

基于主成分贡献度的道路事故热点成因分析

曾令秋¹, 王瑞梅¹, 韩庆文², 曾 孜³, 朱颖祥², 张 程¹

(1.重庆大学计算机学院, 重庆 400044; 2.重庆大学通信工程学院, 重庆 400044; 3.重庆市第七中学校, 重庆 400030)

摘要: 事故热点成因分析是解决交通安全问题的关键和难点, 不同成因的事故热点直接决定驾驶人或自动驾驶的通过策略。针对DTH3N算法聚类得到的事故热点展开研究, 提出一种基于主成分贡献度的道路事故热点成因分析方法。通过对5类事故热点成因因素: 道路、行人、车辆、环境和管制因素贡献度的分析, 提取事故热点道路物理成因和社会成因的影响权值。基于英国STATs19数据库选取5个测试区域进行实验分析, 结果表明, 由事故热点成因分析方法获取的输出参数能够合理解释事故热点成因, 可有效指导驾驶行为决策和优化交通管制。

关键词: 交通安全; 道路事故热点; 主成分分析; 内特性分析; 道路物理成因; 社会成因

Traffic Accident Causation Analysis at Hotspots

Zeng Lingqiu¹, Wang Ruimei¹, Han Qingwen², Zeng Zi³, Zhu Yingxiang², Zhang Cheng¹

(1.College of Computer Science, Chongqing University, Chongqing 400044, China; 2.College of Communication Engineering, Chongqing University, Chongqing 400044, China; 3.Chongqing No.7 High School, Chongqing 400030, China)

Abstract: The key to and difficulty in solving traffic safety problems lie in the causes of accident at hotspots. The accident hotspots caused by different factors determine the operation strategies for drivers and automatic vehicles. Through researching the accident hotspots with the clustering DTH3N algorithm, this paper proposes an analytical method to identify the causes of road accident hotspots based on main accident contributing factors. By analyzing the contributions of five types of accident causation factors at hotspots, namely road, pedestrian, vehicle, environment, and regulation, the paper calculates the influencing weights of both road physical and social factors for accident hotspots. The paper conducts an experimental analysis on five selected testing areas from the British STATs19 database. The results show that the output parameters obtained from the analytical method on accident hotspot serve as a reasonable explanation for the causes of accident hotspots, and can provide effective guidance in driving behavior decision-making as well as in optimizing traffic control.

Keywords: traffic safety; road accident hotspot; principal component analysis; internal characteristic analysis; road physical factors; social factors

收稿日期: 2016-07-28

基金项目: 国家自然科学基金项目“基于车间作用关系的安全预警信息传输控制策略研究”(61601066)、重庆市研究生科研创新资助项目“基于模型学习的公交线路驾驶行为评价移动App”(CYS16027)

作者简介: 曾令秋(1975—), 男, 重庆人, 博士, 副教授, 主要研究方向: VANET、数字图像处理、计算机图形学与虚拟现实。E-mail: zenglq@cqu.edu.cn

0 引言

事故热点是影响道路通行时间和拥堵状况的重要因素, 其产生原因复杂, 例如, 由于地形因素客观成为事故热点区域(南美玻利维亚北永加斯路被称为“死亡之路”)、由于高车流密度和客流量成为事故热点区域

的闹市区。对驾驶人或自动驾驶决策系统而言, 不同成因的事故热点的决策方案必然不同。因此, 从提升道路交通安全的角度来看, 需要为驾驶人或自动驾驶决策算法提供与事故热点相关的内特性信息, 即需要发掘事故热点成因并以此为依据提供相关的控制决策信息。

热点成因识别是一个历史数据分析问

题,即通过分析海量历史数据,发现对象热点中事故的生成共性因素,并以此为依据识别热点成因。目前,中国描述交通事故的数据局限于某一城市或某一路段的统计,并没有公开的权威数据库可以提供相应的数据。英国国家数据库 STATs19^[1]记录了自1926年起英国发生的所有交通事故,并通过82个属性描述单个事故,STATs19是迄今为止英国记录最详细、最完整、最可靠的提供道路交通碰撞信息资源的数据库,并被广泛应用于道路交通安全评估。由于该数据库能够为热点分析提供充足的事故数据,本文基于该数据库开展研究。

从STATs19数据库82个属性变量中发掘成因因素需要进行降维,本文采用较为经典的降维法——主成分分析法实现事故热点分析。文献[2]采用主成分分析法评估高速公路交通安全状况。文献[3]将主成分分析纳入道路交通事故预测中,消除一些重叠信息,并结合神经网络对道路交通事故进行预测。文献[4]提出评价区域交通安全的主成分分析模型,展现了主成分分析法在交通数据库分析中的可行性。以上研究皆基于大区(如行政区域)展开,虽然分析结果能在一定程度上反映对象区域的特性,但是区域特性可能包含多个事故热点,热点成因也有差异,分析结果难以体现事故热点本身的特性,也难以生成针对特定事故热点的合理控制决策。

热点识别是成因分析的基础,在文献[5]前序研究中笔者提出一种基于自然最近邻聚类的改进算法DTH3N对事故区域的历史事故点进行识别,本文将DTH3N算法识别的事故热点为对象,探索一种新的事故热点成因分析方法。

1 概念阐述

1.1 事故热点

事故热点指具有某种关联特性的交通事故发生位置点的集合,即交通事故多发区域。区域内事故成因有内在关联性,区域具有不规则边界,区域的覆盖范围与道路本身特征(例如交叉口等)有直接关系。

热点是聚类的结果。聚类就是将需要处理的整个数据集划分成多个不同的类簇,类簇与类簇之间距离或者相异性尽量大,使得类簇内部尽量紧凑^[6]。常用的聚类算法有:1)基于距离的K-means, K-medoid 和 Nearest Neighbor Hierarchical Clustering 等, 这些方

法聚合的空间对象是欧氏距离而非网络距离,聚合的空间对象形成的区域形状受限;2)基于密度的DBSCAN和KDE等聚类算法,可以发现任意形状的簇类,但对阈值的设置比较敏感,例如半径阈值和簇密度阈值;3)自然最近邻聚类算法(3N),克服了以上算法存在的问题,但处理大面积区域时面临热点定位困难的问题^[5]。

本文采用文献[5]中提出的一种基于自然最近邻聚类的改进算法DTH3N,并结合具有区域约束等优点的3N聚类方法来识别事故热点。图1为多事故点按DTH3N算法聚类生成事故热点的示例,图中黑色圆点表示事故点,而多个相连接的事故点构成一个事故热点。

1.2 事故热点成因

事故热点成因具有多样化特征,可归结为两个大类:道路物理成因和社会成因。道路物理成因指导致事故多发区域生成的道路因素,例如路况、道路设计和区域环境。社会成因指导致事故多发区域生成的社会因素,例如区域治安条件、人口素质和人口密度。

1.3 事故热点成因因素

事故热点成因因素指影响事故发生的因素,本文根据STATs19数据库提供的属性参数定义5个因素:道路因素 X_1 、行人因素 X_2 、车辆因素 X_3 、环境因素 X_4 和管制因素 X_5 (见表1)。

1.4 事故热点成因因素贡献权重

根据事故热点成因因素与道路物理成因和社会成因之间的贡献关系设定贡献权重,以此为依据构造道路物理成因矩阵 CR 和社会成因矩阵 CS 。表1中道路因素 X_1 仅与道

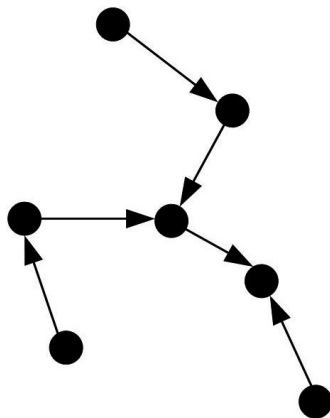


图1 事故热点与事故点的关系

Fig.1 Relationship between accident hotspots and accident points

路物理成因相关，因此设置 X_1 对应的 $CR=1$ 。相较于道路特性，区域的社会因素如犯罪率、人口密度等更容易获取，因此本文从社会因素入手计算热点成因贡献权重。为了充分表现社会因素与道路事故的关联性，选择高社会因素区域和低社会因素区域进行对比计算，具体方法如下：

1) 对比区域选取。根据人口密度和犯罪率两个社会因素，分别选择高人口密度和高犯罪率、低人口密度和低犯罪率区域作为对比区域。

2) 典型道路区域选取。分别从对比区域内筛选高事故发生率主要道路作为典型道路区域。

3) 分别统计典型道路区域的 X_1 ， X_2 ， X_3 ， X_4 ， X_5 对应数据库指标涉及的总事故数 A 。

4) 根据人口密度和犯罪率分别计算选择区域的社会因子

$$F_s(i) = \frac{Density(i)}{Density_{MAX}} + \frac{Crime(i)}{Crime_{MAX}}, \quad (1)$$

式中： $Density(i)$ 为典型道路区域 i 的人口密

度/(人·km²)； $Density_{MAX}$ 为所有典型道路区域中人口密度最大值/(人·km²)； $Crime(i)$ 是典型道路区域 i 的犯罪率； $Crime_{MAX}$ 为所有典型道路区域中犯罪率最高值。

5) 计算两个典型道路区域的 A 比值和 F_s 比值，获得社会因素贡献权重。其中， A 比值为不同区域中事故数的比值，即 $AccNum(i)/AccNum(j)$ ； F_s 比值为不同区域社会因子的比值，即 $F_s(i)/F_s(j)$ 。

在此选择 Islington 区域和 Sutton 区域作为对比区域，根据公式(1)可得 $F_s(Islington)=1.91$ ， $F_s(Sutton)=0.77$ 。选取 Islington 的 A501 道路区域和 Sutton 的 Green Wrythe Lane 道路区域进行事故数统计，统计时段为 2011—2015 年，统计结果见表 2。

表 2 典型道路区域事故数据统计

Tab.2 Data statistics of accidents in typical road areas

影响因素	X_2	X_3	X_4	X_5
A501	940	900	232	1 010
Green Wrythe Lane	220	680	135	493

表 3 CS 值

Tab.3 CS value

影响因素	X_2	X_3	X_4	X_5
计算值	0.58	0.53	0.69	0.82
实际取值	0.60	0.50	0.70	0.80

表 4 事故热点成因因素贡献权重

Tab.4 Contributing weight of causation factors at accident hotspots

影响因素	X_1	X_2	X_3	X_4	X_5
CR	1.0	0.4	0.5	0.3	0.2
$CS^{(1)}$	0	0.6	0.5	0.7	0.8

1) 本文社会因子仅考察区域人口密度和犯罪率，也可加入其他参数修订准确度，例如受教育程度、人口年龄分布等参数。

表 5 测试对象区域人口密度和犯罪率

Tab.5 Population density and crime rate in the surveyed areas

对象区域	人口密度	犯罪率
Islington	非常高	较高
Wandsworth	较高	较低
Hillingdon	非常低	高
Westminster	较高	非常高
Sutton	较低	较低

表 1 事故热点成因因素

Tab.1 Causation factors of accident hotspots

影响因素	STATs19 数据库属性参数
道路因素 (X_1)	路面类型
	交叉口类型
	路面状况
	现场特殊情况
	交叉口位置
车道障碍区(Carriageway Hazards, 例如中间隔离带、交通警示桩等)	
行人因素 (X_2)	人行道道路维修
	人行横道物理设施
	行人位置
	行人移动状态
车辆因素 (X_3)	车型
	拖车(Towing and Articulation)
	车辆操纵
环境因素 (X_4)	光照条件
	天气情况
管制因素 (X_5)	人为控制
	交叉口人行横道控制
	日期
	事故时间点

社会因素权重 CS 表现为社会因子 F_s 对相关事故数的影响程度：

$$CS=\frac{F_s(i)AccNum(j)}{F_s(j)AccNum(i)} \tag{2}$$

计算得到 CS 值与本文中实际取值如表 3 所示，进而可得事故热点成因因素贡献权重(见表 4)。

2 事故热点成因因素计算方法

计算事故热点成因因素是识别热点成因的关键，计算方法为：

第一步：根据表 1 对 STATs19 属性参数进行整合计算，获得 X_1, X_2, X_3, X_4, X_5 。

第二步：计算 X_1, X_2, X_3, X_4, X_5 的贡献率，此处采用主成分分析法获得，具体过程如下：

1) 构建原变量矩阵。

利用事故热点成因因素 $X_j(j=1,2,\cdots,5)$ 构造原变量矩阵；假设热点中包含 n 个事故点，基于此构建参数矩阵

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{15} \\ x_{21} & x_{22} & \cdots & x_{25} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{n5} \end{bmatrix} \tag{3}$$

2) 构建参数矩阵 X 的标准归一化矩阵 $\overline{X}$ ，

$$\overline{X}_{ij}=\frac{X_{ij}-X_j}{S_j}, (i=1,2,\cdots,n;j=1,2,\cdots,5)$$

$$X_j=\sum_{i=1}^n X_{ij}/n, (j=1,2,\cdots,5) \tag{4}$$

$$S_j^2=\frac{1}{n-1}\sum_{k=1}^n (x_{kj}-x_j)^2$$

式中： $\overline{X}_{ij}$ 为 $\overline{X}$ 的元素值； X_j 为参数矩阵 X 中列的平均值； S_j 为参数矩阵 X 中每一列的方差。

3) 构建协方差矩阵

$$R=\frac{\overline{X}^T\overline{X}}{n-1} \tag{5}$$

4) 获取协方差矩阵 R 的特征值矩阵 $\lambda_{1\times 5}$ 和特征向量矩阵 $V_{5\times n}$ ，对特征值进行排序使得 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_5$ 。各成分与标准归一化矩阵的线性组合为

$$F_i=V_i\lambda_i, \tag{6}$$

式中： V_i 为协方差矩阵 R 的第 i 个特征向量； λ_i 为所对应的特征值。

5) 提取主成分个数 m ，计算公式为

$$\frac{\sum_{i=1}^m \lambda_i}{\sum_{j=1}^5 \lambda_j} \geq 85\% \tag{7}$$

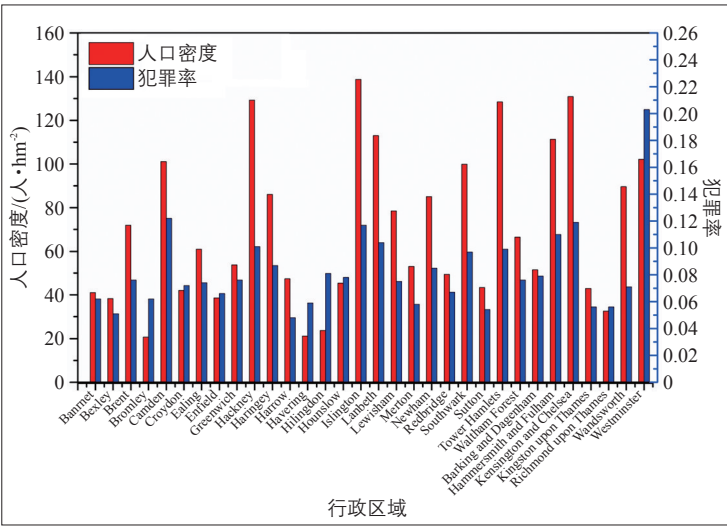


图2 伦敦市各行政区域人口密度和犯罪率
Fig.2 Population density and crime rate in different administrative districts in London

表 6 测试对象区域经纬度范围

目标区域	经度范围	纬度范围
Islington	-0.116403~-0.086439	51.520145~51.537165
Wandsworth	-0.169623~-0.138866	51.419022~51.436691
Hillingdon	-0.424706~-0.393354	51.505523~51.523149
Westminster	-0.214452~-0.180258	51.520533~51.537458
Sutton	-0.244186~-0.213441	51.366352~51.385587

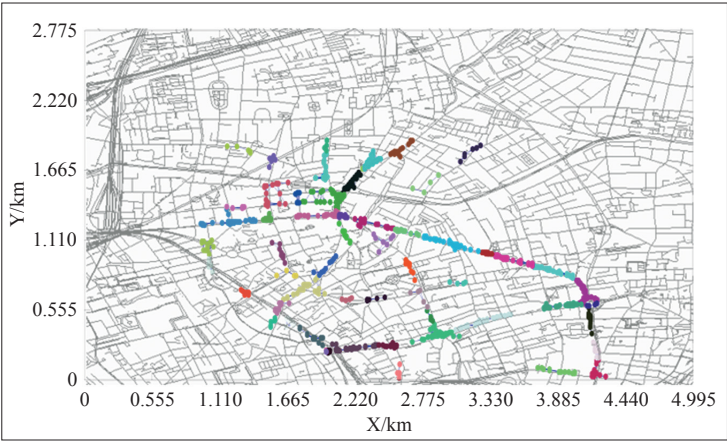
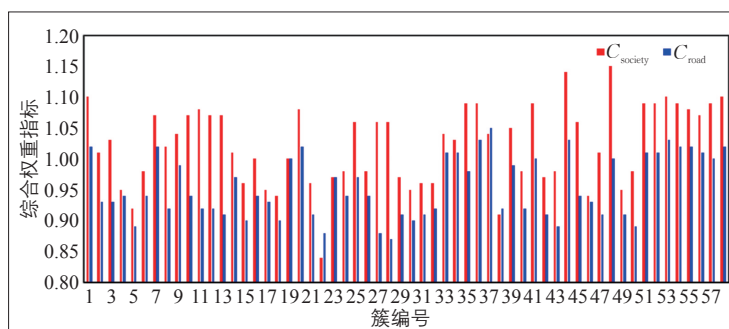


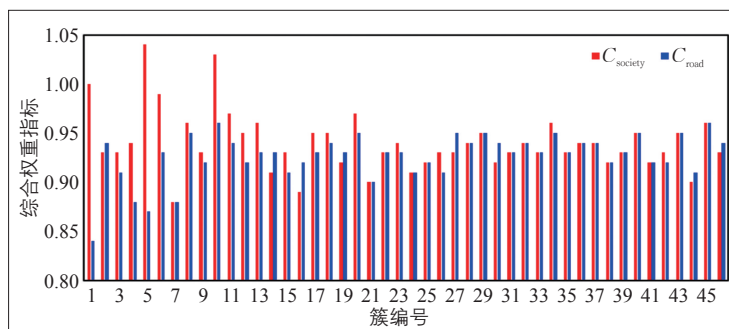
图3 Islington事故热点识别聚类结果
Fig.3 Results of accident hotspot clustering in Islington

表 7 测试对象区域事故热点数量

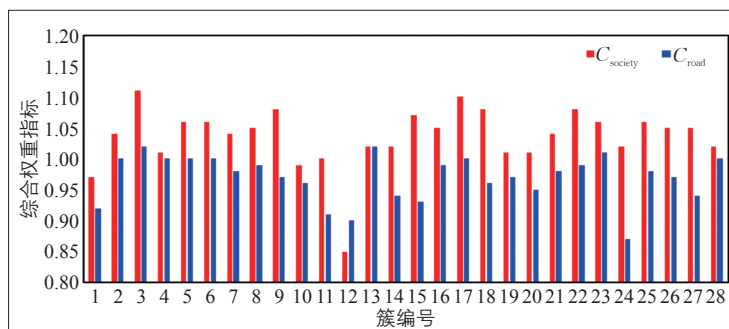
测试对象区域	事故热点数量
Islington	58
Wandsworth	46
Hillingdon	28
Westminster	43
Sutton	17



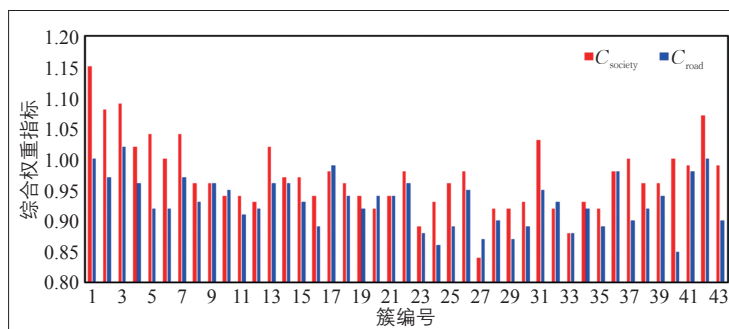
a Islington



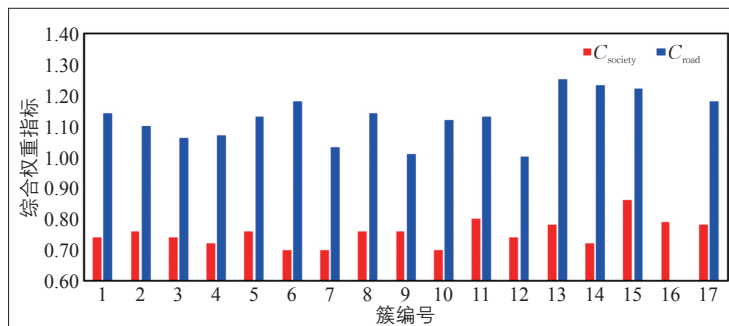
b Wandsworth



c Hillingdon



d Westminster



e Sutton

图4 各测试对象区域的 C_{road} 和 $C_{society}$ Fig.4 C_{road} and $C_{society}$ in different surveyed areas

基于主成分分析, 得到主成分特征值矩阵 $\lambda_{1 \times m} (\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m)$ 所对应的主成分特征向量矩阵 $V_{5 \times m}$ 。

6) 获取主成分特征值的方差贡献矩阵

$$W_{1 \times m} = [w_1, w_2, \dots, w_m],$$

$$w_i = \frac{\lambda_i}{\sum_{j=1}^m \lambda_j} \quad (8)$$

7) 构建事故热点事故影响指标综合权重矩阵

$$F_{5 \times 1} = |V_{5 \times m}| W_{1 \times m}^T, \quad (9)$$

$F_{5 \times 1}$ 矩阵相应的 5 个权重是事故影响参数 X_1, X_2, X_3, X_4, X_5 所对应的贡献率, 即参数变量对事故热点整体的贡献率。

第三步: 计算事故热点成因因素。

$$C_{road} = CR_{1 \times 5} F_{5 \times 1} = CR_{1 \times 5} (|V_{5 \times m}| W_{1 \times m}^T), \quad (10)$$

$$C_{society} = CS_{1 \times 5} F_{5 \times 1} = CS_{1 \times 5} (|V_{5 \times m}| W_{1 \times m}^T). \quad (11)$$

3 数据处理流程

为了检验前文提出方法的有效性, 基于 STAs19 数据库事故历史数据信息进行验证。首先需要对数据进行处理, 包括实验区域筛选、基于区域事故点聚类的热点识别, 以及成因因素计算。

3.1 实验区域筛选

文献[3]阐述了基于道路物理成因和社会成因两个因素预测中国道路交通事故发生的可行性。本文关注事故热点的道路物理成因和社会成因, 因此, 实验区域的筛选可采用基于道路因素和基于社会因素两种模式展开。

基于道路因素的筛选以路况条件作为参照, 路况条件不仅与道路本身有关, 也与道路地理特性相关, 较难获得准确的评价参数。

基于社会因素的筛选以区域社会因素(如区域人口密度、犯罪率及受教育程度)作为参照, 这类数据较易获得, 因此本文选择基于社会因素的筛选模式。基于 STAs19 数据库选择区域人口密度和犯罪率两个社会因素进行区域筛选。图2为伦敦市 32 个行政区域的人口密度和犯罪率分布。

对人口密度和犯罪率进行分析, 筛选 5 个区域进行实验测试(见表5)。在 5 个区域内, 各选择一片面积基本相等的测试对象区域, 区域经纬度如表6所示。

以上根据社会因素筛选获得的 5 个测试对象区域, 其道路特性也表现出多样性, 例

如 Sutton 目标区域为山区，而 Westminster 则为商业区，满足实验的基本条件。

3.2 基于DTH3N算法的热点识别

采用 DTH3N 算法对 5 个测试对象区域进行事故热点聚类识别，Islington 聚类结果如图 3 所示，事故点发生的位置用 (X, Y) 来表示，X 和 Y 分别为事故发生点的经度和维度抽象后得到的数值型坐标，图中不同颜色代表聚类形成的不同的簇，同一颜色代表同一个簇。

文献[5]中，对 5 个测试对象区域的事故点进行聚类获取事故热点，聚类后获得各测试对象区域的热点数量(见表 7)。分析数据可知，区域的人口密度与事故热点数量密切相关，由此证明了区域筛选方法的合理性。

3.3 成因因素计算

根据表 1 定义，从 STAts19 数据库中筛选相关属性参数，累加计算可得道路事故热点成因综合权重指标 C_{road} 和 $C_{society}$ 。

4 实验结果及分析

实验通过 MATLAB 对 5 个测试对象区域进行事故热点成因分析，验证本文提出的基于主成分贡献度的道路事故热点成因分析方法在事故热点评估中的性能。

各测试对象区域中事故热点成因综合权重指标 C_{road} 和 $C_{society}$ 如图 4 所示。可以看出，Islington, Hillingdon, Westminster 地区大部分事故热点 $C_{society}$ 比 C_{road} 大，因此驾驶人行至这些区域时需更多地注意社会成因对行车的影响；Wandsworth 中一小部分事故热点的 $C_{society}$ 和 C_{road} 差距比较明显，其他部分相差不大，因此在此区域中行车需要注意社会成因和道路物理成因的影响；而 Sutton 中所有事故热点的 C_{road} 大于 $C_{society}$ ，因此在此区域中行车更需注意道路物理成因的影响。

为了更深入地分析各事故热点成因与其所在区域道路特性的联系，计算分析各测试对象区域 $C_{society}$ 和 C_{road} 的均值及其方差均值(见图 5)。可以看出：区域人口密度相对较高的 Islington, Wandsworth 以及 Westminster, $C_{society}$ 均值大于 C_{road} 均值，但 Wandsworth 的 $C_{society}$ 和 C_{road} 均值差距并不明显。区域犯罪率相对较高的 Islington 和 Westminster 的 $C_{society}$ 方差均值远大于 C_{road} 方差均值；而区

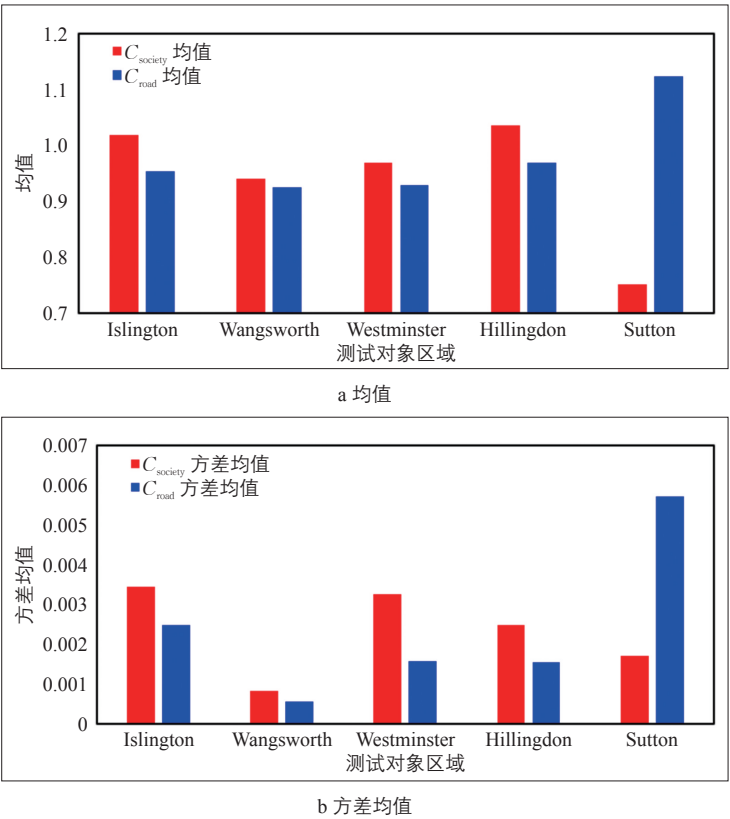


图 5 C_{road} 和 $C_{society}$ 均值和方差均值

Fig5 Mean and variance mean of C_{road} and $C_{society}$

域犯罪率较低的 Wandsworth 中 $C_{society}$ 方差均值略大于 C_{road} 方差均值。因此，可以宏观地认为 Islington 和 Westminster 中事故热点的形成主要受到 $C_{society}$ 影响，而 Wandsworth 中事故热点的形成受 $C_{society}$ 和 C_{road} 共同影响。此外，人口密度和犯罪率都相对较低的 Sutton, C_{road} 均值及其方差均值都远远大于 $C_{society}$ ，可以认为 Sutton 事故热点的形成主要受 C_{road} 影响；区域人口密度非常低而高犯罪的 Hillingdon 事故热点的形成主要受到 $C_{society}$ 的影响。由图 5b 可以看出，Hillingdon 事故热点成因综合权重指标的方差特性与 Islington 等高犯罪率区域类似；图 4 中 Hillingdon 的 $C_{society}$ 值大于 C_{road} 。因此可以推断 Hillingdon 的 $C_{society}$ 较高是受到较高犯罪率的影响。因此，高犯罪率是测试对象区域事故热点形成主要受社会成因影响的主要因素。

5 结语

在道路交通安全研究领域，事故热点成因探究是解决交通安全问题关键的一步。通过运用事故热点成因因素对事故热点进行评估，可以有效地指导驾驶行为决策和优化交

(下转第 75 页)