

乘坐高铁还是传统火车的行为分析

摘要

高铁，因其速度快、乘坐舒适等特征，已经成为了当代人出行的主要交通工具之一。同时，相比于高铁，传统火车也具有独特的优势。在这一背景下，当人大学生选择乘坐高铁还是火车返乡时会综合考虑多方因素。我们需要建立数学模型，探索这些因素的影响，以便实现高铁及火车网络的最优管理。

针对问题一，首先，我们根据文献资料，获取影响选择高铁还是火车的主要因素。随后，我们使用 R 型聚类算法，建立“购票行为影响因素评价模型”，可以定性描述各个因素与抉择结果之间的关系大小。

针对问题二，首先，应用“购票行为影响因素评价模型”，得到对抉择影响最大的因素为“高铁/火车行程时间”。随后，我们使用最小二乘回归算法，建立“购票行为预测模型”，对影响抉择的因素从高到低排序。最后，我们设立“选择高铁还是火车”的指标 Q ，用其与 0.5 的关系预测选择高铁还是火车。

针对问题三，首先，我们运用上述“购票行为预测模型”，计算了下一个寒假学生购票行为的预测值。随后，我们判断若该值大于 0.5，则认为乘坐高铁，若小于 0.5，则认为乘坐火车。最后，我们给出了以及买高铁与火车票的具体人数及百分比。

针对问题四，首先，我们计算了每省份的乘坐高铁人数与乘坐火车人数的比值，利用数据绘图，得到距离与抉择结果之间的关系。随后，我们整合上述关系及前三个问题得到的结论，为铁路管理部门撰写了建议书。

综上，我们解决了题目所给问题，并将在未来完善模型，以扩大适用范围。

关键词：高铁；火车；R 型聚类；最小二乘回归。

目录

0 引言 1

0.1 背景 1

0.2 分析 1

0.3 假设 2

1 问题一 3

1.1 筛选影响抉择结果的主要因素 3

1.2 “购票行为影响因素评价模型” 3

1.2.1 变量相似性度量 3

1.2.2 变量聚类 4

2 问题二 4

2.1 对“购票行为影响因素评价模型”的应用 4

2.2 “购票行为预测模型” 5

2.2.1 最小二乘法的基本原理 5

2.2.2 算法结果 5

2.2.3 判断选择高铁还是火车的准则 6

3 问题三 6

3.1 计算“选择高铁还是火车”的指标 Q 6

3.2 得到每位同学出行方式的预测结果 7

4 问题四 7

4.1 计算每个省份乘坐高铁人数与乘坐火车人数之比 7

5 模型的评价与改进 11

5.1 优点 11

5.2 缺点 11

A 代码 12

插图

1	流程图	2
2	聚类图	4
3	中国各省高铁、火车乘客人数之比	9
4	运行界面	13
5	运行界面	15
6	运行界面	17

表格

1	拟合式中各个因素的系数以及截距	6
2	每个省份乘坐高铁人数与乘坐火车人数之比	8

问题

问题 1	1
问题 2	1
问题 3	1
问题 4	2

§ 引言

0.1 背景

2018 年底中国高铁运营里程超过 2.9 万公里，占全球高铁运营里程的三分之二以上，超过其他国家总和。高铁，因其速度快、乘坐舒适等特征，已经成为了当代人出行的主要交通工具之一。同时，相比于高铁，传统火车也具有独特的优势，例如，它的价格更加低廉，也有人喜欢在乘坐速度较慢的火车时享受路边风景。在这一背景下，高铁仍然无法完全取代传统火车。当人们面临出行选择的时候，会综合考虑经济状况、里程长度、舒适度等多方因素后决定乘坐高铁还是传统火车。[1]

0.2 分析

人们会根据各种因素来抉择：乘坐高铁还是火车。

问题 1

筛选影响抉择的主要因素及理由，建立数学模型描述抉择结果与因素之间的关系。

回答 1

首先，我们根据文献资料，获取影响选择高铁还是火车的主要因素。随后，我们使用 R 型聚类算法，建立“购票行为影响因素评价模型”，可以定性描述各个因素与抉择结果之间的关系大小。

问题 2

只考虑附件一给出的因素，根据数据估计上述模型参数，建立可供计算的乘客购票行为数学模型。着重对影响抉择的因素及因素间关系进行量化分析；建立准则，对影响抉择的因素从高到低排序。

回答 2

首先，应用上述“购票行为影响因素评价模型”，得到对抉择影响最大的因素。随后，我们使用最小二乘回归算法，建立“购票行为预测模型”对影响抉择的因素从高到低排序，对影响抉择的因素及因素间关系进行量化分析。流程图见图1。

问题 3

根据附件二，预测并列表给出每个人的购票结果，以及买高铁与火车票的具体人数及百分比。

回答 3

首先，我们运用上述“购票行为预测模型”，计算了下一个寒假学生购票行为的预测值。随后，我们判断若该值大于 0.5，则认为乘坐高铁，若小于 0.5，则认为乘坐火车。最后，我们给出了以及买高铁与火车票的具体人数及百分比。

问题 4

针对一定区域、特定阶段、并具有代表性人群的购票行为，分析高铁客运量与传统火车客运量的规律，为铁路管理部门撰写建议书。

回答 4

首先，我们计算了每省份的乘坐高铁人数与乘坐火车人数的比值，利用数据绘图，得到距离与抉择结果之间的关系。随后，我们整合上述关系及前三个问题得到的结论，为铁路管理部门撰写建议书。

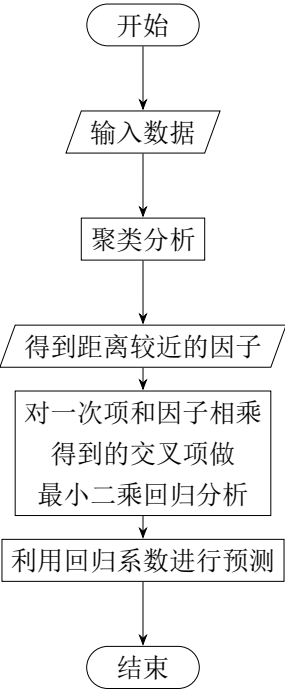


图 1: 流程图

0.3 假设

1. 学生在填写附件中从南京回程购票信息调查表时均按照真实意愿和实际情况填写。

2. 高铁及火车的价格、路线长度、行进速度、购票方式、舒适度在上次假期到下次寒假的时间段内不发生变化。
3. 学生在考虑购买高铁还是火车时的思维方式、侧重点、偏好等在上次假期到下次寒假的时间段内不发生变化。
4. 学生购票时不考虑自然灾害、人为破坏等情况对行进速度、舒适度等因素造成的影响。

§1 问题一

1.1 筛选影响抉择结果的主要因素

通过对文献进行研究，影响乘客选择高铁还是火车的因素主要来自两方面：一是客运方式选择的需求特性，与客运的主体（学生）有关；二是客运方式选择的供给特性，与交通运输方式有关。

综合以上两个层面的因素后，得到影响选择高铁还是火车的主要因素为：职业、出行目的、路程、收入水平、行进时间、票价、舒适度、安全性。

1.2 “购票行为影响因素评价模型”

我们使用 R 型聚类算法来评价乘客购票行为（结果）与这些因素之间关系。

1.2.1 变量相似性度量

我们使用相关系数来度量变量相似性。

记变量 x_j 的取值 $(x_{1j}, (x_{2j}, \dots, (x_{nj})^T \in R^n (j = 1, 2, \dots, m)$ 。则可计算变量 x_j 和 x_k 的相关系数：

$$r_{jk} = \frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}} \quad (1)$$

从而可以得到相关系数矩阵。

1.2.2 变量聚类

我们运用最长距离法对各因素聚类。定义两类变量的距离为：

$$R(G_1, G_2) = \max_{x_j \in G_1, x_k \in G_2} \{d_{jk}\} \quad (2)$$

式中：

$$d_{jk} = 1 - |r_{jk}| \quad (3)$$

或

$$d_{jk}^2 = 1 - r_{jk}^2 \quad (4)$$

§2 问题二

2.1 对“购票行为影响因素评价模型”的应用

我们选区附件一中所列因素以及结果“选择高铁或火车”作为 R 型聚类因素，使用 MATLAB 软件编程得到聚类图：

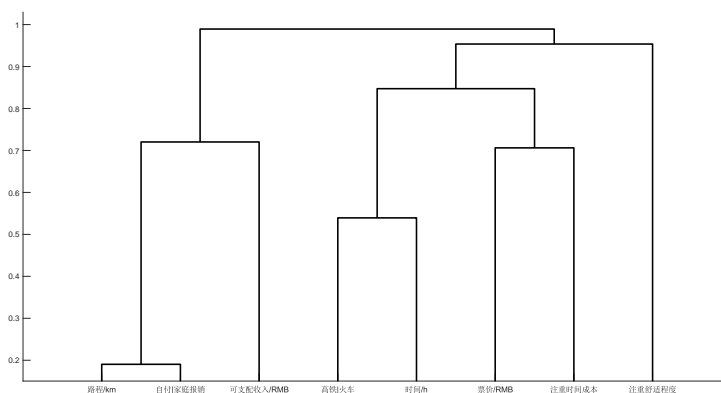


图 2: 聚类图

从聚类图中可以看出，各个因素与“选择高铁还是火车”的相似性可以分为三档：

1. 第一档（相似度最大）：高铁/火车行程时间
2. 第二档（相似度较大）：票价

3. 第三档（相似度较小）：舒适程度

4. 第四档（相似度较小）：路程、自付/家庭报销、可支配收入

可见，高铁/火车行程时间是影响抉择结果的最大因素。大部分大学生群体追求快节奏生活，不希望在路途中花费过多的时间，希望用最高效的方式来完成旅途。同时，行程时间短，也是高铁被发明的最初目的，也是它与火车相区别的主要因素。因此聚类结果符合大学生实际与生活实际。大学生对于时间的看重最大程度决定了对高铁还是火车的选择。由于高铁的时间更短，所以更加追求快速的学生会更倾向于高铁，而对速度不看重的同学会倾向于价格（第二档）低廉的火车。

2.2 “购票行为预测模型”

我们使用线性最小二乘回归算法，得到选择高铁还是火车与各个因素之间的定量关系。

2.2.1 最小二乘法的基本原理

已知一组数据，即平面上的 n 个点 $(x_i, y_j), i = 1, 2, \dots, n$, x_i 互不相同，寻求一个函数（曲线） $y = f(x)$ ，使得 $f(x)$ 在某种准则下与所有数据点最为接近，则曲线拟合得最好。

而线性最小二乘法是解决曲线拟合最常用的方法，基本思路是，令：

$$f(x) = a_1 r_1(x) + a_2 r_2(x) + \dots + a_m r_m(x) \quad (5)$$

式中：

$r_k(x)$ 为事先选定好的一组线性无关的函数；

a_k 为待定系数（ $k = 1, 2, \dots, m; m < n$ ）。

首先，应用上述“购票行为影响因素评价模型”，得到对抉择影响最大的因素。随后，我们使用最小二乘回归算法，建立“购票行为预测模型”对影响抉择的因素从高到低排序，对影响抉择的因素及因素间关系进行量化分析。

2.2.2 算法结果

通过编程，我们确定了“乘坐高铁还是火车”作为因变量的拟合式中各个因素的系数：

表 1: 拟合式中各个因素的系数以及截距

标记	因素	单位	系数
x_1	注重时间成本	/	0.19777
x_2	可支配收入	RMB	0.060269
x_3	票价	RMB	-0.05013
x_4	票价 \times 注重时间成本	RMB	-0.047532
x_5	时间	h	-0.041519
x_6	自付/家庭报销	/	0.0017471
x_7	注重舒适程度	/	4.26×10^{-5}
x_8	路程	km	-8.64×10^{-6}
x_9	路程 \times 自付/家庭报销	km	-2.83×10^{-7}

截距:

$$b = 0.37519 \quad (6)$$

因此, 若设决定“选择高铁还是火车”的指标为 Q , 则:

$$Q = 0.198x_1 + 0.060x_2 - 0.050x_3 - 0.048x_4 - 0.042x_5 + 0.00174x_6 + 4.26 \times 10^{-5}x_7 - 8.64 \times 10^{-6}x_8 - 2.83 \times 10^{-7}x_9 \quad (7)$$

且从拟合结果可以看出, 影响顾客购票行为的因素按从高到低排序为:
时间 > 可支配收入 > 票价 > 自付/家庭报销 > 舒适程度 > 路程

2.2.3 判断选择高铁还是火车的准则

若 $Q \geq 0.5$, 则预测选择乘坐高铁;

若 $Q < 0.5$, 则预测选择乘坐火车;

§3 问题三

3.1 计算“选择高铁还是火车”的指标 Q

利用附件 2 的数据, 运用式7, 计算每位同学的 Q 值。

3.2 得到每位同学出行方式的预测结果

通过将 Q 与 0.5 进行比较，得到每位同学出行方式的预测结果。

在附件 2 的 85 名同学中，下一个寒假要购买高铁票的人数为 68 人，购买火车票的人数为 17 人。

购买高铁票的人数百分比为：

$$\eta_1 = \frac{68}{85} = 80\% \quad (8)$$

购买火车票的人数百分比为：

$$\eta_1 = \frac{17}{85} = 20\% \quad (9)$$

§4 问题四

4.1 计算每个省份乘坐高铁人数与乘坐火车人数之比

为了能够得到对于前往不同地域，学生对高铁还是火车的选择的差别，我们分别计算了每个省份乘坐高铁人数与乘坐火车人数之比。

表 2: 每个省份乘坐高铁人数与乘坐火车人数之比

省份	高铁乘客人数/火车乘客人数比值
北京	8
天津	1
河北	3
山西	0.67
内蒙古	0
辽宁	6
吉林	1.5
黑龙江	3
上海	1
江苏	1.25
浙江	14
安徽	0.75
福建	5
江西	0.6
山东	2.5
河南	1
湖北	5
湖南	3.5
广东	2
广西	1
海南	1
重庆	2.5
四川	2.3
贵州	1
云南	0.1
西藏	0.1
陕西	0.8
甘肃	1
青海	0.75
宁夏	0.6
新疆	0.1

从结果可以看出，从南京去往的省份距离南京越远，乘坐火车的比例越多；离南京越近，乘坐火车的人越多。

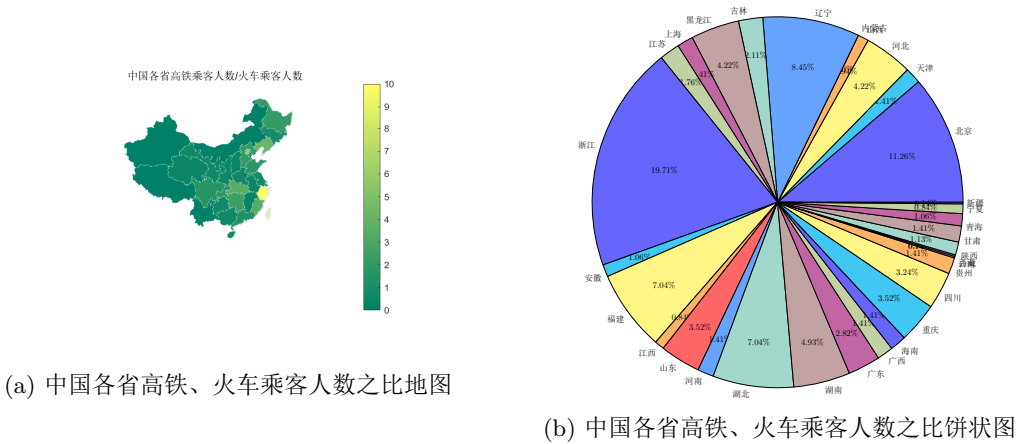


图 3: 中国各省高铁、火车乘客人数之比

供铁路管理部门参考与决策的建议书

尊敬的铁路部门领导及工作人员：

我们是三名本科学学生组成的调研小组。近期，我们通过问卷调查及数学模型建立与运用，对我校学生从南京返程时选择乘坐高铁还是火车的行为决策进行了分析，研究了一定区域、特定阶段、并具有代表性人群的购票行为，分析高铁客运量与传统火车客运量的规律，得到了一些结论，并希望向铁路部门提出一些在高铁/火车数量安排等问题上的建议。

首先，经过聚类分析，高铁/火车行程时间是影响大学生抉择乘坐高铁还是火车的最主要因素。大部分大学生群体追求快节奏生活，不希望在路途中花费过多的时间，希望用最高效的方式来完成旅途，因此会更加倾向于乘坐高铁。同时，行程时间短，也是高铁被发明的最初目的，也是它与火车相区别的主要因素。因此，我们建议铁路部门在保证安全的前提下通过改进高铁相关技术、优化运营线路等措施，提升高铁行进速度，从而满足更多学生返程的需求。

第二，经过最小二乘法拟合，影响大学生购票行为的因素按从高到低排序为：时间 > 可支配收入 > 票价 > 自付/家庭报销 > 舒适程度 > 路程。可以发现，在排在首位的“时间”之后，连续三个因素都与金钱有关，表明大学生购买高铁还是火车很大程度上取决于票价、家庭经济状况及支付来源等。因此，我们建议铁路部门在上述“提升高铁行进速度的同时”，减小技术成本，适当降低高铁票的价格，在为乘客提供快速旅行服务的同时，让更多大学生能够负担高铁的价格。

最后，经过统计，从南京去往的省份距离南京越远，乘坐火车的比例越多；离南京越近，乘坐火车的人越多。这表明对于大学生而言，更倾向于短途高铁及长途火车。原因可能是随着距离的增加，高铁票价的增速大于火车票价的增速，长途高铁的价格非常昂贵，对于大学生来说出行成本太高。因此我们建议铁路部门可以适当降低长途高铁的票价，尤其寒假放假、暑假放假、国庆节放假等学生返乡高峰期，或增大学生票优惠力度，以满足更多学生需求。

以上是我们对大学生选择乘坐高铁还是火车的行为决策进行的分析，以及有关建议，希望可以对铁路管理部门的管理工作提供参考与帮助，为学生返乡旅途提供更大的便利。如有不足，还望领导与工作人员扶正！

感谢您的阅读！

学生调研小组

2019 年 6 月 30 日

§5 模型的评价与改进

5.1 优点

1. 运用了 R 型聚类、最小二乘法等算法，通过客观数据得到结论。
2. 最小二乘拟合使得“选择高铁还是火车”有了一个定量的判断准则，可以明确地预测学生下一寒假的出行方式。

5.2 缺点


1. 没有运用精确的算法来筛选影响购买高铁票还是火车票的主要因素，而是仅仅根据文献资料的结论。
2. “选择高铁还是火车”的结果与各个因素之间的关系不一样符合线性关系，因此线性拟合得到的预测结果可能不准确。
3. 在判断目标省份离南京的距离与选择高铁还是火车之间的关系时，有些省份没有开通高铁，或者有的省份的数据非常少，因此结果存在极端化情况，不够准确。

参考文献

[1] 巩慧琴, “高铁时代下旅客交通工具选择行为研究,” Ph.D. dissertation, 辽宁师范大学, 2012.

1

§A 代码

```
程序清单 1: R.m      Matlab
```

```
1 tic;
2 clc;
3 clear;
4 close;
5
6 %%in
7 data = csvread('data.csv', 1, 2);
8 data(data(:, 4) == 1, :) = [];
9 data(:, 4) = [];
10 data(:, [1 2])=data(:, [2 1]);
11 name = {" 高铁|火车", " 路程/km", " 时间/h", " 注重舒适程度", " 可支配收入/RMB", " 自付|家庭报销", " 票价/RMB", " 注重时间成本"};
12 numClust = 2;
13
14 %%process
15 d=1-abs(corrcoef(data)); % 进行数据变换, 把相关系数转化为距离
16 d=tril(d); % 提出 d 矩阵的下三角部分
17 d=nonzeros(d); % 去掉 d 中的 0 元素
18 z=linkage(d', 'complete'); % 按最长距离法聚类
19 y=cluster(z, 'maxclust', numClust);
20 for i = 1:numClust
21     clust{i}={name{y(:)==i}};
```

```
22 end
23 h=dendrogram(z); % 画聚类图
24 set(h,'Color','k','LineWidth',2.0);% 把聚类图线的颜色修改成黑色，线宽加粗
25 xtick = get(gca,'Xticklabel');
26 Xtick = cell(1, length(xtick));
27 for i = 1:length(xtick)
28     Xtick{xtick(:)==num2str(i)}=name{i};
29 end
30 set(gca,'Xticklabel', Xtick);
31
32 %%out
33 fprintf('Running time is %f second.\n', toc)
```

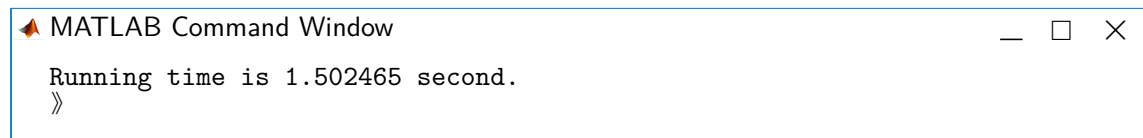
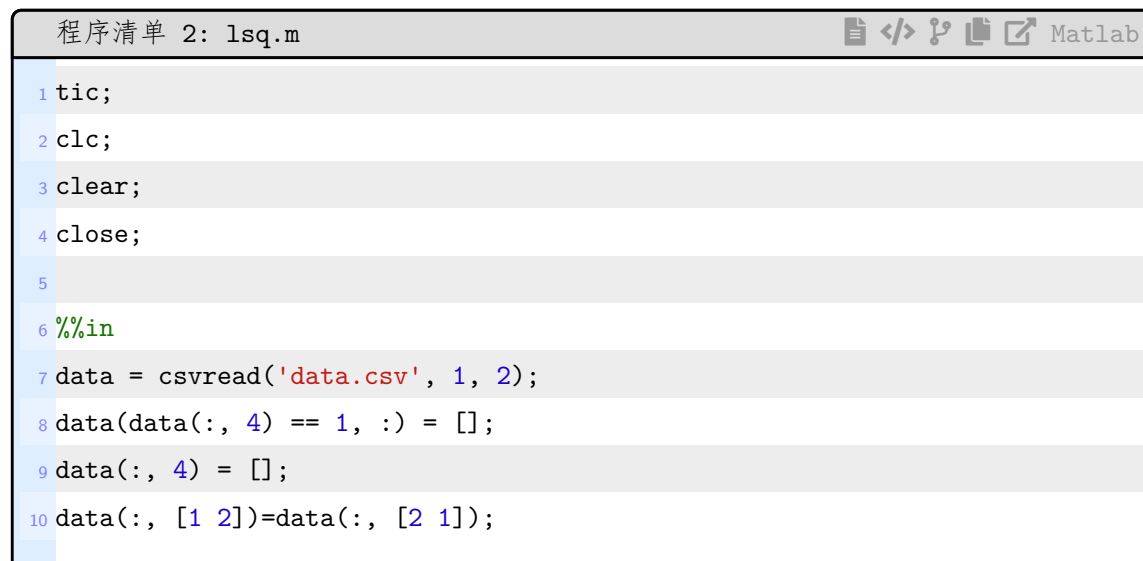


图 4: 运行界面




```
11 name = {" 截距", " 路程/km", " 时间/h", " 注重舒适程度", " 可支配收入/RMB",  
    ↪ " 自付|家庭报销", " 票价/RMB", " 注重时间成本"};  
12 order = string([name strcat(name{2}, strcat('*', name{6}))  
    ↪ strcat(name{7}, strcat('*', name{8})))]);  
13  
14 %%process  
15 [b bint r rint stats]=regress(data(:, 1), [ones(length(data(:, 1)), 1),  
    ↪ data(:, 2), data(:, 3), data(:, 4), data(:, 5), data(:, 6), data(:,  
    ↪ 7), data(:, 8), data(:, 2) .* data(:, 6), data(:, 7) .* data(:, 8)]);  
16 [absb, pos] = sort(abs(b), 'descend');  
17 B = b(pos, :);  
18 order = order(pos, :);  
19 order = [order num2str(B)];  
20 order = [string((1:length(data(1, :)) + 2)') order];  
21 xlswrite('order.xlsx', order);  
22  
23 %%in  
24 data = csvread('data2.csv', 1, 2);  
25 data(data(:, 4) == 1, :) = [];  
26 data(:, 4) = [];  
27 data(:, [1 2])=data(:, [2 1]);  
28 data(:, 1) = 1;  
29 scale = size(data);  
30 data(:, scale(2) + 1) = data(:, 2) .* data(:, 6);  
31 data(:, scale(2) + 2) = data(:, 7) .* data(:, 8);  
32  
33 %%process  
34 y = data * b;  
35 Y = string(num2str(y));  
36 Y(y(:) >= .5)=" 高铁";  
37 Y(y(:) < .5)=" 火车";
```

```
38 xlswrite('predict.xlsx', Y);  
39  
40 %%out  
41 fprintf('Running time is %f second.\n', toc)
```

MATLAB Command Window

Running time is 5.055482 second.
»

图 5: 运行界面

程序清单 3: map.m

```
1 tic;  
2 clc;  
3 clear;  
4 close;  
5  
6 %%in  
7  
8 %% 载入地图数据  
9 sheng=shaperead('bou2_4p.shp', 'UseGeoCoords', true);% 省  
10 %% 使用 importdata 向导导入 2011 年全国 31 个省的数据  
11 d=importdata('ratio.csv', ',');  
12 d.textdata(1, :)=[];  
13 d.textdata(:, 2)=[];  
14 data=d.data;  
15 textdata=d.textdata; % 相对应的省的名称  
16  
17 %%process  
18 %% 定义地图参数  
19 % 针对不同省份, 分别设置不同的颜色 (FaceColor)
```

```
20 % 定义颜色
21 k=128;
22 mycolormap=summer(k);
23 % 生成不同区域按大小的颜色，按照人口数目多少分别指定不同的颜色
24 % 人口越多，颜色越突出
25 geoname={sheng.NAME}';
26 max_data = max(data);
27 n=length(data);
28 mysymbolspec=cell(1,n); % 预定义变量可以加快处理速度
29 for i=1:n
30     count=data(i);
31     mycoloridx=floor( k * count / max_data );
32     mycoloridx(mycoloridx<1)=1;
33     myprovince=textdata{i};
34     geoidx=strmatch(myprovince, geoname);
35     if numel(geoidx) > 0
36         province_name=geoname( geoidx(1) );
37         mysymbolspec{i} = {'NAME', char(province_name),
↪ 'FaceColor', mycolormap( mycoloridx, :) };
38     end
39 end
40 %% 显示地图
41 figure
42 ax=worldmap('china'); % 使用 worldmap 的坐标轴作图
43 setm(ax,'grid','off') % 关闭 grid
44 setm(ax,'frame','off') % 关闭边框
45 setm(ax,'parallellabel','off') % 关闭坐标轴标记
46 setm(ax,'meridianlabel','off') % 关闭坐标轴标记
47
48 % 最关键的两个语句
```

```
49 symbols=makesymbolspec('Polygon',{'default','FaceColor',[0.9 0.9 0.8],  
    ↳ 'LineStyle','--','LineWidth',0.2, 'EdgeColor',[0.8 0.9 0.9]},  
    ↳ mysymbolspec{:});  
50 geoshow(sheng,'SymbolSpec',symbols); % 此处用 mapshow 投影会不正确  
51  
52 %% 图的标注  
53 % 在图像右侧显示 bar  
54 colormap(summer(k));  
55 hcb=colorbar('EastOutside');  
56 step=round(max_data/11);  
57 set(hcb,'YTick',(0:.1:1));  
58 set(hcb,'YTickLabel',num2cell(0:step:max_data));  
59  
60 %%out  
61 fprintf('Running time is %f second.\n', toc)
```

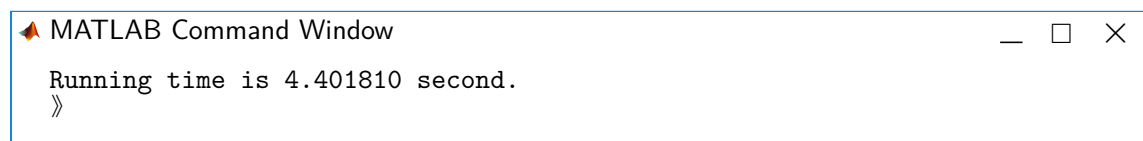


图 6: 运行界面