究把融合注意力 LSTM 与 ARIMA 以及单一 LSTM 在测试集上进行对比评估，指标选用 MAE、RMSE 与 MAPE 的微平均统计。表 1 给出的综合表现显示，融合模型在三项误差上均呈现更低水平，其中 MAE 为 0.85，相比 ARIMA 降低 31%，相对单一 LSTM 降低 13%，RMSE 与 MAPE 亦在相同方向上同步下降。从时段适配性的角度来看，该优势在早晚高峰的波谷切换以及突发扰动窗口内保持稳定，平峰阶段的误差波动进一步收窄，体现出对短期惯性与外生冲击的同时表征能力。需重点关注的是，性能差异可归因于两类机制：一方面，多源特征把轨道与公交的基底强度叠加出租车与共享单车的弹性供给，进一步配合节律与事件嵌入，使短时波动与结构变化得以同时建模；另一方面，注意力层在路网邻接先验约束下强化了相邻小区的协同权重分配，降低了换乘集聚带来的过拟合偏差。由此推导，在 5、10 与 15 min 等不同预测步长设置下，三类模型的相对排序保持一致，融合模型在短步长上的改善更为突出，在长步长上仍能维持优势；空间分布层面，中心城核心站点的改善幅度高于边缘小区，稀疏供给区域的 MAPE 回落较为有限，呈现数据密度对颗粒度预测的约束。结合 3.1 节的数据切分与特征缩放流程，验证与测试阶段固定训练统计量的设置弱化了信息泄露可能性，因而更能反映结构性差异带来的泛化收益。

表1　不同预测模型效果对比表

模型	MAE	RMSE	MAPE
ARIMA	1.23	2.38	18.1%
单一 LSTM	0.98	1.93	14.5%
融合注意力 LSTM	0.85	1.72	12.6%

4 结语

本文围绕分钟级出行需求预测构建了多源融合的完整方案，包括面向公交刷卡、地铁闸机、共享单车与出租车轨迹的清洗、时空对齐与标准化流程，设计覆盖时间、空间与跨方式交互的特征体系，并在路网先验约束下实现注意力加权的 LSTM 时序编码。以上海中心城区一周数据的实证表明，融合模型在 5 分钟粒度下显著优于基线，微平均 MAE 0.85、RMSE 1.72、MAPE 12.6%，相较 ARIMA 与单一 LSTM 均有明显下降，且在高峰切换与扰动窗口保持稳定，核心站点收益更大。性能提升源于多源互补对基底强度与弹性供给的合成表达，以及邻接掩码引导下的空间协同建模。尽管如此，稀疏供给区域的相对误差回落有限，长步长预测的优势有所收敛，在线流式部署与机制可解释性仍需加强。未来工作将探索基于在线学习的概念漂移自适应、联邦与隐私保护训练、结合图结构与稀疏注意力的统一时空编码、概率预测与不确定性评估，以及接驳与事件感知的联合优化，以进一步提升广域推广与实时应用价值。

参考文献

[1]黄海平,陈亚杰,周超然,常舒予.基于深度学习的交通预测：多技术融合的策略，挑战与未来愿景[J].南京邮电大学学报(自然科学版),

[2]李淑庆,张沛喆.基于LC-Transformer模型的高速公路短时交通流预测研究[J].工程科学与技术,

[3]李韫辉,李清泉,涂伟.城市多模式交通的时空可达性分布特征分析[J].测绘地理信息,2025,50(03):125-129.

[4]唐进君,段一鑫,商淑杰,王骋程,陈群.多特征融合时空图卷积的高速公路交通流量预测[J].公路交通科技,2025,42(06):22-31.

[5]李之红,郄堃,王健宇,许晗,陈金政.建成环境影响下的城市轨道交通客流多步短时预测[J].交通运输系统工程与信息,2025,25(01):160-172.

[6]侯越,王甜甜,张鑫,尹杰.多分辨率趋势周期解耦交互的交通流预测[J].浙江大学学报(工学版),

[7]郑创锐,邓秀勤,陈磊.基于解耦自适应动态图卷积的交通预测模型[J].计算机科学,2025,52(S1):684-691.

[8]于江燕,王倩,孟宪静,张瑞敏,耿蕾蕾.基于动态时空Transformer的城市网络流量预测方法[J].计算机工程与应用,

[9]潘义勇,徐家聪,尤逸文,全勇俊.网约车出行需求影响因素多尺度空间异质性分析[J].吉林大学学报(工学版),

作者简介：刘红波（1987—），男，汉族，本科，研究方向为大数据工程。