

时间异质性下共享单车需求的共形预测 ——分组覆盖偏差的分解与校准

奚 宁¹, 张雨祺²

¹上海理工大学管理学院, 上海

²沈阳体育学院管理与新闻传播学院, 辽宁 沈阳

收稿日期: 2026年9月9日; 录用日期: 2026年10月19日; 发布日期: 2026年10月29日

摘 要

共享单车需求具有明显的日内周期和时间异质性, 预测区间的边际覆盖率达到目标水平, 并不意味着不同时段都具有相同的覆盖表现。以首尔共享单车小时需求数据为例, 采用线性回归和梯度提升树作为点预测模型, 在按时间顺序划分的训练集、校准集和测试集上构造90%的分割共形预测区间, 并比较边际覆盖率和分组覆盖率。结果显示, 两类模型的边际覆盖率分别为0.974和0.930, 均达到目标水平, 但梯度提升树在早、晚高峰的分组覆盖率仅为0.800和0.767。为解释这种现象, 将用于判定分组是否欠覆盖的相对尺度分解为分组尺度成分和尺度错配成分。结果表明, 某一时段是否出现欠覆盖, 不仅与该时段误差尺度相对测试集整体的高低有关, 也与校准集和测试集之间的整体误差尺度错配有关。进一步采用分组共形校准后, 梯度提升树早高峰覆盖率由0.800提高到0.907, 而晚高峰仅提高到0.806, 两个高峰时段的改善幅度与校准集中各时段的误差尺度差异能否反映测试集中的相应差异有关。上述分析为解释边际覆盖达标而部分时段仍出现欠覆盖的现象提供了一个简单框架, 也可作为计量经济学和时间序列课程中由点预测拓展到区间预测与不确定性量化的教学案例。

关键词

共形预测, 共享单车需求, 边际覆盖率, 分组覆盖率, 时间异质性, 不确定性量化

Conformal Prediction for Bike-Sharing Demand under Temporal Heterogeneity —Decomposition and Calibration of Group Coverage Deviations

Ning Xi¹, Yuqi Zhang²

¹Business School, University of Shanghai for Science and Technology, Shanghai

²School of Management and Journalism, Shenyang Sport University, Shenyang Liaoning

Received: September 9, 2026; accepted: October 19, 2026; published: October 29, 2026

Abstract

Bike-sharing demand exhibits pronounced intraday patterns and temporal heterogeneity. Achieving the target marginal coverage rate for prediction intervals does not necessarily imply comparable coverage across different time periods. Using hourly bike-sharing demand data from Seoul, linear regression and gradient boosting are employed as point prediction models, and 90% split conformal prediction intervals are constructed using chronologically ordered training, calibration, and test sets. The marginal coverage rates of the two models are 0.974 and 0.930, respectively, both reaching the target level, whereas the group coverage rates of the gradient boosting model are only 0.800 and 0.767 during the morning and evening peak periods. To explain this phenomenon, the relative scale used to determine whether a group is undercovered is decomposed into a group-specific scale component and a scale-mismatch component. The results show that whether a given period is undercovered depends not only on its error scale relative to that of the overall test set, but also on the mismatch in overall error scale between the calibration and test sets. After applying group-wise conformal calibration, the morning-peak coverage rate of the gradient boosting model increases from 0.800 to 0.907, whereas the evening-peak coverage rate rises only to 0.806. The extent of improvement in each peak period is related to whether the between-group differences in error scale observed in the calibration set reflect the corresponding differences in the test set. These findings provide a simple framework for explaining why marginal coverage can reach the target while some time periods remain undercovered, and they also offer a teaching case for extending econometrics and time-series courses from point prediction to interval prediction and uncertainty quantification.

Keywords

Conformal Prediction, Bike-Sharing Demand, Marginal Coverage, Group Coverage, Temporal Heterogeneity, Uncertainty Quantification

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

共享单车的运营需要根据不同时段的需求合理安排车辆投放和调度。需求预测不准确, 可能导致部分时段车辆供给不足, 也可能造成车辆闲置并增加调度成本。已有研究利用多种机器学习回归模型对共享单车需求进行预测, 并采用平均绝对误差(MAE)、均方根误差(RMSE)和决定系数(R^2)等指标评价点预测效果[1]。然而, 点预测只能给出单一预测值, 难以反映预测结果的不确定性, 因此还需要通过预测区间刻画这种不确定性。传统预测区间通常依赖误差分布或模型形式等假设。共形预测则不同: 在校准样本与新观测满足可交换性的条件下, 它可以在已有点预测模型的基础上构造预测区间, 而不必预先设定具体的误差分布, 并提供有限样本下的边际覆盖保证。这一方法对点预测模型的形式没有特殊要求, 线性模型、树模型等均可使用[2]-[5]。

将共形预测应用于首尔共享单车小时需求数据后发现, 以线性回归和梯度提升树为点预测模型构造的 90%分割共形预测区间, 其边际覆盖率均达到目标水平, 但梯度提升树在早、晚高峰的分组覆盖率均不高于 0.800。这表明, 边际覆盖保证针对的是总体上的覆盖概率, 而不是各组的覆盖率; 仅依据边际覆盖率评价预测区间, 可能会掩盖特定时段的覆盖不足[6]。针对这类问题, 已有研究提出了不同的改进方

法。对于预先划分的不同时段或不同组, 可以采用 **Mondrian** 共形预测等方法在组内分别进行校准, 以适应不同组之间的误差尺度异质性[2]; 对于误差分布随时间发生变化的情形, 则可以采用自适应共形预测、动态调整等方法, 使预测区间随新的误差信息进行更新[7]-[9]。这两类方法针对的数据特征并不相同, 因此, 在选择具体调整方法之前, 需要先判断这里的分组欠覆盖在多大程度上与组间误差尺度差异有关, 以及校准集与测试集之间整体误差尺度的变化发挥了什么作用。

为此, 基于同一数据与模型, 将用于判定分组是否欠覆盖的相对尺度分解为分组尺度成分和尺度错配成分。分组尺度成分反映测试集内部某一时段的误差尺度相对于整体的高低, 用于刻画不同时段之间的误差尺度异质性; 尺度错配成分反映校准集与测试集之间整体误差尺度的差异。由此, 某一组是否出现欠覆盖, 可以转化为其分组尺度成分是否超过由尺度错配成分决定的门槛。分解结果显示, 不同时段的误差尺度存在明显差异, 早、晚高峰的分组尺度成分均处于较高水平, 表明时段误差尺度异质性是分组欠覆盖的重要特征; 同时, 尺度错配程度也会改变各组出现欠覆盖的判定门槛。例如, 早高峰中两类模型的分组尺度成分十分接近, 但尺度错配程度不同, 因此呈现出不同的分组覆盖结果。基于对时段误差尺度异质性的识别, 进一步采用分组共形校准, 在各运营时段内分别确定校准分位数[2]。结果显示, 分组校准能够改善部分高峰时段的欠覆盖, 但其改善幅度与校准集中各时段之间的误差尺度差异能否较好反映测试集中的相应差异有关。对于校准集与测试集之间的整体误差尺度错配, 这里主要进行识别和解释, 不作动态调整。

2. 共形预测方法与覆盖率偏差分解

分割共形预测使用由校准集确定的统一校准分位数为测试集中的观测构造预测区间, 其有限样本保证针对边际覆盖率。当预测误差在不同组之间存在异质性, 或者校准集与测试集之间的整体误差尺度存在差异时, 边际覆盖率与分组覆盖率可能表现不同。为分析这种差异, 以下首先介绍分割共形预测及其覆盖率指标, 再构造用于判定分组是否欠覆盖的相对尺度, 并将其分解为分组尺度成分和尺度错配成分。

2.1. 分割共形预测

设观测样本为 $\{(X_i, Y_i)\}_{i=1}^n$, 其中, X_i 为第 i 个观测的特征向量, Y_i 为相应的响应变量。给定特征 X , 点预测模型可表示为:

$$\hat{Y} = \hat{f}(X) \quad (1)$$

预测区间在点预测值的基础上给出真实响应值可能落入的范围。记预测区间为:

$$C(X) = [L(X), U(X)] \quad (2)$$

其中 $L(X)$ 和 $U(X)$ 分别为区间的下界和上界。设 (X_{new}, Y_{new}) 为一个新的观测, 给定显著性水平 α , 所构造的预测区间应满足:

$$P\{Y_{new} \in C(X_{new})\} \geq 1 - \alpha \quad (3)$$

其中, $1 - \alpha$ 为目标边际覆盖率。

将样本划分为三个互不重叠的子集: 训练集 D_{train} 、校准集 D_{cal} 和测试集 D_{test} , 样本量分别为 n_{train} 、 n_{cal} 和 n_{test} , 且三者之和为 n 。训练集用于估计点预测模型, 校准集用于确定预测区间的宽度, 测试集用于评价预测表现。训练得到的点预测模型记为:

$$\hat{f} = \mathcal{A}(D_{train}) \quad (4)$$

其中, $\mathcal{A}$ 表示学习算法。对于回归问题, 可取绝对残差作为不符合度分数:

$$S_i = |Y_i - \hat{f}(X_i)|, (X_i, Y_i) \in \mathcal{D}_{cal} \quad (5)$$

不符合度分数越大, 表示预测误差越大。设校准集共有 n_{cal} 个不符合度分数, 将其从小到大排序:

$$S_{(1)} \leq S_{(2)} \leq \cdots \leq S_{(n_{cal})} \quad (6)$$

给定显著性水平 α , 取:

$$k = (n_{cal} + 1)(1 - \alpha) \quad (7)$$

其中, $\lceil \cdot \rceil$ 表示向上取整。令:

$$q_c = S_{(k)} \quad (8)$$

q_c 是校准集绝对残差的第 k 个次序统计量, 用于确定预测区间的宽度。对于测试集中的第 i 个观测, 其特征向量记为 X_i , 首先得到点预测值:

$$\hat{Y}_i = \hat{f}(X_i) \quad (9)$$

再利用 q_c 构造预测区间:

$$C(X_i) = [\hat{f}(X_i) - q_c, \hat{f}(X_i) + q_c] \quad (10)$$

在上述分割共形预测中, 测试集各观测均使用同一个校准分位数 q_c 。当响应变量非负时, 可将区间下界截断至 0。若校准样本与新观测满足可交换性, 则在有限样本下, 新观测的真实响应值落入预测区间的概率至少为 $1 - \alpha$ 。这一保证针对边际覆盖率, 并不意味着各组的覆盖率均能达到目标水平[6]。因此, 评价预测区间不能只考察边际覆盖率, 还需要考察分组覆盖率。

2.2. 边际覆盖率与分组覆盖率

区间预测首先考察测试集中真实响应值落入预测区间的比例, 即经验边际覆盖率, 记为 $\hat{\pi}$:

$$\hat{\pi} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} I\{Y_i \in C(X_i)\} \quad (11)$$

其中, $I(\cdot)$ 为示性函数。式(3)给出理论上的边际覆盖率, $\hat{\pi}$ 是其在测试集上的经验估计, 下文简称为边际覆盖率。若 $\hat{\pi} < 1 - \alpha$, 则称为边际欠覆盖。

边际覆盖率反映测试集整体的覆盖表现, 但不能说明不同组的覆盖情况, 因此需要进一步计算各组的经验覆盖率。设测试集中第 g 组的观测集合为 G_g , 观测数为 n_g , 其经验分组覆盖率记为 $\hat{\pi}_g$:

$$\hat{\pi}_g = \frac{1}{n_g} \sum_{(X_i, Y_i) \in G_g} I\{Y_i \in C(X_i)\} \quad (12)$$

下文简称为分组覆盖率。若 $\hat{\pi}_g < 1 - \alpha$, 则称第 g 组出现欠覆盖。各组观测数不同, 分组覆盖率的抽样波动也不同, 因此后文主要关注明显偏离目标水平的情形。

在不截断区间下界时, 预测区间宽度为 $2q_c$; 若对非负响应变量将下界截断至 0, 不同观测的实际区间宽度可能有所不同。平均区间宽度记为 $\bar{W}$:

$$\bar{W} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} [U(X_i) - L(X_i)] \quad (13)$$

平均区间宽度反映预测区间的宽窄程度。由于覆盖率与区间宽度之间存在权衡, 比较不同预测区间构造方法时, 应先考察边际覆盖率和分组覆盖率, 在覆盖水平具有可比性时再比较平均区间宽度。

2.3. 覆盖率偏差分解

将测试集全体观测和第 g 组观测绝对残差的经验 $(1-\alpha)$ 分位数分别记为 q_t 和 $q_{t,g}$ 。具体地, q_t 和 $q_{t,g}$ 分别取第 $\lceil n_{test}(1-\alpha) \rceil$ 和第 $\lceil n_g(1-\alpha) \rceil$ 个次序统计量。这里不采用式(7)的共形位次, 以使下文式(15)成为严格的等价关系。二者仅用于事后诊断, 不参与预测区间的构造。

定义第 g 组的相对尺度为:

$$R_g = \frac{q_{t,g}}{q_c} \quad (14)$$

由于 $\hat{\pi}_g \geq 1-\alpha$ 等价于第 g 组中至少有 $n_g(1-\alpha)$ 个绝对残差不超过 q_c , 而按定义, $q_{t,g}$ 为该组绝对残差的第 $n_g(1-\alpha)$ 个次序统计量, 因此

$$\hat{\pi}_g \geq 1-\alpha \Leftrightarrow q_{t,g} \leq q_c \Leftrightarrow R_g \leq 1 \quad (15)$$

式(15)表明, R_g 可以直接用于判定第 g 组的覆盖率是否达到目标水平。进一步将 R_g 分解为:

$$R_g = \frac{q_{t,g}}{q_t} \cdot \frac{q_t}{q_c} = H_g \Delta \quad (16)$$

其中, $H_g = q_{t,g}/q_t$ 为分组尺度成分, 反映第 g 组误差尺度相对于测试集整体的高低, 各组 H_g 的差异即测试集内部的误差尺度异质性; $\Delta = q_t/q_c$ 为尺度错配成分, 反映校准集与测试集之间整体误差尺度的差异。

将整个测试集视为一个组, 由式(15)及 q_t 的定义可得:

$$\hat{\pi} \geq 1-\alpha \Leftrightarrow q_t \leq q_c \Leftrightarrow \Delta \leq 1 \quad (17)$$

因此, 边际覆盖率达到目标水平当且仅当 $\Delta \leq 1$; 第 g 组的覆盖率达到目标水平当且仅当 $H_g \Delta \leq 1$, 即 $H_g \leq 1/\Delta$ 。各组是否达到目标水平由 H_g 和 Δ 共同决定。当 $\Delta \leq 1$ 时, 边际覆盖率达到目标水平, 但若某一组的 $H_g > 1/\Delta$, 该组仍会出现欠覆盖; 当 $\Delta > 1$ 时, 测试集出现边际欠覆盖, 具体哪些组欠覆盖仍取决于各组 H_g 与判定门槛 $1/\Delta$ 的相对大小。若各组共同覆盖全部测试集且互不重叠, 边际覆盖率就是各组覆盖率按观测数加权后的平均值, 因此不同组的过覆盖和欠覆盖可能相互抵消。这也是边际覆盖率达到或接近目标水平时, 部分组仍可能出现明显欠覆盖的原因。对非负响应变量, 区间下界截断至 0 只会减小区间宽度, 而被截去的部分不可能包含真实响应值, 因此上述判定关系均不受影响。

上述分解将分组欠覆盖与分组尺度成分 H_g 和尺度错配成分 Δ 联系起来。当不同组之间存在明显的误差尺度异质性时, 可以进一步采用分组校准。在这一分解框架下, 分组校准调整各组的校准分位数, 而 H_g 和 Δ 本身不变。

3. 实证结果与分析

首先介绍数据与实证设计, 给出两类模型的点预测表现、校准集上的预测误差及相应的校准分位数; 随后考察边际覆盖率与分组覆盖率, 利用覆盖率偏差分解解释不同组的覆盖差异, 并进一步分析分组共形校准的覆盖效果及其影响因素。

3.1. 数据与实证设计

数据来自 UCI Machine Learning Repository 发布的 Seoul Bike Sharing Demand 数据集[10], 共包含 8760 个逐小时观测, 时间范围为 2017 年 12 月 1 日至 2018 年 11 月 30 日。数据中的 Functioning Day 用于标记共享单车系统是否正常运行。Functioning Day = No 的 295 个观测中, 租赁数量均为 0。由于这些

记录对应系统非正常运营状态, 与正常运营条件下的需求预测问题不同, 予以剔除。最终样本包含 8465 个观测。

被解释变量为每小时租赁数量。解释变量包括时刻、温度、湿度、风速、能见度、露点温度、太阳辐射、降雨量、降雪量和节假日。其中, 时刻和节假日作为分类变量处理, 时刻按 0~23 时划分为 24 个类别, 以反映需求的日内周期变化。主要变量的描述统计见表 1。

Table 1. Summary statistics of main variables in functioning observations
表 1. 正常运营观测主要变量描述统计

变量	含义	均值	标准差	最小值	中位数	最大值
Rented Bike Count	每小时租赁数量	729.157	642.351	2.000	542.000	3556.000
Temperature	温度	12.771	12.104	-17.800	13.500	39.400
Humidity	湿度	58.147	20.485	0.000	57.000	98.000
Wind speed	风速	1.726	1.034	0.000	1.500	7.400
Visibility	能见度	1433.873	609.051	27.000	1690.000	2000.000
Dew point temperature	露点温度	3.945	13.242	-30.600	4.700	27.200
Solar Radiation	太阳辐射	0.568	0.868	0.000	0.010	3.520
Rainfall	降雨量	0.149	1.126	0.000	0.000	35.000
Snowfall	降雪量	0.078	0.444	0.000	0.000	8.800

注: 各变量观测数均为 8465。时刻按 0~23 时划分为 24 个类别, 节假日为二值分类变量, 二者均作为分类变量进入模型, 因此不列入表中。

为保持样本的时间顺序, 并避免同一天的观测被划入不同子集, 样本按完整日期进行划分。训练集、校准集和测试集的日期数量依次约占 60%、20%和 20%。353 个正常运营日期中, 前 211 个日期作为训练集, 随后 70 个日期作为校准集, 其余 72 个日期作为测试集。三个子集的日期范围、观测数和季节构成见表 2。训练集包含冬季、春季和夏季观测, 校准集以夏季为主并包含部分秋季观测, 测试集则全部属于秋季, 校准集与测试集的季节构成存在明显差异。季节变量不纳入模型, 主要基于两点考虑。一方面, 温度、露点温度和太阳辐射等气象变量已经包含较多季节信息; 另一方面, 训练集中没有秋季观测, 加入季节分类变量会使测试集中出现训练集中未见的类别。为考察不同日内时段的覆盖表现, 预先将每天划分为 00~05、06~08、09~16、17~20 和 21~23 五个时段, 其中 06~08 和 17~20 分别作为早、晚高峰时段。下文所称分组均指上述五个时段。实证分析统一取 $\alpha = 0.10$, 即目标覆盖率设为 90%; 后文实证结果均基于上述样本划分和时段分组。

Table 2. Time span, number of observations and seasonal composition of the three subsets
表 2. 三个子集的日期范围、观测数与季节构成

子集	日期范围	观测数	季节构成
训练集	2017-12-01~2018-07-01	5064	冬 2160、春 2160、夏 744
校准集	2018-07-02~2018-09-09	1680	夏 1464、秋 216
测试集	2018-09-10~2018-11-30	1721	秋 1721

3.2. 点预测结果与校准分位数

采用线性回归(LR)和梯度提升树(GB)作为点预测模型。LR 作为线性基准, 并通过时刻虚拟变量刻画日内周期变化; GB 用于刻画变量之间可能存在的非线性关系。同时采用线性模型和非线性模型, 是为了

考察后续覆盖表现是否随点预测模型而变化，而非比较两类模型的优劣。

点预测表现采用平均绝对误差(MAE)、均方根误差(RMSE)和决定系数(R^2)衡量，均在测试集上计算，结果见表 3。LR 的 MAE、RMSE 和 R^2 分别为 289.952、385.383 和 0.593，GB 的相应指标分别为 311.356、423.089 和 0.509。LR 的 MAE 和 RMSE 均低于 GB， R^2 则高于 GB。上述结果仅反映该测试集上的点预测表现，不能据此认为 LR 的预测能力整体优于 GB。

Table 3. Test-set point prediction performance of the two models
表 3. 两种点预测模型的测试集结果

模型	MAE	RMSE	R^2
线性回归(LR)	289.952	385.383	0.593
梯度提升树(GB)	311.356	423.089	0.509

分割共形预测的预测区间由校准集绝对残差所确定的校准分位数 q_c 构造，而不是由测试集的点预测误差决定。表 4 比较了两类模型在校准集和测试集上的平均绝对误差，并列出了由校准集得到的校准分位数 q_c 。在校准集上，LR 和 GB 的 MAE 分别为 523.3 和 401.6，GB 较低；相应的 q_c 分别为 974.0 和 819.2。进入测试集后，LR 的 MAE 降至 290.0，GB 的 MAE 为 311.4。可见，LR 在测试集上的 MAE 低于 GB，其校准分位数却明显更大。校准分位数由校准集绝对残差的分布确定，与测试集的误差尺度未必一致，这一差异将在 3.4 节中通过尺度错配成分加以刻画。

Table 4. Prediction errors in the calibration and test sets and calibration quantiles
表 4. 校准集与测试集的预测误差及校准分位数

模型	平均绝对误差(MAE)		校准分位数 q_c
	校准集	测试集	
LR	523.3	290.0	974.0
GB	401.6	311.4	819.2

注：测试集平均绝对误差与表 3 一致，此处保留一位小数。

3.3. 边际覆盖率与分组覆盖率的实证比较

表 5 给出了当 $\alpha = 0.10$ (目标覆盖率 90%)时，分割共形预测在测试集上的结果，包括校准分位数 q_c 、边际覆盖率、五个时段的分组覆盖率和平均区间宽度。由于租赁数量非负，预测区间下界均截断至 0。

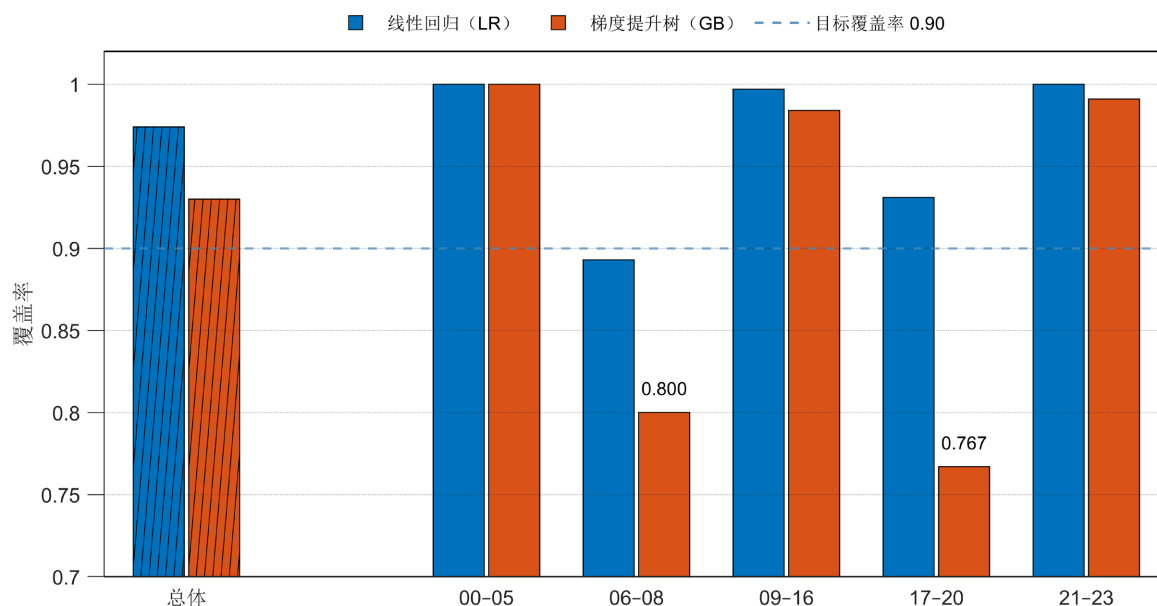
Table 5. Marginal and group coverage rates of split conformal prediction on the test set (90% target coverage)
表 5. 测试集的边际覆盖率与分组覆盖率(目标覆盖率 90%)

模型	校准分位数 q_c	边际覆盖率 $\hat{\pi}$	分组覆盖率 $\hat{\pi}_g$					平均区间宽度 $\bar{w}$
			00~05 ($n = 426$)	06~08 ($n = 215$)	09~16 ($n = 576$)	17~20 ($n = 288$)	21~23 ($n = 216$)	
LR	974.0	0.974	1.000	0.893	0.997	0.931	1.000	1663.8
GB	819.2	0.930	1.000	0.800	0.984	0.767	0.991	1375.0

注：截断前区间宽度恒为 $2q_c$ ，对于 LR 和 GB 模型，该宽度分别为 1948.1 和 1638.3。由于下界截断会减少部分观测的区间宽度，测试集平均区间宽度低于该值。

从边际覆盖率看，两类模型均达到目标水平，LR 和 GB 分别为 0.974 和 0.930。分组覆盖率则表现出

明显差异。LR 各组覆盖率介于 0.893 和 1.000 之间, 其中 06~08 时段为 0.893, 略低于目标水平; 其余各组均达到目标水平。GB 在 06~08 和 17~20 时段的覆盖率分别为 0.800 和 0.767, 明显低于 0.90, 而其余三个时段均在 0.984 以上。由于各组覆盖率按观测数加权后得到边际覆盖率, 其他时段较高的覆盖率抵消了早、晚高峰的欠覆盖, 使 GB 的边际覆盖率仍达到目标水平。图 1 对不同模型的边际覆盖率与分组覆盖率进行了直观比较。



注: 最左一组为全体测试观测, 对应边际覆盖率; 虚线表示 0.90 的目标覆盖率; 纵轴自 0.70 起。

Figure 1. Comparison of marginal and group coverage rates in the test set

图 1. 测试集的边际覆盖率与分组覆盖率比较

需要注意的是, 较高的分组覆盖率并不一定意味着预测区间更加理想。统一校准分位数 q_c 由全体校准样本的预测误差确定, 因此当不同组的误差尺度存在差异时, 统一尺度可能导致部分组区间相对过宽而产生过覆盖, 而另一部分组区间不足而产生欠覆盖。由此可见, 过覆盖和欠覆盖均反映了统一尺度校准在面对组间误差异质性时的局限。

进一步比较两类模型可以发现, LR 的校准分位数和平均区间宽度均大于 GB, 其边际覆盖率也相对更高。值得注意的是, LR 在测试集上的预测误差小于 GB (290.0 对 311.4), 校准分位数却反而更大 (974.0 对 819.2)。这一反差说明, 分组覆盖表现不仅受到组间误差尺度差异的影响, 也与校准集和测试集之间的整体误差尺度变化有关。下一节利用 2.3 节的覆盖率偏差分解, 分别分析分组尺度成分与尺度错配成分的作用。

3.4. 覆盖率偏差的实证分解

3.3 节表明, 两类模型的边际覆盖率均达到 0.90 的目标水平, 但分组覆盖表现存在明显差异, 其中 GB 模型在早、晚高峰时段存在明显欠覆盖。为进一步分析这种差异, 表 6 根据式(14)~(16)列出尺度错配成分 Δ 、判定门槛 $1/\Delta$, 以及早、晚高峰的分组尺度成分 H_g 和相对尺度 R_g 。由式(15)可知, 第 g 组覆盖率达到目标水平当且仅当 $R_g \leq 1$; 因此, 当 $R_g > 1$ 时, 第 g 组出现欠覆盖。结合式(16), 这一条件等价于 $H_g > 1/\Delta$ 。

从表 6 可以看出, 两类模型在早、晚高峰时段的 H_g 均高于 1, 说明在相应模型下, 这些时段的误差

尺度高于该模型测试集整体水平。但判定门槛 $1/\Delta$ 由各模型自身的尺度错配成分决定, 因此两类模型的判定门槛并不相同。早高峰中, LR 和 GB 的 H_g 分别为 1.56 和 1.53, 数值十分接近; 但由于两类模型的 Δ 分别为 0.650 和 0.857, 对应判定门槛 $1/\Delta$ 分别为 1.54 和 1.17。LR 的 H_g 略高于判定门槛, 因此出现轻微欠覆盖, 覆盖率为 0.893, 仅略低于目标水平; GB 的 H_g 明显高于判定门槛, 因此出现更明显的欠覆盖, 覆盖率低至 0.800。这表明, 早高峰中两类模型的分组尺度成分较为接近, 但尺度错配成分不同, 对应的判定门槛也不同, 因而呈现出不同的覆盖表现。晚高峰中, 分组尺度成分和尺度错配成分的差异同时存在。LR 的 H_g 为 1.40, 低于 1.54 的判定门槛, 因此覆盖率达到目标水平, 为 0.931; GB 的 H_g 不仅高于 LR, 而且达到 1.60, 超过 1.17 的判定门槛, 因此出现欠覆盖, 覆盖率为 0.767。由此可见, 分组覆盖结果并不能仅由某一时段误差尺度的高低解释, 而取决于分组尺度成分 H_g 与尺度错配成分 Δ 的共同作用。

Table 6. Decomposition of coverage deviations

表 6. 覆盖率偏差分解

模型	尺度错配成分 Δ	判定门槛 $1/\Delta$	06~08		17~20	
			分组尺度成分 H_g	相对尺度 R_g	分组尺度成分 H_g	相对尺度 R_g
LR	0.650	1.54	1.56	1.02	1.40	0.91
GB	0.857	1.17	1.53	1.31	1.60	1.37

注: R_g 根据未舍入的分位数计算, 与表中 H_g 和 Δ 的显示值直接相乘可能存在末位差异。

上述结果表明, 测试集内部不同时段之间存在明显的误差尺度异质性。针对这种异质性, 可以在校准集中按时段分别确定校准分位数, 使不同时段不再使用同一个校准分位数。需要注意的是, 这种调整针对的是时段之间的差异, 而校准集与测试集之间的整体误差尺度错配仍然存在。下一节进一步考察分组共形校准对各时段覆盖表现的改善情况。

3.5. 分组共形校准及其覆盖效果

对于第 g 组, 按照式(7)的共形位次在组内计算校准分位数 $q_{c,g}$, 并对属于该组的测试观测构造预测区间

$$C_g(X_i) = [\hat{f}(X_i) - q_{c,g}, \hat{f}(X_i) + q_{c,g}] \quad (18)$$

对于非负响应变量仍将区间下界截断至 0。为比较分组校准与统一校准之间的差异, 定义第 g 组的分组校准因子为

$$A_g = \frac{q_{c,g}}{q_c} \quad (19)$$

其中 q_c 为统一校准时使用的校准分位数。 $A_g > 1$ 表示该组使用的校准分位数大于统一校准分位数, $A_g < 1$ 表示该组的校准分位数小于统一校准分位数。测试集的分组尺度成分 H_g 和尺度错配成分 Δ 仍沿用 2.3 节的定义。于是, 分组校准后第 g 组的相对尺度记为 $R_g^{(G)}$, 可写为

$$R_g^{(G)} = \frac{q_{t,g}}{q_{c,g}} = \frac{q_{t,g}}{q_t} \cdot \frac{q_t}{q_c} \cdot \frac{q_c}{q_{c,g}} = \frac{H_g}{A_g} \Delta \quad (20)$$

将式(15)中的 q_c 换为 $q_{c,g}$, 可得分组校准后第 g 组覆盖率达到目标水平当且仅当 $R_g^{(G)} \leq 1$, 即

$H_g/A_g \leq 1/\Delta$ 。因此,与统一校准下比较 H_g 和 $1/\Delta$ 不同,分组校准后需要比较 H_g/A_g 与同一判定门槛 $1/\Delta$ 。在上述定义下,分组校准不改变 H_g 和 Δ 的取值,而是通过 A_g 调整各组使用的校准分位数。统一校准与分组校准的覆盖率和平均区间宽度比较见表 7。

GB 的结果可以用式(20)直接解释。早高峰的 A_g 为 1.32,说明分组校准将该时段未截断前的区间半宽提高到统一校准的 1.32 倍。测试集的 H_g 为 1.53,因此 $H_g/A_g = 1.15$, 低于 GB 的判定门槛 1.17。相应地,校准分位数由 819.2 提高到 1081.8,早高峰覆盖率由 0.800 升至 0.907,达到目标水平。晚高峰则不同, A_g 仅为 1.08,而测试集的 H_g 达到 1.60,因而 $H_g/A_g = 1.49$, 仍明显高于 1.17 的门槛。虽然分组校准扩大了晚高峰的预测区间,覆盖率由 0.767 提高到 0.806,但仍低于目标水平。

LR 呈现出另一种情况。其早高峰的 A_g 为 1.00,说明校准集中该时段的校准分位数与统一校准分位数基本相同,而测试集的 H_g 达到 1.56,因而 $H_g/A_g = 1.57$, 略高于 LR 的判定门槛 1.54。分组校准几乎没有改变该时段的校准分位数,覆盖率仍为 0.893。21~23 时段则相反, A_g 只有 0.69,分组校准相应收窄了该时段的区间,覆盖率由 1.000 降至 0.968,同时平均区间宽度由 1792 降至 1313。这说明,分组校准既可能扩大某些时段的区间,也可能收窄另一些时段的区间;各时段在分组校准后是否达到目标覆盖率,可由 H_g/A_g 与判定门槛 $1/\Delta$ 的相对关系进行判断。

Table 7. Comparison of pooled and group-wise conformal calibration (90% target coverage)
表 7. 统一校准与分组校准的比较(目标覆盖率 90%)

模型	时段	分组校准因子 A_g	分组尺度成分 H_g	H_g/A_g	覆盖率		平均区间宽度	
					统一校准	分组校准	统一校准	分组校准
LR	00~05	0.83	0.69	0.83	1.000	1.000	1386	1208
LR	06~08	1.00	1.56	1.57	0.893	0.893	1583	1581
LR	09~16	1.11	0.75	0.67	0.997	1.000	1729	1864
LR	17~20	1.04	1.40	1.34	0.931	0.938	1908	1980
LR	21~23	0.69	0.81	1.18	1.000	0.968	1792	1313
LR	边际	—	—	—	0.974	0.972	1664	1616
GB	00~05	1.05	0.49	0.47	1.000	1.000	1136	1178
GB	06~08	1.32	1.53	1.15	0.800	0.907	1284	1590
GB	09~16	0.88	0.72	0.81	0.984	0.976	1487	1337
GB	17~20	1.08	1.60	1.49	0.767	0.806	1521	1621
GB	21~23	0.67	0.76	1.13	0.991	0.912	1442	1040
GB	边际	—	—	—	0.930	0.937	1375	1340

注: 分组校准下, 各组截断前的区间宽度为 $2q_{c,g}$; H_g/A_g 根据未舍入的各分位数计算, 与表中 H_g 和 A_g 两列保留两位小数的数值直接相除可能存在末位差异; 两列“平均区间宽度”均为截断后实际预测区间的平均宽度。

从整体结果看, 分组校准并不是通过对所有时段统一加宽区间来改善覆盖表现。GB 的边际覆盖率由 0.930 提高到 0.937, 平均区间宽度反而由 1375 降至 1340。分组校准实际上重新分配了不同时段区间宽度: 早、晚高峰的区间有所加宽, 而部分夜间和日间时段相应收窄。因此, 在提高高峰时段覆盖率的同时, 整体平均区间宽度并未增加。LR 也呈现出区间重新分配的特征, 平均区间宽度由 1664 降至 1616, 边际覆盖率由 0.974 略降至 0.972; 但由于其欠覆盖仅出现在早高峰一组, 且该组的 A_g 接近 1, 分组校准对其覆盖表现的影响很小。

从 GB 早高峰、GB 晚高峰和 LR 早高峰的结果可以看出, 三种情况下分组校准因子 A_g 的调整幅度

存在明显差异。GB 早高峰的组别调整幅度足以使 H_g/A_g 降至判定门槛以下, 而 GB 晚高峰和 LR 早高峰的调整幅度不足, H_g/A_g 仍高于相应判定门槛。由此可见, 分组校准能够利用校准集中观察到的组间差异调整各组的校准分位数, 但其改善程度取决于这种组别调整与测试集分组尺度之间的对应程度。这里的比较主要用于说明上述分解框架如何解释分组校准后的覆盖变化, 并不构成对分组共形预测一般性能的评价。

4. 结论与启示

共享单车需求具有明显的日内周期和时间异质性, 仅考察预测区间的边际覆盖率, 可能掩盖特定时段的覆盖不足。对首尔共享单车小时数据的实证分析表明, 以线性回归和梯度提升树为点预测模型构造的 90% 分割共形预测区间, 其边际覆盖率分别达到 0.974 和 0.930, 但分组覆盖表现存在明显差异。梯度提升树在早、晚高峰的覆盖率分别只有 0.800 和 0.767, 而其余三个时段均在 0.984 以上。覆盖率偏差分解进一步表明, 某一时段是否出现欠覆盖, 不仅与该时段相对于测试集整体的误差尺度有关, 也与校准集和测试集之间的整体误差尺度错配有关。早高峰中, 两类模型的分组尺度成分十分接近, 但由于尺度错配成分不同, 对应的判定门槛也不同, 因而呈现出明显不同的覆盖结果。

在具有时间异质性的预测问题中, 边际覆盖率达到目标水平并不意味着预测区间在各个时段都具有同样的可靠性。因此, 评价预测区间时, 除考察边际覆盖率和平均区间宽度外, 还应检查具有实际意义的分组覆盖率。分组共形校准为改善这类偏差提供了一种简单的处理方式, 其改善程度取决于校准集中得到的组别调整与测试集分组尺度之间的对应程度。实证中, 梯度提升树早高峰的覆盖率经分组校准后由 0.800 提高到 0.907, 而晚高峰仅由 0.767 提高到 0.806, 说明分组校准并不能保证各组均达到目标覆盖率。同时, 梯度提升树在测试集上的总体平均区间宽度由 1375 降至 1340, 表明分组校准主要体现为对不同时段区间宽度的重新分配, 而不是对所有时段统一加宽区间。因此, 判断分组校准的效果, 需要结合分组尺度、尺度错配和组别调整综合分析。

从实际应用看, 上述分解也提供了一个简单的诊断思路。首先考察边际覆盖率和具有实际意义的分组覆盖率; 若存在明显的分组欠覆盖, 可进一步计算分组尺度成分 H_g 和尺度错配成分 Δ 。当不同组的 H_g 差异较为明显时, 可考虑通过分组校准调整各组的校准分位数; 当 Δ 显示校准集与测试集之间存在较明显的整体尺度错配时, 仅进行分组校准可能不足, 还需要考虑滚动校准或自适应共形预测等能够随时间更新校准尺度的方法; 若两类差异同时存在, 则需要同时关注组间异质性和随时间变化的整体误差尺度。由此, 可以按照“检查分组覆盖 - 分解覆盖偏差 - 识别差异来源 - 选择相应校准方式”的顺序进行诊断和调整。

从教学角度看, 这一案例可以作为传统计量经济学和时间序列课程中点预测内容的扩展。线性回归、预测误差以及训练集和测试集划分等都是学生较熟悉的内容, 在此基础上引入校准集和分割共形预测, 可以自然地由点预测拓展到预测区间和不确定性量化。进一步比较边际覆盖率和分组覆盖率, 又能够让学生直观看到边际指标可能掩盖特定时段的覆盖不足, 并理解组间误差尺度异质性以及校准集与测试集之间的尺度错配为什么会影响预测区间的覆盖表现。整个案例只需要基本的回归模型、树模型、残差分位数和简单比例分解, 同时使用公开数据, 便于重复计算和课堂实验, 因而适合作为连接传统计量方法与机器学习中不确定性量化方法的教学案例。

上述分析仍有一定局限。实证结果来自单一共享单车数据集, 训练集、校准集和测试集采用固定的时间顺序划分, 分组方式也预先按照运营时段设定。不同数据、不同时间划分以及其他分组方式下, 覆盖率偏差的表现可能有所不同。覆盖率偏差分解使用测试集经验分位数进行事后诊断, 主要用于解释分组覆盖偏差的来源, 并不直接构成实际预测中的动态调整规则。后续可以在更多数据和不同时间划分下

考察上述分解结果是否稳定。此外, 滚动校准、自适应共形预测等动态方法能否改善分组覆盖表现, 以及不同校准方法在覆盖率和区间宽度之间如何权衡, 也值得进一步研究。

基金项目

本文受上海理工大学 2023 年度本科教学研究与改革重点项目“新文科建设背景下管理科学专业人才培养路径”(项目编号: JGXM202302)资助。

参考文献

- [1] Sathishkumar, V.E. and Cho, Y. (2020) A Rule-Based Model for Seoul Bike Sharing Demand Prediction Using Weather Data. *European Journal of Remote Sensing*, **53**, 166-183. <https://doi.org/10.1080/22797254.2020.1725789>
- [2] Vovk, V., Gammernan, A. and Shafer, G. (2005) Algorithmic Learning in a Random World. Springer.
- [3] Shafer, G. and Vovk, V. (2008) A Tutorial on Conformal Prediction. *Journal of Machine Learning Research*, **9**, 371-421.
- [4] Angelopoulos, A.N. and Bates, S. (2023) Conformal Prediction: A Gentle Introduction. *Foundations and Trends in Machine Learning*, **16**, 494-591. <https://doi.org/10.1561/22000000101>
- [5] Fontana, M., Zeni, G. and Vantini, S. (2023) Conformal Prediction: A Unified Review of Theory and New Challenges. *Bernoulli*, **29**, 1-23. <https://doi.org/10.3150/21-bej1447>
- [6] Barber, R.F., Candès, E.J., Ramdas, A. and Tibshirani, R.J. (2021) The Limits of Distribution-Free Conditional Predictive Inference. *Information and Inference: A Journal of the IMA*, **10**, 455-482. <https://doi.org/10.1093/imaiai/iaaa017>
- [7] Stankeviciute, K., Alaa, A.M. and van der Schaar, M. (2021) Conformal Time-Series Forecasting. *Advances in Neural Information Processing Systems*, **34**, 6216-6228.
- [8] Zaffran, M., Féron, O., Goude, Y., Josse, J. and Dieuleveut, A. (2022) Adaptive Conformal Predictions for Time Series. *2022 Proceedings of the 39th International Conference on Machine Learning, Baltimore*, 17-23 July 2022, 25834-25866.
- [9] Xu, C. and Xie, Y. (2021) Conformal Prediction Interval for Dynamic Time-Series. *2021 Proceedings of the 38th International Conference on Machine Learning, Online*, 18-24 July 2021, 11559-11569.
- [10] UCI Machine Learning Repository (2026) Seoul Bike Sharing Demand. <https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand>