DOI: 10.54254/3049-5377/2025.25916

基于随机参数的居民出行方式选择预测

杨　程[1]，肖绪冠[2]，刘珍珍[1*]

（1.广西计算中心有限责任公司，广西壮族自治区南宁市，530000；

2.桂林电子科技大学，广西壮族自治区桂林市，541000；

* 通讯作者，164829228@qq.com ）

摘　要：随着城市化的加速发展，居民的出行需求呈现多样化趋势，分析居民出行方式的选择行为成为优化交通系统的重要课题。该研究以多项Logit回归模型为核心，建立基于距离、时间、费用、年龄和天气等特征的居民出行方式选择预测模型。通过生成模拟数据，研究了电动车、摩托车、小汽车和公交车在不同条件下的选择概率。实验结果表明，该模型能够较高精度地预测居民出行方式的选择行为，准确率约93.58%。

关键词：交通方式选择；多项Logit回归；数据模拟；预测模型

引言

在现代城市中，居民的出行方式选择受到多种因素的影响，包括出行目的、交通工具的可获得性、出行费用、出行时间等。近年来，交通方式预测的研究主要集中在理解和预测居民的出行行为，尤其是考虑多维度的因素来预测居民选择不同交通方式的概率。李文权等人 [1]探索了中型城市居民出行方式选择行为，提出了粒子群优化随机森林模型来提高出行方式选择的预测精度。付琪 [2]提出了基于改进BP（Back Propagation）神经网络的出行方式选择预测方法，解决了传统BP神经网络在特征选择、结构设计和参数优化上的不足。张璐宇 [3]采用了基于BP神经网络的交通方式分担预测方法，结合Spearman相关性分析和XGBoost特征重要性分析，选择了交通方式选择的关键影响因素，最终构建了一个高效的交通方式分担预测模型。郭季 [4]基于山西省的城市居民出行选择数据，研究了不同规模城市居民的出行方式选择行为及其机理。通过结构方程模型（Structural Equation Modeling，简称SEM）分析，个人属性、出行活动属性、环境属性等因素对居民出行选择产生了重要影响。Huanfa Chen和Yan Cheng [5]提出了一种针对严重类别不平衡问题的机器学习方法，通过结合过采样/欠采样技术与多种预测方法，提出了一个系统化的评估框架。这一方法能有效提高少数类别（如共享单车等）在出行方式选择预测中的预测性能。

传统的交通方式选择模型，尤其是基于传统回归分析的模型，往往未能充分考虑到个体特征对选择的影响。基于此背景，如何综合考虑出行距离、出行时间、费用、个体年龄和天气等多重因素，建立一个精确的出行方式选择预测模型，是本研究的核心问题。

本文基于多项Logit回归模型的预测方法 [6]，探索居民在不同出行条件下（包括出行距离、出行时间、出行费用、个体年龄和天气等）对交通方式的选择行为。期望能够为政策制定者提供一种灵活且高效的工具，助力未来城市交通的智能化和个性化发展。

1. 数据生成与预处理

1.1. 数据生成

在实际交通出行方式选择的研究中，通常需要大量的个体选择数据以进行建模和分析。由于缺乏实际的调查数据，本研究采用了模拟数据生成的方法，通过编写MATLAB代码模拟生成四种交通方式（电动车、摩托车、小汽车、公交车）的出行数据，每种交通方式生成1000条数据，并赋予每条数据对应的交通方式标签。这些数据包括了出行距离、出行时间、出行费用、个体年龄和天气等特征。在数据生成过程中，为了模

拟出行行为的多样性和随机性，定义了每种交通方式的出行特征范围，包括出行距离、速度范围、起步费用和每公里费用等：

（1）出行距离：不同交通方式的出行距离范围被预设为不同的区间，这些范围根据交通工具的特性进行合理划分。电动车的出行距离范围为1-10公里，摩托车为2-20公里，小汽车为5-30公里，公交车为1-30公里。

（2）速度范围：每种交通方式的速度区间被定义为一定的随机范围，电动车的速度范围为20-30公里/小时，摩托车的速度范围为30-50公里/小时，小汽车的速度范围为40-80公里/小时，公交车的速度范围为20-40公里/小时。

（3）费用模型：费用的计算基于起步费用和每公里费用，与实际交通工具的运营成本相关。除了基本的费用计算外，费用值还添加了一定的随机波动，以模拟实际出行中的费用变化。

（4）年龄：不同年龄段的居民对交通方式的偏好差异较大。年轻人（15-22岁）往往倾向于选择电动车或公交车，而中年人（23-45岁）则更偏好摩托车或小汽车。年长者（46-60岁）通常倾向于选择公交车。

（5）天气：两轮车对恶劣天气最敏感，大雨时两轮车使用概率仅约为5%。小汽车对天气不敏感，大雨时使用率约10%，且小雨比例也更高。公交在雨天吸引力上升大雨时使用率10%，小雨时使用率约为30%。

在数据生成过程中，通过在时间和费用计算中引入正态分布随机波动，模拟了交通方式在真实环境下的随机性。电动车和摩托车的时间数据加入标准差为2的正态分布噪声，费用数据加入标准差为0.5的随机波动。这种处理增强了模拟数据的真实性，并避免生成数据过于理想化，具体代码如下：

数据生成与预处理：

```
traffic_modes = {'eBike', 'Motorcycle', 'Car', 'Bus'};
distance_ranges = [1 10; 2 20; 5 30; 1 30]; % 出行距离范围(km)
speed_ranges = [20 30; 30 50; 40 80; 20 40]; % 不同交通方式的速度范围(km/h)
base_fee = [1, 4, 10, 2]; % 起步价
fee_per_km = [0.5, 0.8, 1.5, 0.2]; % 每公里费用
bus_min_time = 10; % 公交车最小基础时间(分钟)
n_samples = 1000; % 每种交通方式的样本数量
age_means = [28, 33, 40, 35];
age_stds = [6, 5, 8, 7];
% 每行对应一种交通方式，四列为四种天气的概率，行和为1
weather_probs = [0.55 0.25 0.15 0.05; % eBike
0.50 0.30 0.15 0.05; % Motorcycle
0.35 0.30 0.25 0.10; % Car
0.30 0.30 0.30 0.10]; % Bus
```

1.2. 数据存储与处理

生成的数据被存储为表格形式，方便进一步的处理和分析。每条生成的出行数据包含五个字段：交通方式（Mode）、出行距离（Distance_km）、出行时间（Time_min）、出行费用（Cost_Yuan）、年龄（Age）和天气（Weather）。所有生成的数据被存储在一个单元格数组中，并且通过cell2table函数转换为MATLAB的表格格式。表格中列的顺序和数据类型被明确指定，以便后续进行分析和建模。

数据存储后，使用writetable函数将表格保存为Excel文件TrafficData_Logit.xlsx，还通过containers.Map对象实现对生成数据的交通方式编码映射，将四种交通方式的名称映射为相应的数字编码（1，2，3，4）。具体代码如下：

数据存储与处理：

```
data = {};
for i = 1:length(traffic_modes)
mode = traffic_modes{i};
```

```
% 随机生成距离
  distances = rand(n_samples, 1) * (distance_ranges(i, 2) - distance_ranges(i, 1))
+   distance_ranges(i, 1);
% 根据距离和速度范围计算时间
  avg_speed = rand(n_samples, 1) * (speed_ranges(i, 2) - speed_ranges(i, 1)) +
speed_ranges(i, 1);
times = distances ./ avg_speed * 60;
if strcmp(mode, 'Bus')
% 公交车时间包含基础时间和站点停留时间
  times = max(times + randn(n_samples, 1) * 3, bus_min_time); % 加入基础时间
else
% 其他交通方式加入随机波动
  times = times + randn(n_samples, 1) * 2;
end
% 根据距离计算费用
  fees = base_fee(i) + distances * fee_per_km(i);
  fees = fees + randn(n_samples, 1) * 0.5; % 加入随机波动
% 将生成的数据存储到表中
for j = 1:n_samples
  data{end+1, 1} = mode;
  data{end, 2} = round(distances(j), 2);
  data{end, 3} = round(times(j), 1);
  data{end, 4} = round(fees(j), 2);
  age = max(round(age_means(i) + age_stds(i) * randn()), 16);
  data{end, 5} = min(age, 75);
  weather = randsample(1:4, 1, true, weather_probs(i,:));
  data{end,6} = weather;
 end
end
% 将数据转换为表格格式
T = cell2table(data, 'VariableNames', {'Mode', 'Distance_km', 'Time_min',
'Cost_Yuan','Age','Weather'});
% 保存生成的数据到 Excel 文件
output_file = 'TrafficData_Logit.xlsx';
if exist(output_file, 'file')
  delete(output_file);
end
writetable(T, output_file, 'FileType', 'spreadsheet');
```

```
disp('数据生成完成，已保存为 TrafficData_Logit.xlsx');
% 将交通方式转换为数值编码
unique_modes = unique(T.Mode);
mode_mapping = containers.Map(unique_modes, 1:length(unique_modes));
T.ModeCode = cellfun(@(x) mode_mapping(x), T.Mode);
```

1.3. 数据划分

在数据预处理的过程中，生成的数据集被划分为训练集和测试集。通过交通方式的编码对数据进行分类，确保每种交通方式都有足够的样本量。为了评估模型的泛化能力，通过逻辑索引，每种交通方式的前70%条数据被分配为训练集，剩余的30%条数据作为测试集，确保了训练集和测试集的样本分布与整体数据集一致，避免过拟合的问题。

2. 模型与方法

2.1. 多项Logit回归模型

多项Logit回归（Multinomial Logit Regression, MLR）是一种广泛应用于分类问题的统计模型，尤其适用于因变量为类别型的情况。在该交通方式选择预测问题中，居民的出行方式（电动车、摩托车、小汽车、公交车）属于一个多类别问题，该模型通过学习不同交通方式的特征与选择之间的关系，预测给定特征下某个交通方式被选择的概率。

多项Logit回归模型是基于Logit函数扩展到多类别的情形。对于每一个交通方式，模型会计算一个类别的概率，最终选择概率最大的类别作为预测结果。假设有K个交通方式类别，模型通过以下方式计算每个类别的概率，如公式（1）所示：

$$P(y=k|X)=\frac{exp(\beta_k^T X)}{\sum_{j=1}^{K} exp(\beta_j^T X)} \tag{1}$$

其中，y 为交通方式类别，X 为特征向量（出行距离、时间和费用等），k 为类别 y 的回归系数。通过训练数据优化回归系数，使得模型在每个类别的概率预测上尽可能准确。

2.2. 模型实现

使用MATLAB软件提供的mnrfit函数训练多项Logit回归模型。通过最大似然估计来拟合数据，得到各类别回归系数。具体过程如下：

（1）数据准备：将生成的数据集中的特征作为输入特征矩阵 X，将交通方式编码作为目标变量 y。

（2）数据标准化：为了确保特征的尺度一致并提高模型训练的效果，使用标准化方法将特征矩阵 X 进行标准化处理。标准化的过程是对每个特征减去其均值并除以标准差，确保每个特征的均值为0，标准差为1，从而提升多项Logit回归模型的收敛速度和稳定性。

（3）模型训练：使用mnrfit函数训练多项Logit回归模型。该函数接受特征矩阵 X 和目标变量 y，通过最优化过程拟合回归系数，输出训练好的模型。

（4）模型预测：训练完成后，用mnrval函数对测试数据进行预测。该函数返回每个类别的预测概率，然后选择概率最大的类别作为最终预测结果。

（5）结果分析：通过计算预测结果与真实结果之间的准确率，评估模型的性能。

2.3. 模型性能评估

为了评估多项Logit回归模型的性能，使用预测准确率作为主要的评估指标。预测准确率是通过计算模型预测结果与真实标签之间的匹配程度来得到的。准确率的计算方法如公式（2）所示：

$$Accuracy=\frac{\sum_{i=1}^{n} I(y_i^*=y_i)}{n} \tag{2}$$

其中， n 是测试集的样本数量， $\mathrm{I}(y_i^* = y_i)$ 是指示函数，若预测结果 y_i^* 与真实标签 y_i 相同，则其值为1，否则为0。

在模型训练完成后，使用测试集对模型进行评估。通过对测试集的预测结果与测试集真实标签进行比较，得到模型的预测准确率。

多项Logit回归模型通过训练数据学习交通方式选择的规律，结合实际出行特征来预测居民的出行方式。通过准确率等指标的评估，该模型能够提供对不同交通方式选择的可靠预测，为交通规划和管理提供一定的数据支持。

模型与方法：

```
% 使用mnrfit训练多项Logit回归模型
disp('开始训练多项Logit回归模型...');
B = mnrfit(X_train, categorical(y_train)); % 训练模型
% 预测测试数据
disp('对测试集进行预测...');
P_test = mnrval(B, X_test); % 预测概率
[~, predictions] = max(P_test, [], 2); % 获取概率最大值对应的类别
% 输出预测准确率
accuracy = mean(predictions == y_test);
fprintf('测试集预测准确率: %.2f%%\n', accuracy * 100);
```

3. 结果与分析

3.1. 预测结果

通过多项Logit回归模型对测试集数据进行预测，得到每个样本选择特定交通方式的概率，得到的测试集预测准确率为93.58%，预测结果如图1和图2所示。该结果表明多项Logit回归模型在此数据集上具有较好的性能，能够有效地区分不同的交通方式。其能够利用出行距离、时间和费用等特征，准确预测居民在不同情境下选择交通方式的概率。

ActualMode	PredictedMode	Distance_km	Time_min	Cost_Yuan	Age	Weather	eBike_Probability	Motorcycle_Probability	Car_Probability	Bus_Probability	CorrectPrediction
Bus	Bus	8.02	22.4	4.7	26	2	0.131141225	0	2.985E-119	0.868858775	TRUE
Bus	eBike	3.51	4.8	1.44	36	2	0.832529328	0	2.70756E-81	0.167470672	FALSE
Bus	Bus	5.01	11.9	4.16	31	1	0.025748577	0	8.79152E-51	0.974251423	TRUE
Bus	Bus	4.91	13	3.14	18	3	0.08231114	0	3.1094E-93	0.91768886	TRUE
Bus	Bus	2.27	9.3	2.64	40	1	0.194045641	0	1.4446E-45	0.805954359	TRUE
Bus	Bus	3.75	9.6	3.12	29	3	0.160606554	0	7.37666E-60	0.839393446	TRUE
Bus	Bus	1.76	1.3	1.75	22	2	0.055136263	2.8045E-305	3.25957E-43	0.944863737	TRUE
Bus	Bus	8.1	24.1	4.48	16	4	0.173281102	0	1.0371E-141	0.826718898	TRUE
Bus	Bus	8.49	15.7	5.61	31	1	0.026761847	0	2.50385E-67	0.973238153	TRUE
Bus	Bus	1.01	0.8	1.2	28	4	0.498674824	0	9.40041E-48	0.501325176	TRUE
Bus	Bus	5.65	14.2	3.14	35	1	0.410838959	0	1.52584E-95	0.589161041	TRUE
Bus	Bus	2.05	4.4	1.59	16	2	0.042994017	1.1079E-268	3.18009E-67	0.957005983	TRUE
Bus	Bus	2.76	7.6	2.25	16	1	0.012847405	2.6432E-273	1.49964E-68	0.987152595	TRUE
Bus	Bus	5.13	15.7	3.59	24	3	0.122528097	0	7.98511E-91	0.877471903	TRUE
Bus	eBike	3.64	14	1.58	33	2	0.818897598	0	5.8327E-118	0.181102402	FALSE
Bus	Bus	6.49	16.1	3.86	29	1	0.119536499	0	6.83838E-96	0.880463501	TRUE
Bus	Bus	6.51	16.7	4.41	25	1	0.02475068	0	8.23399E-85	0.97524932	TRUE
Bus	Bus	7.06	20.9	4.24	20	1	0.029696359	0	2.8209E-116	0.970303641	TRUE
Bus	Bus	8.36	26.5	5.08	34	1	0.183427106	0	1.4587E-123	0.816572894	TRUE
Bus	Bus	4.39	13	2.35	16	1	0.048869872	0	8.7636E-108	0.951130128	TRUE
Bus	Bus	4.4	9.5	3.64	31	2	0.072970128	0	6.34365E-50	0.927029872	TRUE
Bus	Bus	9.13	26.4	4.83	28	1	0.195324893	0	4.986E-143	0.804675107	TRUE
Bus	Bus	8.59	15.4	5.5	27	3	0.068913183	0	1.63033E-75	0.931086817	TRUE
Bus	Bus	9.56	31.7	5.67	16	1	0.012493875	0	5.1294E-152	0.987506125	TRUE
Bus	Bus	8.86	17.3	5.98	30	1	0.016301214	0	2.84586E-68	0.983698786	TRUE
Bus	Bus	6.53	15.4	4.03	23	1	0.034646312	0	9.02236E-92	0.965353688	TRUE
Bus	Bus	6.02	14.3	4.3	25	3	0.068929245	0	4.81932E-75	0.931070755	TRUE
Bus	Bus	5.14	11.5	4.28	30	3	0.070285698	0	1.0405E-50	0.929714302	TRUE
Bus	Bus	7.89	12.7	5.59	20	2	0.004897917	0	1.54334E-55	0.995102083	TRUE
Bus	Bus	5.23	12.4	3.41	34	1	0.185318574	0	1.2344E-75	0.814681426	TRUE
Bus	eBike	8.16	18.4	4.59	33	4	0.669554495	0	2.1498E-107	0.330445505	FALSE
Bus	eBike	5.33	13	2.96	34	4	0.828272271	0	2.91192E-97	0.171727729	FALSE

图1 预测结果

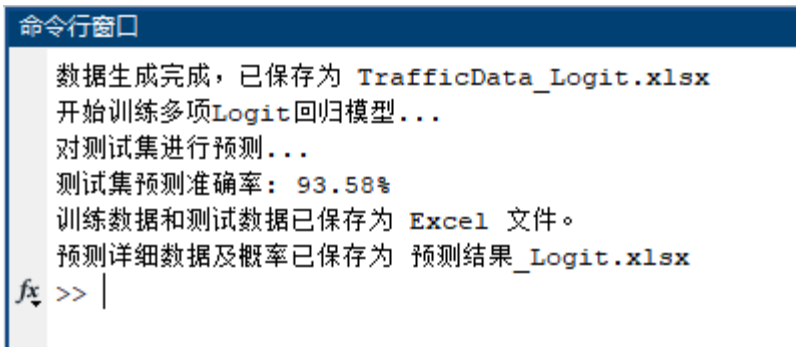

图2 预测结果

3.2. 交通方式选择分析

在模型训练过程中，特征对交通方式选择的影响被自动学习并量化为回归系数。通过对这些系数的分析，可以得出以下几点结论：

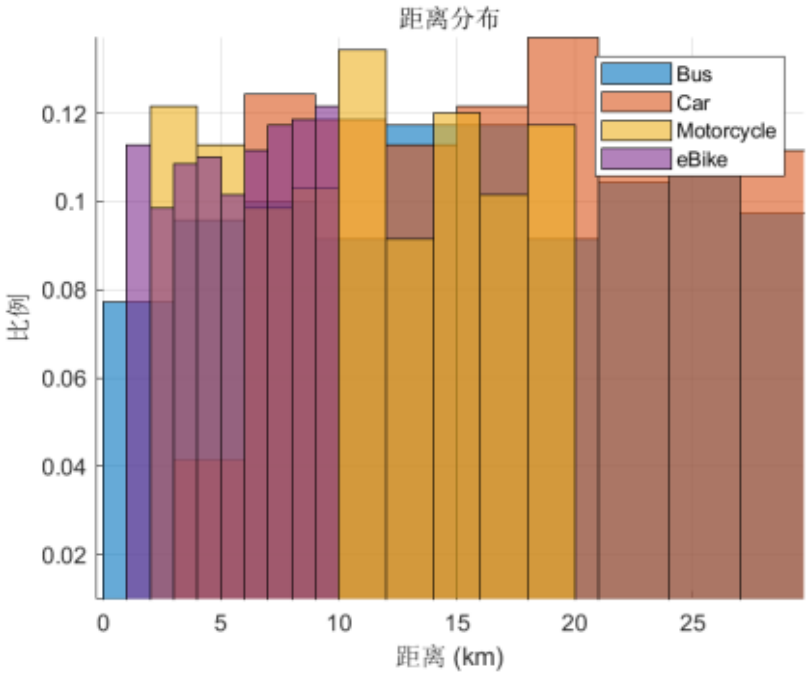

图3 距离分布

（1）出行距离：如图3，不同出行方式在距离维度呈现各异特征。公交（Bus）在短距离（0-5km左右）有一定占比，长距离（15-30km）占比也较突出，说明公交适用于多种距离出行；汽车（Car）在中长距离（10-30km）占比较高，体现其适合稍远出行；摩托车（Motorcycle）在10-25km区间占比大，是中距离出行常用选择；电动自行车（eBike）集中在短距离（0-15km），是短程通勤得力工具。

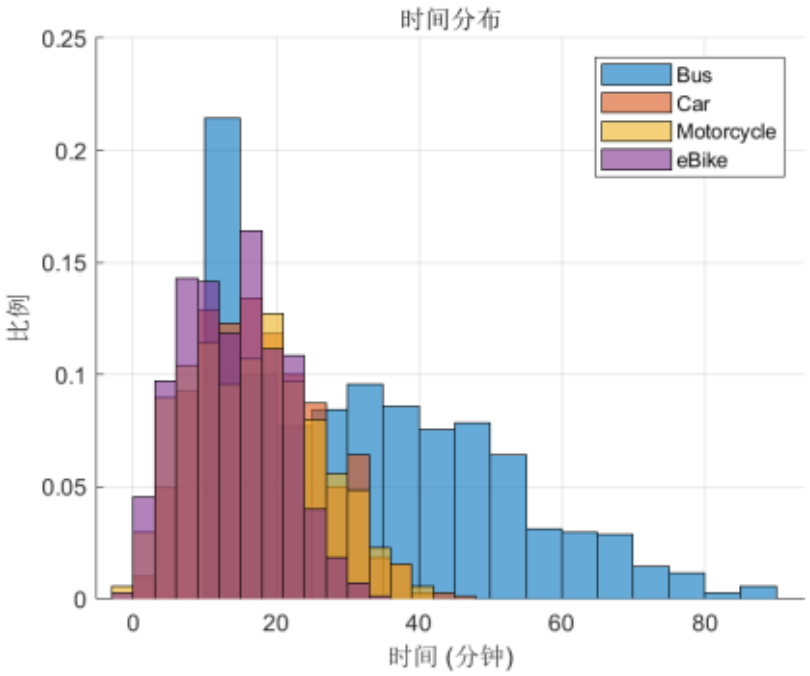

图4 时间分布

（2）出行时间：如图4，公交（Bus）出行时间分布较广，20-80分钟均有占比，且在20-30分钟左右有相对集中区域，反映公交受线路、站点等影响，出行时长跨度大；汽车（Car）在0-40分钟占比较高，说明中短时间出行选汽车较便捷；摩托车（Motorcycle）主要集中在0-30分钟，中短耗时出行适配；电动自行车（eBike）集中于0-20分钟，短时间出行优势明显，适合快速短途通勤。

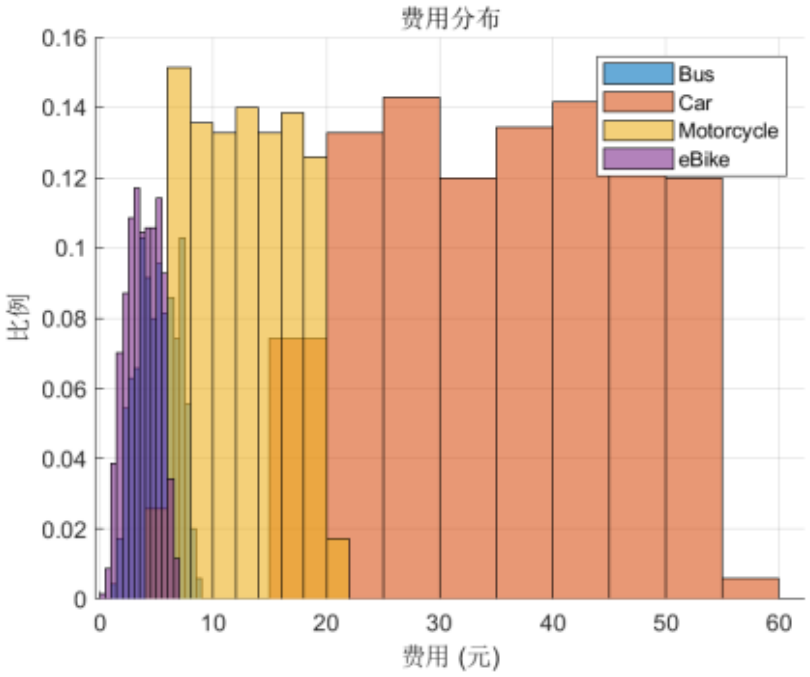

图5 费用分布

（3）出行费用：如图5，公交车（Bus）费用最低且分布窄，会吸引大量选择低预算出行的居民。相对而言，小汽车（Car）费用分布较高且宽，由于高昂的油费和停车费用，通常在高收入群体中更为常见。而电动

车（eBike）和摩托车（Motorcycle）在费用上介于公交车和小汽车之间，是一种既具有经济性又具备一定灵活性的选择。

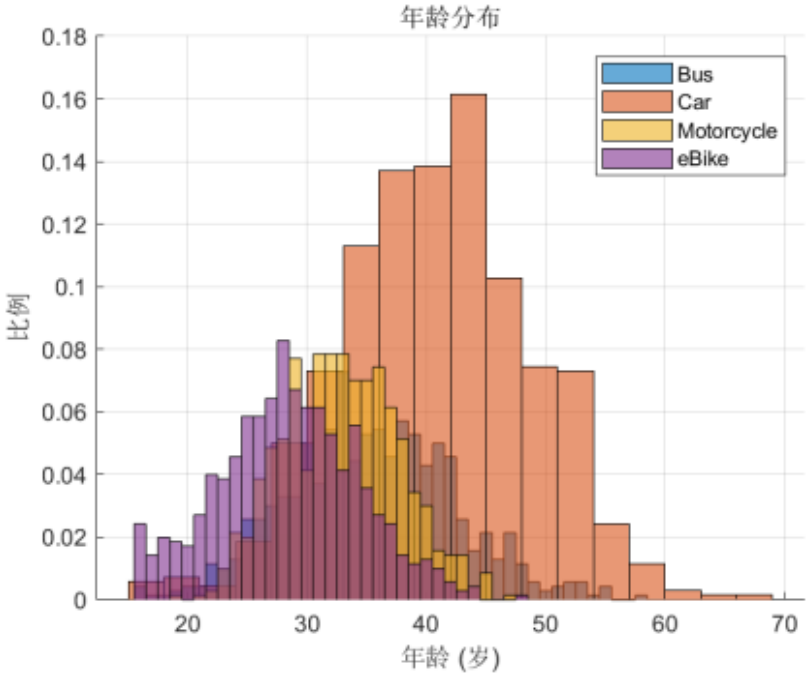

图6 年龄分布

（4）年龄：如图6，公交（Bus）在各年龄段均有使用，中青年（30-50岁左右）占比相对突出，不同年龄层因出行需求、习惯等，公交是通用选择；汽车（Car）使用人群集中在25-55岁，该年龄段有较强经济与出行需求，依赖汽车出行；摩托车（Motorcycle）使用者多为20-45岁，年轻群体因追求便捷、灵活，对其偏好高；电动自行车（eBike）在20-50岁分布，尤其受中青年短程通勤青睐，便捷且成本低。理解不同年龄段居民的交通方式选择特点对于优化交通系统、提供针对性服务具有重要意义。通过制定有针对性的交通政策，可以更好地满足不同群体的出行需求，推动城市交通的可持续发展。

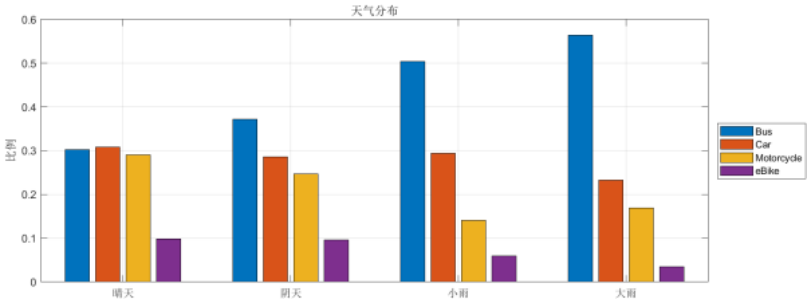

图7 天气分布

（5）天气：如图7，晴天时，公交（Bus）、汽车（Car）、摩托车（Motorcycle）使用比例相对均衡，电动自行车（eBike）占比稍低，说明好天气下多种出行方式竞争；阴天时，公交占比提升，汽车、摩托车等略有下降，可能部分人因天气选公共交通；小雨天气，公交占比明显高，汽车其次，摩托车、电动自行车因防雨不便占比低；大雨时，公交占比显著最高，汽车次之，摩托车、电动自行车受天气影响大，民众优先选公交、汽车等更具防护性的出行方式。

通过这些分析，不仅能够理解模型的预测结果，还能够对居民出行方式选择的潜在规律做出合理推测。若出行者的预算较紧张并且距离较短，则很可能选择电动车或公交车；如果出行时间要求较高且预算允许，则小汽车或摩托车可能会成为首选。交通费用和出行时间的相互作用，以及居民的个性化需求，都会影响最终的选择。

4. 结语

本研究基于多项Logit回归模型，探讨了城市居民出行方式选择的预测问题。通过MATLAB代码生成与现实较接近的四种交通方式出行特征模拟数据，作为模型训练输入。应用多项Logit回归模型对城市居民交通方式选择建模，预测居民在不同出行特征下选择交通方式的概率，该模型可依出行特征预测选择概率，在训练集与测试集均有高准确率，验证了其有效性。本研究为交通方式选择预测提供可行建模方法，对未来交通规划有实践意义，随着实际数据和先进模型应用，有望在智能交通系统、公共政策优化等领域发挥重要作用。

参考文献

[1] 李文权, 邓安鑫, 郑炎, 等. 基于机器学习的中型城市居民出行方式选择行为研究 [J]. 交通运输系统工程与信息, 2024, 24(2): 13-23.

[2] 付琪. 基于改进BP神经网络的出行方式选择预测方法研究 [D]. 西华大学, 2023.

[3] 张璐宇. 基于BP神经网络的交通方式分担预测方法研究 [D]. 石家庄铁道大学, 2022.

[4] 郭季. 不同规模城市居民出行方式选择行为机理研究 [D]. 长安大学, 2020.

[5] CHEN H, CHENG Y. Travel mode choice prediction using imbalanced machine learning [J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(4): 3795-3808.

[6] 郝小妮, 石文瀚, 刘建荣, 等. 基于随机系数Logit模型的城市群城际出行方式选择行为研究 [J]. 交通信息与安全, 2022, 40(5): 139-146.