DOI:10.14048/j.issn.1671－2579.2022.06.049

基于 K-Means 聚类算法的高速公路事故多发路段鉴别

吴志敏[1]，黄觉[2]，向崎[1]

（1.广东华路交通科技有限公司，广东 广州 510420；2.广东省路桥建设发展有限公司，广东 广州 510420）

摘要：为鉴别高速公路事故多发路段，该文提出了一种基于 K-Means 聚类算法的事故多发路段鉴别方法。针对事故严重程度，引入路产损失与平均伤亡赔偿金作为当量事故数评定指标，对传统的当量事故数进行改进。根据改进当量事故数统计分布特征确定路段划分长度，结合累积频率法对事故多发路段进行初步鉴别。采用 K-Means 聚类算法对初选的事故多发路段进行聚类分析，得出最终的事故多发路段。为验证所提方法的正确性，对广(州)—梧(州)高速公路河口至平台段事故多发路段进行鉴别。结果表明：相比传统当量事故数，改进当量事故数更能反映事故严重程度；改进当量事故数服从负二项分布，可根据其统计分布特征得出客观的路段划分长度；采用 K-Means 聚类算法筛选结果优于 DBSCAN 算法，其筛选的事故多发路段总长度占初选结果的 66.7%，该方法可为高速公路事故多发路段治理提供强有力的理论依据。

关键词：事故多发路段；路段划分长度；聚类分析；负二项分布；高速公路

中图分类号：U416.2 **文献标志码**：A

根据 2014 年国家统计局的资料显示，中国每年发生的交通事故将近 20 万起，其中驾龄为 1～5 年的驾驶员占死亡人数的 45.3%[1]。高速公路事故多发路段是指受复杂道路环境的影响，在某个路段范围内发生的事故次数相较于其他路段频繁[2－3]。通常事故多发路段长度相对于路网总长较短，但发生的事故量占比较大。如何有效鉴别高速公路事故多发路段是高效治理的理论基础，对提升高速公路运营管理水平具有重要意义。

目前，国内外专家学者对事故多发路段鉴别相关研究取得了很多成果。孟祥海等[4]采用统计方法对事故多发路段进行了鉴别，并识别出其突出影响因素；瞿庆亮等[5]考虑事故多发路段空间位置与空间分布特征，提出聚类算法对事故多发路段进行鉴别；张长生等[6]对山区高速公路的事故形态与时间分布特征进行了统计，以当量事故率作为判定指标，采用质量控制法鉴别事故多发路段；Elvik[7]考虑多项事故致因因素，建立事故与致因因子之间的回归模型，以相对危险程度作为评价指标，可用于对事故多发路段进行鉴别；Lord 等[8]验证了事故数服从泊松分布，根据其统计特征确定了事故多发路段的临界点；Wright 等[9]以事故死亡人数作为判定标准，通过建立危险度模型对事故多发路段进行了鉴别。

以上研究从不同角度对事故多发路段鉴别进行了阐述，但对于事故多发路段的长度确定多采用固定单位划分，存在较强的主观性。因此该文采用负二项分布与累积频率法相结合对事故多发路段进行鉴别，同时引入 K-Means 聚类算法进一步确定事故多发路段的真实长度，所取得事故多发路段的长度具有一定的客观性，可为事故多发路段高效治理提供理论依据。

1 事故多发路段鉴别方法概述

事故多发路段鉴别方法通常建立在概率统计思想上，以历年事故数据为依据，对其进行统计分析。给定相应的标准值(事故数临界点、事故严重程度等)作为判定标准，如超过标准值，将其确定为事故多发路段，分析事故发生成因，制定相应的改善措施。常见的鉴别方法[10]如图 1 所示。

以上鉴别方法在实际运用中均存在不同程度的局限性，往往得出的结果对后期治理作用有限[11]。相对于其他方法，累积频率法可根据在不同情况下得出相

收稿日期：2020－10－29

基金项目：广东省交通运输厅科技项目(编号：科技－2016－02－040)

作者简介：吴志敏，男，硕士，工程师. E－mail:1060705467@qq.com

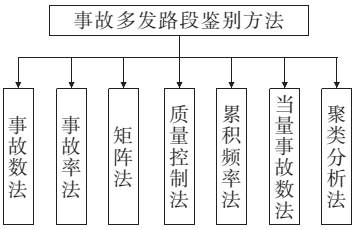

图 1　常见的事故多发路段鉴别方法

对应的判定标准值，适用范围较广，是目前事故多发路段鉴别运用较多的方法。与此同时，K-Means 聚类算法具有精度高、分析结果不易失真等特点，可对初选的事故多发路段进行再分析，确定最终的事故多发路段。因此该文将以累积频率法作为事故多发路段初步选定方法，结合事故多发路段统计分布特征，确定事故多发路段长度。对于事故多发路段初选结果，采用 K-Means 聚类算法对其进一步筛选，以确定较为准确的事故多发路段长度。

2　基于改进当量事故数—累计频率法的事故多发路段鉴别

传统的当量事故数法通过对死伤人数进行赋值来表征事故数，用于说明事故的严重程度，而未考虑事故造成的路产损失。随着行车安全技术的不断提升，车辆发生事故引起的伤亡人数有所降低。但事故发生时车辆可对路面、护栏以及其他交通工程沿线设施造成损害，往往事故越严重造成的路产损失越大，因此路产损失作为事故严重程度的评估指标是有必要的。

为了对路产损失进行事故当量化，该文引入同样可表征事故经济损失的伤亡赔偿金，将两者之间的比值用作事故当量加权值。采用路段总路产损失与该路段的平均伤亡赔偿金之间的比值对伤亡人数事故当量进行加权，改进了传统的当量事故数法，其计算公式如下：

$$E_A = A + (1+k)(\alpha B + \beta C) \tag{1}$$

$$k = \frac{\sum_{i=1}^{A} F_i}{\sum_{i=1}^{A} S_i/(B+C)} \tag{2}$$

式中：E_A 为总事故数；A 为路段相应的事故发生数；B 为路段事故发生时死亡人数；C 为路段事故发生时受伤人数；α，β 分别为事故死亡人数、受伤人数的权重，事故死亡人数的权重取值通常为 2.0，而受伤人数的权重取值通常为 1.5[12]；k 为路段总路产损失与该路段的平均伤亡赔偿金之间的比值；F_i 为路段发生第 i 次事故时路产损失值；S_i 为路段发生第 i 次事故时伤亡赔偿金。

相关研究认为高速公路事故数服从负二项分布[13]。若当量事故数服从负二项分布，可通过计算负二项分布参数，对事故路段进行划分，得出路段划分长度，其计算公式如下：

$$\delta = \lambda l / [V_{ar}(x)/\lambda l - 1] \tag{3}$$

式中：δ、λ 为负二项分布参数，可通过贝叶斯估计计算得出；l 为路段划分长度；$V_{ar}(x)$ 为事故的方差值。

在确定事故路段长度的基础上，该文采用累积频率法对事故多发路段进行鉴别。以累积频率作为纵坐标，每单位长度发生的事故次数作为横坐标，绘制累积频率图。相关研究表明[14]，累积频率通常在 5%～20%范围内将存在一个突变点，其计算方法为对累积频率曲线进行二次求导，找出零点，从而得出相应的突变点。在突变点上方，累积频率虽在增加，但事故次数不断递减；在突变点下方，事故次数急剧上升。因此，将突变点下方对应的路段作为事故多发路段。同时，按照路段线形及结构物对其划分为平直路段、平曲线路段、纵坡路段、弯坡组合路段、隧道路段和互通立体交叉路段 6 个路段，分别求取累积频率突变点，实现对各路段事故多发路段的初步确定。

3　基于 K-Means 聚类算法的事故多发路段选取

K-Means 聚类算法的中心思想[15]：事先确定聚类类别数为 K，随机选取初始点为聚类中心，并通过计算每一个样本与聚类中心之间的距离，将样本点归到最相似的类别中，重复上述过程，直到聚类中心不再改变，最终确定了每个样本的属性类别以及每个类的聚类中心，聚类分析过程如图 2 所示。

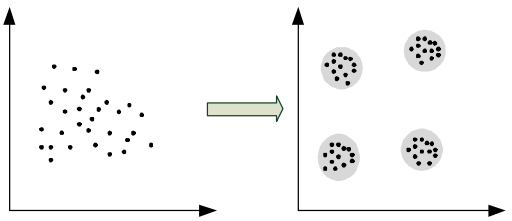

图 2　K-Means 聚类算法分析结果示意图

该文采用 K-Means 聚类算法对已选取的事故多发路段样本进一步筛选，得出最具代表性的事故多发路段，其计算步骤如下：① 将 6 个划分路段中初步确

定的事故多发路段样本分别选取 K 个对象作为聚类中心；② 计算各事故样本与所选取的聚类中心之间的聚类，将各事故样本按照最小距离原则归到最相似的聚类类别；③ 根据②中聚类结果，重新选取 K 个聚类中心；④ 重复步骤②与③，直到确定最终聚类中心。

根据聚类分析结果可得聚类中心阈值 N 与聚类中心半径 R，其中当初选的事故多发路段事故数大于阈值 N 时，可得到筛选的事故多发路段。为了确定相应的事故多发路段长度，采用已筛选出的事故多发路段中各事故样本与各聚类中心之间的误差平方和确定，当误差平方和最小时，该事故样本群落入的聚类中心半径即为其事故多发路段长度，计算公式如下：

$$Y=\sum_{i=1}^{N}\sum_{k=1}^{K}\varphi\left|x_i-m_k\right|^2 \tag{4}$$

式中：x_i 为第 i 个事故样本点；m_k 为聚类类别 k 的平均聚类中心；φ 为指数函数值，当样本 x_i 属于第 k 类时，取值为 1，否则取值为 0。

4 实例分析

该文以广(州)—梧(州)高速公路河口至平台段作为研究对象，对其事故多发路段进行鉴别。首先，对 6 个路段事故进行统计，分析了改进当量事故数，同时分别检验各路段改进当量事故数统计分布，确定路段划分长度；然后采用累积频率法对事故多发路段进行初步选取；最后运用 K-Means 聚类算法对事故多发路段进一步筛选，确定最终的事故多发路段，并与其他方法进行对比，证明该文所提方法的适用性。

4.1 工程概况

广梧高速公路河口至平台段全长 98.49 km，运营桩号为 K89＋168～K187＋658，路基宽度 24.5 m，设计速度 80 km/h，平曲线最小半径 740 m，设计最大纵坡 3.5%，竖曲线最小半径为凸曲线 7 000 m，凹曲线 10 000 m。沿线主要构造物有：隧道 18 座，互通立交 8 处。

4.2 事故多发路段鉴别

(1) 改进当量事故次数结果分析

收集了广梧高速公路河口至平台段广州方向 2014 年 1 月至 2016 年 12 月的事故资料，经统计该项目 3 年内发生的事故数为 504 起，死亡人数为 8 人，受伤人数为 121 人。为了得到符合该文要求的事故数据，对该项目中车辆碰撞程度较小、与动物相撞等未造成人员伤亡与路产损失事故进行了剔除，最终得到事故数为 348 起。依据传统当量事故数法以及改进的当量事故数法，得出该项目的传统当量事故次数以及改进当量事故次数分别为 546 起、649 起。

将该项目路段划分为平直路段、平曲线路段、纵坡路段、弯坡组合路段、隧道路段以及互通立交路段，对上述路段按照传统当量事故数与该文所提的改进当量事故数进行统计，结果如表 1 所示。

表 1 各路段的传统当量事故次数与改进当量事故次数统计结果

路段名称	传统当量事故次数/起	改进当量事故次数/起
平直路段	22	26
平曲线路段	51	68
纵坡路段	86	108
弯坡组合路段	81	97
隧道路段	128	149
互通立交路段	178	201
总量	546	649

从表 1 可看出：改进当量事故次数总量远远高于传统当量事故次数，说明引入路产损失与平均伤亡赔偿金作为当量事故次数的评估指标，可进一步反映事故严重程度。同时，相对于传统当量事故次数，隧道路段、互通立交路段以及纵坡路段的改进当量事故次数变化较大，主要原因在于以上路段发生的侧翻、冲向护栏等事故造成的路产损失远远高于其他路段，且受到车流量、道路环境等因素影响，路产修复工作较为困难，因此对事故治理提出了更高的要求[16]。

(2) 改进当量事故次数的统计分布特征

为了确定划分的 6 个路段改进当量事故次数的统计分布特征，采用拟合优度检验[17]，检验结果如表 2 所示，其中 χ^2 为卡方检验值，v 为自由度，p 为显著性水平(当 $p>0.05$ 时，服从负二项分布，否则服从其他分布)。

表 2 当量事故次数统计分布的拟合优度检验

路段名称	事故总数/起	χ^2	v	p
平直路段	26	9.72	6	0.68
平曲线路段	68	7.56	8	0.79
纵坡路段	108	14.21	12	0.45
弯坡组合路段	97	13.54	13	0.51
隧道路段	149	16.14	16	0.33
互通立交路段	201	18.42	19	0.26

从表 2 可看出：各路段当量事故次数均服从负二项分布。在拟合优度检验中，卡方检验值可反映检验对象集散程度，卡方检验值越大，说明检验对象聚集程度越高。隧道路段与互通立交路段占全网路段长度的 40.17%，卡方检验值相较其他路段要高，说明事故相对集中，而纵坡路段与弯坡组合路段长度虽占比不大(11.23%)，但该项目处在山岭地区，坡度变化频繁，引发的事故较多；平直路段与平曲线路段卡方检验值较小，说明事故相对分散。

(3) 事故多发路段鉴别的初步选取

采用式(3)计算各路段的事故划分长度，依据划分路段采用累积频率法鉴别事故多发路段，得出初选结果，各路段事故－累积频率曲线如图 3 所示。

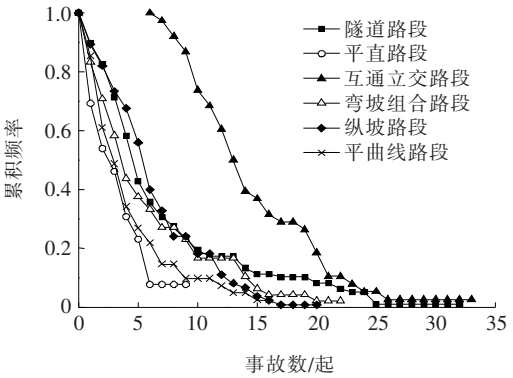

图 3　各路段事故－累积频率曲线示意图

经计算可得出隧道路段、平直路段、互通立交路段、弯坡组合路段、纵坡路段以及平曲线路段的事故－累积频率突变点的事故数分别为 20 起、5 起、21 起、14 起、11 起、9 起。对当量事故数高于突变点的路段归为事故多发路段，具体事故多发路段初选结果以及路段划分长度如表 3 所示。

4.3　事故多发路段的确定

采用 K-Means 聚类算法对初选的事故多发路段进一步筛选，得出最终的事故多发路段。对各路段初选的事故多发路段进行聚类分析，以事故多发路段个数作为聚类中心个数，对大于所求取的聚类阈值作为最终事故多发路段。通过计算各事故样本与聚类中心的事故当量次数误差最小值对应的聚类中心半径 R 作为最终事故多发路段的长度。为了体现该文所提算法的优势，该文采用文献[18]中 DBSCAN 算法与 K-Means 聚类算法所选结果进行对比分析，最终选取的事故多发路段结果如表 4 所示。由于篇幅受限且筛选路段较多(31 个)，该文仅给出筛选路段的最前桩号与最终桩号。

表 3　事故多发路段鉴别结果一览表

路段名称	起点桩号	终点桩号	路段划分长度/m	改进当量事故次数/起
隧道路段	K93＋800	K94＋700	900	20.6
	K111＋129	K111＋939	810	23.4
	K118＋876	K119＋546	670	21.8
	K134＋054	K135＋034	980	24.9
	K142＋158	K142＋948	790	26.2
互通立交路段	K101＋920	K102＋960	1 040	22.8
	K115＋580	K116＋180	600	24.5
	K155＋480	K156＋230	750	28.4
	K168＋150	K168＋870	720	30.5
	K176＋270	K176＋940	670	31.4
	K186＋210	K186＋970	760	36.4
平直路段	K158＋310	K159＋420	1 110	9.4
纵坡路段	K139＋030	K139＋660	630	12.7
	K149＋880	K150＋680	800	21.1
弯坡组合路段	K139＋860	K140＋780	920	15.4
	K135＋170	K135＋680	510	18.7
	K186＋270	K186＋930	660	23.8
平曲线路段	K89＋200	K89＋800	600	10.1
总和			13 920	402.1

从表 4 可看出：采用 K-Means 聚类算法得到的事故多发路段的总长度低于 DBSCAN 算法，其总长度仅占初选事故多发路段总长度的 66.7%；而当量事故总数几乎相等，与初选事故多发路段的总数均相差不到 10%，说明最终确定的事故多发路段的当量事故数变化不大。

由于 DBSCAN 算法参数中邻域半径作为事故多发路段最小长度，其取值需根据不同的路线线形进行选取，选取标准存在主观性，容易偏离选取范围，将其他非事故多发路段纳入范围内，降低了事故多发路段鉴别精度。而该文所提方法中聚类半径不受路线线形影响，仅与所统计的当量事故数高度相关，其鉴别精度较高。因此相比 DBSCAN 算法，该文所选算法对初选的事故多发路段所隐藏的非事故多发路段筛选精度较高，说明了该文所提算法的适用性与科学性，可为事故多发路段高效治理提供理论依据。

5　结论

(1) 对传统的当量事故法进行了改进，将路产损

表 4 两种算法确定的事故多发路段结果

路段名称	K-Means 聚类算法				DBSCAN 算法			
	最前桩号	最终桩号	总长度/m	改进当量事故总数/起	最前桩号	最终桩号	总长度/m	当改进量事故总数/起
隧道路段	K93+945	K142+845	2 970	121.4	K93+933	K142+759	3 589	123.5
互通立交路段	K101+955	K186+640	3 180	141.0	K101+975	K186+858	3 945	145.7
平直路段	K158+475	K159+135	660	9.1	K158+569	K159+126	715	8.7
纵坡路段	K139+350	K150+405	765	38.9	K139+451	K150+298	849	39.5
弯坡组合路段	K135+255	K186+655	1 255	41.8	K135+199	K186+721	1 456	38.8
平曲线路段	K89+245	K89+695	450	9.4	K89+215	K89+715	500	9.5
总和			9 280	361.6			11 054	365.7

失转化为当量事故指标，更能客观地反映事故严重程度。

(2) 将全路段依据线形及结构物进行划分，采用改进当量事故法对各路段事故数进行统计，经检验当量事故数服从负二项分布；利用负二项分布特征，确定事故路段划分长度，结合累积频率法得到高准确度的事故多发路段。

(3) 通过 K-Means 聚类算法对初选事故多发路段进行筛选，得到更真实的事故多发路段；同时与前人研究对比发现，该文所提模型识别精度更高，为事故多路段治理提供了理论依据。

(4) 考虑 K-Means 聚类算法选取初始聚类中心的随机性可能导致聚类结果较弱的稳定性，下一步将对 K-Means 聚类算法进行改进，降低对初始聚类中心的依赖性，以期得到更为客观的事故多发路段。

参考文献：

[1] 孙丽璐，陈甜，赵娟，等. 我国交通事故损失影响因素及地区特征分析：基于全国 31 个省市自治区 2004—2015 年面板数据[J]. 西南大学学报(自然科学版)，2019，41(8)：114－123.

[2] 王颖志，王立君. 基于网络时空核密度的交通事故多发点鉴别方法[J]. 地理科学，2019，39(8)：1 238－1 245.

[3] 金霞，雷桂荣，刘峰，等. 交通事故黑点鉴别与改善排序研究[J]. 公路与汽运，2018(2)：45－48.

[4] 孟祥海，覃薇，霍晓艳. 基于统计与假设检验的高速公路交通事故数据分布特性[J]. 交通运输工程学报，2018，18(1)：139－149.

[5] 瞿庆亮，曲国庆，徐工. 车载 GPS 在公路交通事故多发路段判别中的应用[J]. 公路，2016，61(7)：224－228.

[6] 张长生，马荣国. 山区高速公路交通事故分析及多发路段鉴别[J]. 长安大学学报(自然科学版)，2010，30(6)：76－80.

[7] ELVIK R . Evaluations of Road Accident Blackspot Treatment：A Case of the Iron Law of Evaluation Studies? [J]. Accident Analysis & Prevention，1997，29(2)：191－199.

[8] LORD D，WASHINGTON S，IVAN J N，et al. Poisson，Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes：Balancing Statistical Fit and Theory[J]. Accident Analysis & Prevention，2005，37(1)：35－46.

[9] WRIGHT C C，ABBESS C R，JARRETT D F. Estimating the Regression-to-Mean Effect Associated with Road Accident Black Spot Treatment：Towards a more Realistic Approach[J]. Accident Analysis & Prevention，1988，20(3)：199－214.

[10] 彦伟，宇仁德，孙连超. 鉴别道路交通事故黑点的聚类分析方法研究[J]. 交通标准化，2011(11)：110－113.

[11] 邵祖峰. 交通事故黑点鉴别方法研究综述[J]. 道路交通与安全，2008(2)：44－49.

[12] 王钱. 基于改进累计频率曲线法的高速公路事故多发点鉴别研究[D]. 西安：长安大学，2018.

[13] 孟祥海，刘振博. 基于 Tobit 回归的山区高速公路事故率分析模型[J]. 中外公路，2020，40(2)：294－299.

[14] 王龙健，成嘉琪. 累计频率法鉴别道路事故多发点中单位取样长度的限定[J]. 公路与汽运，2015(5)：36－39.

[15] 高曼，韩勇，陈戈，等. 基于 K-Means 聚类算法的公交行程速度计算模型[J]. 计算机科学，2016，43(S1)：422－424，439.

[16] 邓国忠，曹帆，吴勇，等. 互通式立交与隧道出口小间距路段事故影响因素分析[J]. 中外公路，2019，39(4)：283－287.

[17] CHANG L. Analysis of Freeway Accident Frequencies：Negative Binomial Regression Versus Artificial Neural Network[J]. Safety Science，2005，43(8)：541－557.

[18] 耿超，彭余华. 基于动态分段和 DBSCAN 算法的交通事故黑点路段鉴别方法[J]. 长安大学学报(自然科学版)，2018，38(5)：131－138.