居民出行调查中交通小区划分方法的改进

赵锦焕 李文权

东南大学，交通学院，南京 210096

摘 要：为了减少居民出行调查的工作量和费用，利用以往调查内容中的年龄、职业、收入和车辆拥有率等指标，根据聚类分析的等价矩阵模糊聚类方法，通过MATLAB编程进行聚类分析，用F-统计量确定最佳模糊聚类类数，对交通小区重新分类。在调查时，只需在每类小区选择一个交通小区，就可满足调查精度要求。最后用模糊C-均值聚类方法验证。计算结果表明，模糊聚类分析方法可以应用在居民出行调查的交通小区划分中。

关键词：交通小区；模糊聚类；出行调查；类数

中图分类号：U491.1+21 **文献标识码**：A **文章编号**：1672-4747(2009)02-0110-06

Improvement of the Traffic District Partition in Resident Trip Investigation

ZHAO Jin-huan LI Wen-quan

Transportation College, Southeast University, Nanjing 210096, China

Abstract: In order to reduce the workload and cost in resident trip investigation, data including age, profession, incomes and vehicle ownership rate of the late survey were used. According to the equivalent matrix fuzzy of cluster analysis, and with MATLAB programming, the optimal classification numbers were determined using F-statistics, and the traffic districts were re-classified. By this way, only a traffic district was chosen from each type traffic zone, the accuracy will meet the demands when conduct a traffic investigation. To verify the method, a fuzzy C-means cluster was used. The results showed that the fuzzy cluster analysis method can be conveniently applied to the partition of traffic zone in resident trip investigation.

收稿日期：2008-05-05.

基金项目：国家高科技研究发展计划项目（2007AA11Z200）

作者简介：赵锦焕（1985-），男，汉族，浙江丽水人，东南大学交通学院硕士研究生，主要研究方向：交通运输规划与管理。

Key words: Traffic districts, fuzzy cluster analysis, trip investigation, classification number

0 引　言

城市居民出行调查是以掌握城市居民的个体出行特征，了解城市交通主体和人的出行活动规律为主要目的的调查活动，是交通规划中需要收集的基础资料之一。居民出行调查的内容主要包括每次出行的资料（如起点、终点、出行目的、出行方式、出行起始时间、出行到达时间、出行距离）和个人基本资料（性别、职业、年龄、收入、居住地等情况）[1]。

进行交通规划管理时需要全面了解交通源之间的交通流，但交通源太多，不可能对每个交通源进行单独研究。因此，在交通规划时，需要将交通源合并成若干小区。交通小区划分恰当与否将直接影响到交通调查、分析、预测的工作量和精度。交通小区划分越小，调查准确度越高；但是，交通小区划分过小，调查工作量极大，成本很高。如何利用现有的统计资料（如收入、职业等），或对于已经进行过居民出行调查的城市，如何利用以往调查的数据，在保证调查精度的前提下，尽可能减少交通小区调查数目，从而可以减少居民出行工作量就成为难点。文献[2]、[3]提出系统聚类法在交通小区划分中的应用，但未计算最佳聚类类数，只是依靠经验确定最后分类结果，人为主观因素较多，可能误差较大，为此本文提出用F-统计量方法来确定最佳聚类类数，并用模糊C-均值聚类方法对聚类结果进行验证。

1 交通小区划分的模糊聚类方法

1.1 交通小区划分的依据

随着城市新一轮的房地产开发，不同收入阶层的居民在城市空间内重新选择居住地区，造成城区内不同地区的居民出行特征差异日趋明显。低收入者偏向于公共交通或者骑自行车，收入较高的出行者则倾向于选择出租车和私人小汽车。年龄和职业等都对出行时间和出行方式的选择产生重要影响[4], [5]，因此，可以依靠在不同交通小区内年龄结构、收入、车辆拥有率、职业、出行方式等存在差异，对各个交通小区重新分类，减少调查工作量。

聚类分析取意于“人以群分，物以类聚”的俗语，即按确定标准对客观事物进行分类的数学方法[6]。由于现实的分类往往伴随着模糊性，所以，用模糊理论来进行聚类分析会显得较自然，更符合客观实际，这也就是模糊聚类分析。本文即是基于模糊聚类分析的基础，采用基于模糊等价关系的模糊聚类分析和模糊C-均值聚类分析来讨论城市居民出行调查中交通小区改进问题。

1.2 基于等价矩阵的模糊聚类方法

基于模糊等价关系的聚类分析是依据客观事物间的特征、亲疏程度和相似性，通过建立模糊等价关系对客观事物进行分类的数学方法[7]。

（1）数据变换：设被分类的样本域为 $X=\{x_1,x_2,\cdots,x_n\}$，x_i 所对应的因素向量为 $(x'_{i1},x'_{i2},\cdots,x'_{im})$，为了便于比较或消除量纲的影响，在作聚类之前首先要对数据进行变换。变换的方法有中心化变换、标准化变换、极差标准化变换、极差正则化变换、对数变换等。

最常用的标准化变换为：

$$x_{ij}=\frac{x'_{ij}-\overline{x}_j}{S_j}\qquad i=1,2,\cdots,n\qquad j=1,2,\cdots,m$$

式中，$\overline{x}_j=\frac{1}{n}\sum_{i=1}^{n}x'_{ij}$；$S_j=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x'_{ij}-\overline{x}_j)^2}$，$j=1,2,\cdots,m$。

（2）标定：即给出被分类对象间相似程度的统计量 $r_{ij}(1\leqslant i,\ j\leqslant n)$，从而确定出模糊相似矩阵 $\boldsymbol{R}=(r_{ij})_{n\times n}$。$r_{ij}$ 可根据具体情况按不同方法给出，较常用的如绝对值减数法、数量积法等。本文采用绝对值减数法，即：

$$r_{ij}=1-c\left(\sum_{k=1}^{m}\left|x_{ik}-x_{jk}\right|\right)$$

式中，c 是适当选取的常数，应使 r_{ij} 在 [0，1] 中分散开，目的是用 0 到 1 之间的数值充分凸显它们的差异。

（3）模糊等价矩阵：由标定所得到的矩阵 $\boldsymbol{R}$ 一般为相似矩阵，不满足传递性，故本文采用平方自合成法，求出 $\boldsymbol{R}$ 的传递闭包 $t(\boldsymbol{R})$，则 $t(\boldsymbol{R})$为模糊等价矩阵。具体方法如下：

设 $\boldsymbol{R}_1=(a_{ik})_{m\times s}$，$\boldsymbol{R}_2=(b_{kj})_{s\times n}$，$\boldsymbol{R}=\boldsymbol{R}_1\circ\boldsymbol{R}_2=(c_{ij})_{m\times n}$，则：

$$c_{ij}=\bigvee_{k=1}^{s}(a_{ik}\wedge b_{kj})\qquad a_{ik}\wedge b_{kj}=\min(a_{ik},b_{kj})$$
$$a_{ik}\vee b_{kj}=\max(a_{ik},b_{kj})。$$

计算 $\boldsymbol{R}^2=\boldsymbol{R}\cdot\boldsymbol{R}$，$\boldsymbol{R}^4=\boldsymbol{R}^2\cdot\boldsymbol{R}^2$，$\cdots$，$\boldsymbol{R}^{2^k}=\boldsymbol{R}^{2^{k-1}}\cdot\boldsymbol{R}^{2^{k-1}}$，$k=1,2,\cdots$。当 $\boldsymbol{R}^{2^k}=\boldsymbol{R}^{2^{k-1}}$ 时，停止计算，取 $t(\boldsymbol{R})=\boldsymbol{R}^{2^k}$。

（4）模糊聚类：设定不同的置信水平$\lambda\in[0，1]$，对模糊等价矩阵 $t(\boldsymbol{R})$ 进行聚类处理。若 $t(\boldsymbol{R})$ 中的元素 $c_{ij}\geqslant\lambda$，则将集合 X 中的 X_i 和 X_j 归为一类，从而得到一系列分类结果。

1.3　模糊聚类类数的确定

由 1.2 可知：对于不同的λ，得到不同的分类结果，这对全面了解样本的分类情况是比较形象和直观的。到底应该把样本分为几类是一个比较困难的问题，因为分类问题本身就没有一定的标准，人们可以从不同角度给出不同的分类结果。本文根据 F- 统计量来确定最佳分类个数，该方法是来自统计的方差分析思想，具体如下：

在某一聚类结果中，设分类数为 r，第 j 类的样本数为 n_j，对应的样本为：x_1^j，x_2^j，$\cdots$，$x_{n_j}^j$，其聚类中心向量为 $\overline{x}^j=(\overline{x}_1^j,\overline{x}_2^j,\cdots,\overline{x}_m^j)$。

作 F 统计量：

$$F=\frac{\dfrac{\sum n_j\left\|\overline{x}^j-\overline{x}'\right\|^2}{r-1}}{\dfrac{\sum_{j=1}^{r}\sum_{i=1}^{n_j}\left\|x_i^j-\overline{x}^j\right\|^2}{n-r}}\qquad(1)$$

式中，$\overline{x}'=(\overline{x}_1,\overline{x}_2,\cdots,\overline{x}_m')$；$\overline{x}'=\dfrac{1}{n}\sum_{i=1}^{n}x_{ik}'$ $(k=1,2,\cdots,m)$

$\left\|\overline{x}^j-\overline{x}'\right\|$ 为 $\overline{x}^j$ 与 $\overline{x}'$ 的欧式距离；$\left\|x_i^j-\overline{x}^j\right\|$ 为第 j 类样本 x_i^j 与中心 $\overline{x}^j$ 的欧式距离。易知：$F\sim F(r-1,n-r)$，这里 $F(r-1,n-r)$ 是指自由度为 $r-1$、$n-r$ 的 F- 分布。（1）式分子表征类与类之间的距离，分母表征类内样本间的距离。因此，F 值越大，说明类与类之间的距离就越大，也就是类与类间的差异越大，此时对应的分类是合适的。

从而本文采用如下判断准则：

若 F 的观测值 F_r 满足：$F_r>F_a(r-1,\ n-r)$，则判定对应的分类是合适的。

若多于一个 F_r 使上式成立，那么，计算差值 $F_r-F_\alpha(r-1,\ n-r)$，记 $F_z-F_\alpha=\max_r(F_r-F_\alpha)$，则可判定对应于 F_z 的分类是合适的。

1.4　模糊 C-均值聚类方法

模糊 C-均值聚类方法（FCM）是指把聚类归结成一个带约束的非线性规划问题，通过优化求解获得数据集的模糊划分和聚类。在基于目标函数的聚类算法中，模糊 C- 均值类型算法的理论最为完善，应用最为广泛。

设被分类的样本域为 $X=\{x_1,x_2,\cdots,x_n\}$，$x_i=(x_{i1},x_{i2},\cdots,x_{im})$，退化 C- 模糊矩阵为：

$$\boldsymbol{A}=\begin{pmatrix}A_1\\ \vdots\\ A_c\end{pmatrix}=\begin{pmatrix}a_{11}&\cdots&a_{1n}\\ \vdots&&\vdots\\ a_{c1}&\cdots&a_{cn}\end{pmatrix},$$

令 $v_i=\dfrac{\sum_{j=1}^{n}(a_{ij})^r x_j}{\sum_{j=1}^{n}(a_{ij})^r}$ 为类 A_i 的聚类中心。

$J(A,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}(a_{ij})^{\mathrm{T}}\left\|x_j-v_i\right\|$ 为目标函数，在退化的 C- 模糊划分空间中，目标函数的最优化问题是可解的，而且当 $r\geqslant1$，$x_j\neq v_i$ 时的最优解的迭代算法即是模糊 C- 均值聚类算法。其步骤如下：

（1）对分类数 $c(2<c<n)$，取 c 个初始聚类中心向量：

$$\boldsymbol{V}^{(0)}=\begin{pmatrix}v_1^{(0)}\\ \vdots\\ v_c^{(0)}\end{pmatrix}=\begin{pmatrix}v_{11}^{(0)}&\cdots&v_{1m}^{(0)}\\ \vdots&&\vdots\\ v_{c1}^{(0)}&\cdots&v_{cm}^{(0)}\end{pmatrix}$$

（2）按 $a_{ij}^{(l+1)}=\left(\sum_{k=1}^{c}\left(\frac{\left\|x_j-v_i^{(l)}\right\|^2}{\left\|x_j-v_k^{(l)}\right\|^2}\right)^{\frac{1}{r-1}}\right)^{-1}$（$j=1,2,\cdots,n$，$i=1,2,\cdots,c$）迭代计算 C- 模糊划分矩阵 $\boldsymbol{A}^{(l)}$。

（3）计算第 l 步的聚类中心矩阵：

$$\boldsymbol{V}^{(l)}=\begin{pmatrix}v_1^{(l)}\\ \vdots\\ v_c^{(l)}\end{pmatrix}=\begin{pmatrix}v_{11}^{(l)} & \cdots & v_{1m}^{(l)}\\ \vdots & & \vdots\\ v_{c1}^{(l)} & \cdots & v_{cm}^{(l)}\end{pmatrix},$$

式中：$v_i^{(l)}=\dfrac{\sum_{j=1}^{n}(a_{ij}^{l})^r x_j}{\sum_{j=1}^{n}(a_{ij}^{l})^r}\quad i=1,2,\cdots,c$

（4）给定满意误差限 $\varepsilon>0$，用一个矩阵范数 $\|\bullet\|$，比较 $\boldsymbol{A}^{(l)}$与$\boldsymbol{A}^{(l+1)}$，若 $\left\|\boldsymbol{A}^{(l+1)}-\boldsymbol{A}^{(l)}\right\|\leqslant\varepsilon$，则停止迭代；否则转第二步继续迭代。

1.5　交通小区划分改进的步骤

本文利用交通基础数据，比如交通小区内成员的年龄结构、收入、车辆拥有率、职业、出行方式等，运用基于等价模糊聚类和 F- 统计法得到最佳分类数，但是，由于该方法在各个步骤中都有不同的算法，都会产生误差，可能得到的结果不是最佳，所以，本文用 FCM 算法对结果进行验证。如果两种方法得到的结果误差很小，说明聚类结果比较符合实际；如果相差比较大，则重新选择分类数进行计算，直到满足误差要求。具体的流程图如图 1 所示：

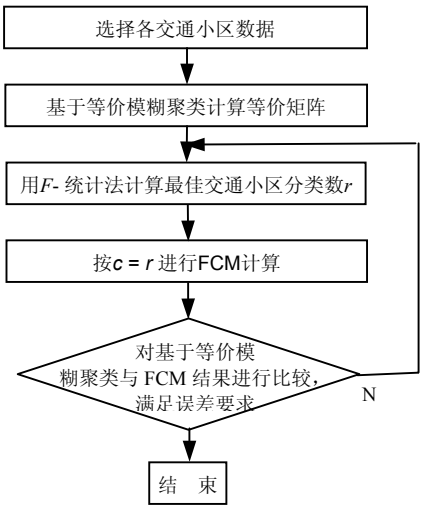

图 1　交通小区划分改进方法流程

Fig.1　Process of the improvement in traffic district partition

2　实例分析

本文以某市八个小区作为研究对象，分别取各小区内老年人数（年龄大于 60 岁）、小学生人数、低收入人数（经济收入小于 1 万）、高收入人数（大于 10 万）、拥有汽车人数与小区总人数的比值作为指标，具体说明本文提出的方法在居民出行调查中交通小区划分中的应用。

2.1　采用模糊等价聚类分析

（1）根据以往居民出行调查的数据，得到各个小区各指标情况如表 1 所示。

表 1　各交通小区指标值/（%）

Tab.1　Index values of each traffic district/（%）

小区编号	小学生比例	老年人比例	低收入比例	高收入比例	车辆拥有率
1	0.086 0	0.149 1	0.047 4	0.128 1	0.203 5
2	0.047 9	0.284 4	0.079 9	0.060 7	0.092 7
3	0.046 2	0.250 0	0.037 8	0.050 4	0.098 7
4	0.072 8	0.140 2	0.010 8	0.107 8	0.164 4
5	0.043 1	0.122 5	0.023 2	0.086 1	0.102 7
6	0.028 8	0.143 8	0.070 3	0.035 1	0.111 8
7	0.047 0	0.146 0	0.042 1	0.056 9	0.126 2
8	0.055 8	0.107 0	0.027 9	0.048 8	0.088 4

（2）采用 MATLAB 软件计算求得各变量间的模糊等价矩阵，如表 2 所示。

表 2　模糊等价矩阵值

Tab.2　Values of the fuzzy equivalence matrix

小区编号	1	2	3	4	5	6	7	8
1	1.000 0	0.876 0	0.876 0	0.974 3	0.929 6	0.890 1	0.890 1	0.890 1
2	0.876 0	1.000 0	0.978 8	0.876 0	0.876 0	0.876 0	0.876 0	0.876 0
3	0.876 0	0.978 8	1.000 0	0.876 0	0.876 0	0.876 0	0.876 0	0.876 0
4	0.974 3	0.876 0	0.876 0	1.000 0	0.929 6	0.890 1	0.890 1	0.890 1
5	0.929 6	0.876 0	0.876 0	0.929 6	1.000 0	0.890 1	0.890 1	0.890 1
6	0.890 1	0.876 0	0.876 0	0.890 1	0.890 1	1.000 0	0.914 3	0.914 3
7	0.890 1	0.876 0	0.876 0	0.890 1	0.890 1	0.914 3	1.000 0	0.958 3
8	0.890 1	0.876 0	0.876 0	0.890 1	0.890 1	0.914 3	0.958 3	1.000 0

（3）取不同置信水平λ，得到分类效果如下：

$\lambda = 0.9788$ 时，样本聚为 7 类，即{1}，{4}，{2，3}，{5}，{6}，{7}，{8}；

$\lambda = 0.9743$ 时，样本聚为 6 类，即{1，4}，{2，3}，{5}，{6}，{7}，{8}；

$\lambda = 0.9583$ 时，样本聚为 5 类，即{1，4}，{2，3}，{5}，{6}，{7，8}；

$\lambda = 0.9296$ 时，样本聚为 4 类，即{1，4，5}，{2，3}，{6}，{7，8}；

$\lambda = 0.9143$ 时，样本聚为 3 类，即{1，4，5}，{2，3}，{6，7，8}；

$\lambda = 0.8901$ 时，样本聚为 2 类，即{1，4，5，6，7，8}，{2，3}；

$\lambda = 0.8760$ 时，样本聚为 1 类，即{1，2，3，4，5，6，7，8}。

（4）确定最佳阈值λ，取$\alpha = 0.05$，不同的分类数对应的 F 统计量值如表 3 所示。

表 3　F 统计量值

Tab.3　Values of F-statistics

聚类类数 r	λ 值	F_r	$F_{0.05}(r-1，n-r)$
1	$\lambda = 0.8760$	—	—
2	$\lambda = 0.8901$	$F_2 = 6.86$	$F_{0.05}(1，6) = 5.99$
3	$\lambda = 0.9143$	$F_3 = 7.39$	$F_{0.05}(2，5) = 5.79$
4	$\lambda = 0.9296$	$F_4 = 5.07$	$F_{0.05}(3，4) = 6.59$
5	$\lambda = 0.9583$	$F_5 = 7.18$	$F_{0.05}(4，3) = 9.12$
6	$\lambda = 0.9743$	$F_6 = 6.01$	$F_{0.05}(5，2) = 19.3$
7	$\lambda = 0.9788$	$F_7 = 5.55$	$F_{0.05}(6，1) = 23.4$

从表 3 可知，仅当$\lambda = 0.8901$ 和$\lambda = 0.9143$ 时，满足 $F_r > F_{0.05}(r-1,n-r)$，并且 $F_2 - F_{0.05}(1，6) = 6.86 - 5.79 = 1.07$，$F_3 - F_{0.05}(2,5) = 7.39 - 5.79 = 1.60$，故最佳阈值取$\lambda = 0.9143$。把原样本聚成三类：{1，4，5}，{2，3}，{6，7，8}，只需在每类小区中选择一个小区作为调查对象。调查小区数量从 8 个变为 3，既能满足精度要求，又能减少一半的工作量。

2.2　采用模糊 C- 均值聚类分析

由 2.1 分析可知：将原样本分三类比较合理，故此法中取分类数 $c = 3$，$r = 2$，误差极限$\varepsilon = 0.000\ 01$，采用 MATLAB 软件编程，分类结果和各聚类中心如表 4 所示。

表 4　分类结果和各聚类中心

Tab.4　Classification results and the each clustering center

分类结果	小学生	年龄大于 60	低收入	高收入	车辆拥有率
{1，4}	0.0791	0.1448	0.0306	0.1175	0.1838
{5，6，7，8}	0.0445	0.1296	0.0399	0.0573	0.1075
{2，3}	0.0470	0.2266	0.0595	0.0557	0.0568

由表 4 可知，原样本聚类结果为：{1，4}，{2，3}，{5，6，7，8}，显然，这与 2.1 的聚类结果（{1，4，5}，{2，3}，{6，7，8}）存在差异。根据统计学知识可知，采用不同的聚类方法会得到不同的分类结果，对任何观测数据都没有唯一正确的分类方法，本文两者分类方法误差的存在是可接受的。

3　结束语

居民出行调查在交通规划和城市综合及专项交通规划中扮演着极其重要的角色，而且投入经费较多。针对城市各个小区间存在的联系与差异，在居民出行调查时，利用基于模糊聚类方法，把交通小区重新分类，再从每一类小区选择一个小区作为调查对象，结果表明交通调查工作量可以大大减少。另外，由表 4 中λ的取值（从 0.8760 到 0.9788）可以看出，各个小区间存在着比较大的相关性，符合我国的基本国情。

文献[3]是按城市的土地利用这一种影响因素进行聚类分析，对于最佳聚类分类数，只是通过主观经验确定，没有经过科学计算，分类的结果可能不是最佳。本文考虑多种影响因素（如收入、职业等），通过 F- 统计量确定最佳聚类数，利用 C- 均值聚类方法对聚类结果进行验证。由于居民出行调查内容包含很多数据[8]，如何选择已有居民出行调查的数据进行聚类分析，使得过程比较简单，结果比较符合实际是下一步研究的重点。

参考文献

[1] 王　炜. 城市交通管理规划指南[M]. 北京：人民交通出版社，2003.

[2] 桂小玲，勒文舟，胡郁葱. 模糊聚类分析方法及其在交通规划中的应用[J]. 交通与计算机，2005，2：80-82.

[3] 杨　波，刘海洲. 基于聚类分析的交通小区划分方法的改进[J]. 交通规划，2007.7.

[4] 曲大义，于仲臣，等. 苏州市居民出行特征分析及交通发展对策研究[J]. 东南大学学报（自然科学版），2001，5：118-122.

[5] 李　民. 基于活动链的居民出行行为分析[D].吉林：吉林大学，2004.

[6] 谢季坚，刘承平. 模糊数学方法及其应用[M]. 武汉:华中科技大学出版社，1999.

[7] 高新波. 模糊聚类分析及其作用[M]. 西安:西安电子科技大学出版社，2004.

[8] 王　瑞. 城市居民出行调查若干问题研究[D]. 西安:长安大学，2006.

（中文编辑：刘娉婷）

上接第 84 页

因此，成渝高速公路 K48+000 至 K52+000 原有沥青混合料的阿布森法沥青回收试验中后续加热时间确定为 30 min。

4　结　论

本文通过对成渝高速公路 K48+000 至 K52+000 原有沥青混合料的沥青抽提、回收试验，对原有的阿布森沥青回收试验提出了一些改进，通过试验对比得出以下结论：

（1）在试验过程中，从加热开始即向蒸馏瓶中通入少量 CO_2，当抽提液温度达到 157°C 以后增大 CO_2 的通气量。试验证明，增大 CO_2 的通入量能有效地控制高温情况下沥青的二次老化现象。

（2）通过对“后续加热时间”的规定，简化了原有规程，使得试验更具操作性；找到了三氯乙烯被完全蒸馏的最佳时间点，提高了试验成功的可靠性。

因此，采用改进后的阿布森法进行沥青回收试验能有效地保证回收沥青的准确性和可靠性。

参考文献

[1] JTJ 052-2000. 公路工程沥青及沥青混合料试验规程[S].

[2] 张　建，肖　维，黄晓明. 沥青路面再生中旧沥青的回收与再生研究[J]. 湖南交通科技，2006，32（4）.

[3] 何文峰，王　欣，刘先淼. 阿布森沥青回收试验探讨[J]. 中南公路工程，2005，30（3）.

[4] 熊出华，张永兴等. 一种新的沥青回收方法探讨[J]. 中外公路，2006，26（2）.

（中文编辑：吴继屏）