2004年某县交通事故数据的挖掘分析

李武选[1],郭岩红[2],李 源[3],李 军[4]

(1. 长安大学 经济与管理学院,陕西 西安 710064;
2. 北京商业管理干部学院 党委办公室,北京 100028;
3. 中兴通讯技术股份有限公司 软件系统测试部,陕西 西安 710065;
4. 陕西彬长矿区物资供销有限责任公司,陕西 咸阳 712000)

摘 要:为了对交通事故实施量化分析,通过图表、聚类、相关、偏相关及回归分析等统计方法对2004年某县交通事故的基本数据进行了分析。分析认为,导致该县交通事故直接经济损失的直接因素主要是受伤人数,而不是死亡人数和事故起数;同时从不同角度进行了事故原因分析,分析结果表明:加强周末交通监管力度,增加疏导交通的人数和时间等;加强好天气的交通管理;注重于驾驶人员和出行人员保持一个良好的心态及车辆自检问题;加大宣传和整治马路市场、事故黑点的力度等。

关键词:交通管理;交通事故;驾驶人员;交通方式
中图分类号:U491.1 **文献标志码**:A **文章编号**:1671-6248(2009)01-0050-05

交通事故已成为当今危害人类健康、家庭幸福和人的生命的罪魁祸首,因此尽快解决好交通安全问题便成为现代社会的当务之急。通过挖掘和分析交通事故数据,从而发现交通规律,避免交通事故就成为一个紧迫而且重要的问题。对于此类问题的研究主要有:秦利燕、邵春福从GIS角度进行安全管理系统建设方面研究[1];刘志强、王兆华和钱卫东就车速问题进行安全管理研究[2];路平从道路和交通条件方面对比分析了道路交通事故的基本规律,并提出了相应对策[3],然而从统计和计量角度对交通事故数据进行挖掘分析还很少见。鉴于此,我们从各个不同的角度就南方某县一年内道路交通事故发生的相关数据做了深入而全面的分析。

一、交通事故的定量分析

由2004年南方某县境内交通事故的统计资料[4]可知,国道线发生事故5起,占交通事故总数的15%,造成5人死亡,4人受伤,直接经济损失2.33万元;省道线发生事故6起,占总数的18%,造成2人死亡,7人受伤,直接经济损失2.6万元;县乡道路发生事故15起,占总数的45%,死亡14人,伤11人,直接经济损失3.6万元;乡镇内道路发生事故7起,占总数的21%,造成6人死亡,8人受伤,直接经济损失4.893万元(表1)。

表1 交通事故情况统计

道路发生交通事故的类型	国道线	省道线	县乡道线	乡镇内道路	全年总计
发生事故数/起	5	6	15	7	33
所占比例/%	15	18	45	21	100
死亡人数/个	5	2	14	6	27
受伤人数/个	4	7	11	8	30
直接经济损失/万元	2.33	2.6	3.6	4.893	13.423

收稿日期:2008-06-12
作者简介:李武选(1962-),男,陕西兴平人,讲师。

(一)图示分析

笔者通过 Excel 软件对 2004 年该县境内发生的交通事故情况进行分组直方图分析(图 1)。

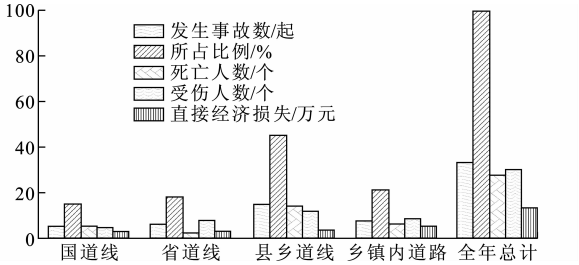

图 1 交通事故五项指标按不同规模道路类型对比直方图

(二)聚类分析

SPSS 提供 2 种聚类方法:一种是层次聚类法,另一种是快速聚类法。层次聚类法又包括 R 型聚类法和 Q 型聚类法,其中前者为样本聚类法,后者为变量聚类法[5]。本文使用的是 Q 型聚类法,其基本思想是在聚类分析的开始,每一个变量自成 1 类;然后按照某种方法度量所有变量之间的亲疏程度,并把最亲密或者称最相似的变量首先聚成一小类;然后再度量剩余的变量和小类(或者小类和小类)之间的亲疏程度,并将当前最亲密的变量或者小类再聚成 1 类;如此进行反复聚类,直到所有变量聚成 1 类为止。

不同道路类型的 Q 型聚类分析从数据分析的结果看,如分成 2 类时,县乡道线的情况更严重,它的事故发生情况相对其他类型规模道路"独树一帜",是重点整治的对象道路。若分成 3 类,则县乡道线和乡镇内道路都将成为重点整治对象(图 2)。

聚类类数	县乡道线		乡镇内道路		省道线		国道线
1类	×	×	×	×	×	×	×
2类	×		×	×	×	×	×
3类	×		×		×	×	×

图 2 发生事故起数按不同规模道路系统聚类结果

(三)相关回归分析

从图 3 可以看出,直接经济损失与发生事故数、死亡人数和受伤人数都有关系。鉴于发生事故所占比例是一个派生数据,所以在下面分析中不予考虑。同时笔者发现该县在乡镇内道路上发生事故的受伤人数高出发生事故起数,说明有严重交通事故发生;在省道线同样存在这一情况。

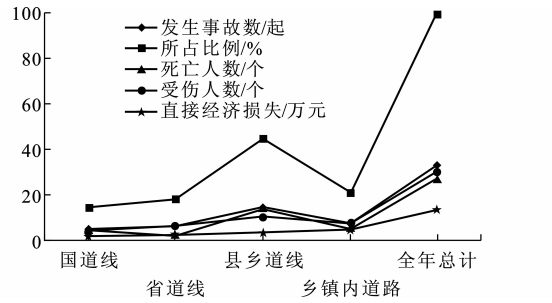

图 3 直接经济损失与发生事故数、死亡人数和受伤人数的相关折线

1. 相关性分析

(1)简单相关性分析。从简单相关系数(表 2)可以看出,在不考虑相互影响的情况下,该县交通事故直接经济损失与全年交通事故总数、死亡人数和受伤人数相关性均呈正向高度相关,这说明上述 3 个因素都是该县直接经济损失的重要因素,其重要次序依次为受伤人数、死亡人数和全年交通事故总数,而且这种正相关关系均是显著的(显著性检验数为括号中数值,均小于0.05)[6]。

表 2 交通事故直接经济损失与影响因素的简单相关系数

简单相关系数	全年发生事故总	死亡人数	受伤人数
直接经济损失	0.941(0.017)	0.909(0.033)	0.976(0.004)
全年发生事故总数		0.985(0.002)	0.987(0.002)
死亡人数			0.952(0.013)

(2)偏相关性分析[7-8]。本文针对偏相关性分析从控制 1 个变量和控制 2 个变量分别阐述。

在控制 1 个变量的情况下,即只考虑死亡人数,那么控制受伤人数时交通事故的直接经济损失与死亡人数之间的偏相关系数见表 3。这个偏相关系数表明:如果剔除受伤人数对直接经济损失和死亡人数的影响后,直接经济损失与死亡人数的净相关性为 -0.308,也就是说该县交通事故直接经济损失随着死亡人数的增加而减少;反之亦然。但是这个结论显著性太差[6](显著性检验数为括号中的 0.692,它明显大于 0.05)。这说明该县死亡人数的赔偿费用非常之低。在控制 1 个变量的情况下,即只考虑受伤人数,那么控制死亡人数时交通事故的直接经

济损失与受伤人数之间的偏相关系数见表 4。这个偏相关系数表明:如果我们剔除死亡人数对直接经济损失和受伤人数的影响后,直接经济损失与受伤人数的净相关性为 0.870,在 0.10 的显著性水平下是统计显著的。这说明该县受伤人数的赔偿费用与交通事故的直接经济损失存在高度的正相关关系。

表 3　控制受伤人数时直接经济损失与死亡人数的偏相关系数

控制变量	直接经济损失	死亡人数
受伤人数 & 直接经济损失	1.000	-0.308(0.692)
死亡人数		1.000

表 4　控制受伤人数时直接经济损失与受伤人数的偏相关系数

控制变量	直接经济损失	受伤人数
死亡人数 & 直接经济损失	1.000	0.870(0.130)
受伤人数		1.000

在控制 2 个变量的情况下,即控制全年发生事故总数和受伤人数时直接经济损失与死亡人数的相关系数见表 5。如果我们剔除全年发生事故起数和受伤人数对直接经济损失和死亡人数的影响,那么这两者的净相关系数提高到 0.944,在 0.215 的显著性水平下,它们是统计显著的。这说明该县直接经济损失与死亡人数的正相关关系是重要的。死亡人数应当是直接经济损失的重要一部分,可在交通事故处理中,受伤的费用具有持久性,甚至持续到受伤者生命的结束,其治疗费用应当是很大的。在控制 2 个变量的情况下,即控制全年发生事故总数和死亡人数时直接经济损失与受伤人数的偏相关系数见表 6。由表 6 可知,如果剔除全年发生事故起数和死亡人数对直接经济损失和受伤人数的影响,那么这两者的净相关系数提高到0.985,在0.111的显著性水平下,它们是统计显著的。这说明该县直接经济损失与受伤人数的正相关关系更为突出。

表 5　控制全年发生事故总数和受伤人数时直接经济损失与死亡人数的偏相关系数

控制变量	直接经济损失	死亡人数
全年发生事故总数 直接经济损失 & 受伤人数	1.000	0.944(0.214)
死亡人数		1.000

表 6　控制全年发生事故总数和死亡人数时直接经济损失与受伤人数的偏相关系数

控制变量	直接经济损失	受伤人数
全年发生事故总数 直接经济损失 & 受伤人数	1.000	0.985(0.110)
受伤人数		1.000

2. 回归分析

通过统计分析软件 SPSS13.0[9] 进行建模,得到该县 2004 年直接经济损失与发生交通事故时的受伤人数、发生事故次数和死亡人数的线性回归模型计量经济表达式[10-12]。

(1)单因素影响的回归模型。模型Ⅰ调整的拟合优度系数 $R^2=0.938$,F 统计量为 61.252,Sig. $F=0.004<0.05$。模型Ⅰ整体上显著性突出(式(1))。

(2)双因素影响的回归模型。模型Ⅱ调整的拟合优度系数 $R^2=0.944$,F 统计量为 34.914,Sig. $F=0.028<0.05$。模型Ⅱ整体上显著性突出(式(2))。

(3)三因素影响的回归模型。模型Ⅲ调整的拟合优度系数 $R^2=0.988$,F 统计量为 105.905,Sig. $F=0.070>0.05$。模型Ⅲ整体上不具有显著性(式(3))。

$$y=0.434x_1+0.156 \quad (1)$$

$$y=0.803x_2-0.330x_2+0.082 \quad (2)$$

$$y=1.258x_1-1.244x_2+0.613x_3 \quad (3)$$

式中:y 为直接经济损失;x_1 为受伤人数;x_2 为发生事故次数;x_3 为死亡人数。

从统计意义上看,要根据相关性检验、F 检验、方差分析表可知,模型Ⅰ是最佳的回归模型;也就是说,该县经济损失的主要原因是由交通事故中的受伤人数引起的,而不是死亡人数和交通事故次数的多寡。但是考虑到我们的目的主要是对该县交通事故直接经济损失的影响因素分析,也可以选择模型Ⅲ因素。鉴于该模型解释变量的检验均不显著,不为 0(显著性检验数均大于 0.05),解释因变量的原因不充分,但它是必要的。

二、交通事故的原因分析

(一)发生事故的日期

笔者对 2004 年该县发生的交通事故按发生事故的日期进行统计并列表(表 7)。由表 7 可知,每月的 17 日事故发生最多,占交通事故总数的 15%,

每月 1 日发生事故占总数的 12%,每月的 5 日、7 日、21 日、24 日发生事故分别占总数的 6%。据我们查询日历,2004 年 12 个月的每月 17 日是星期一有 1 天;是周二有 2 天;是周三有 2 天;是周四有 1 天;是周五有 2 天;是周六有 3 天;是周日有 1 天。显然,当年每月发生事故集中于周二、周三和周五、周六 4 天。实际调查统计数据表明该县全年之中周五事故发生最多,全年共 11 起,占交通事故总数的 33%,周一最少,全年 2 起,占总数的 6%。

表 7 交通事故按日期分布

发生事故的日期分布	1 日	5 日	7 日	17 日	21 日	24 日
所占发生事故总数比例/%	12	6	6	15	6	6

由此可见,个人休息状态是影响交通事故发生的一个主要影响因素(包括司乘人员和相应交通事故涉及的非司乘人员,周一的前一天大多数人休息状况良好,相应人员头脑清醒、注意力集中、反应灵敏);造成交通事故多发的另外一个原因是道路拥挤、饮酒、娱乐等(周末休息时间来临,下班回家或外出皆需要赶时间,情绪影响)。因此,笔者建议:要控制休息状况;控制情绪;分不同情况分段下班,若该措施不可行时,可采用周末加强交通监管力度,增加疏导交通人数和时间等方法。

(二)发生事故的天气状况

笔者对 2004 年该县发生的交通事故按天气状况进行统计并列表(表 8)。由表 8 可知,晴天事故发生最多,共发生 20 起,占交通事故总数的 61%,雨天发生事故 5 起,占总数的 15%,阴天、雪天和其他天气发生事故各 4 起,均占总数的 12%。从该县交通事故发生情况看,似乎天气状况的影响并不算大,阴、雨、雪等天气状况发生交通事故的比例加起来少于晴天发生交通事故的比例。这就要求该县交通管理部门考虑加强好天气的交通管理,这只需加强人力和相关措施即可。

表 8 交通事故按天气状况分布

发生事故天的气状况分布	晴天	雨天	阴天和雪天	其他天气	合计
发生事故数/起	20	5	4	4	33
所占发生事故总数比例/%	61	15	12	12	100

(三)发生事故的天气状况

笔者对 2004 年该县交通事故涉及的驾驶员年龄进行统计列表(表 9)。由表 9 可知,发生事故最多是年龄段处于 21 ~25 岁的驾驶员,共 12 起,占交通事故总数的 36%;45 岁以上驾驶员发生事故 8 起,占总数的 24%;26 ~30 岁发生事故 4 起,占总数的 12%;31 ~35 岁发生事故 5 起,占总数的 15%;36 ~40岁发生事故 1 起,占总数的 3%;41 ~45 岁发生事故 3 起,占总数的 9%。从表 12 数据发生交通事故的比例看,21 ~25 岁年龄段的驾驶员属于情绪易冲动型人群,他们是该县交通事故的高发人群;其次是 45 岁以上的人群;发生事故最少的年龄段当属 36 ~40 岁年龄段的人群。21 ~25 岁年龄段的年轻人,因为驾车上路时间短,交通知识不够,驾驶技术不熟练、经验缺乏,安全意识淡漠,当然还包括路况设施等因素影响。45 岁以上的驾驶员人群多数由于生活压力过大,急于赚钱养家糊口,似乎有一些拼命的因素。笔者建议:加强对 21 ~25 岁的年轻人严格管理和约束,通过惩罚与培训提高其社会责任心;45 岁以上人员则要注重于驾驶车辆心态的重新树立和加强,杜绝疲劳、疏懒、酗酒,在驾驶中始终保持一个良好的心态,还要注意到车辆自身状况。

表 9 2004 年某县发生交通事故按年龄段分布

发生交通事故的年龄段分布	21 ~ 25	26 ~ 30	31 ~ 35	36 ~ 40	41 ~ 45	45 岁以上	全年总计
发生事故数/起	12	4	5	1	3	8	33
所占比例/%	36	12	15	3	9	24	100

(四)发生事故的交通方式

笔者按照不同交通方式对 2004 年该县交通事故进行统计列表(表 10)。由表 10 可知,大客车发生事故 1 起,占交通事故总数的 3%;大型货车发生事故 4 起,占总数的 12%;小型客车发生事故 12 起,占总数的 36%;小型货车发生事故 5 起,占总数的 15%;机动农用三轮、拖拉机、摩托车发生事故 11 起,占总数的 34%。上述数据表明,小型客车和机动农用三轮、拖拉机、摩托车车辆的交通事故发生率最高,二者近乎相等,二者之和达到 70%。小型客运由于当年数量增多,短驾龄司乘人员增加,综合表 12 可看出 21 ~25 岁年龄段的驾驶员为事故高发年龄段,该部分驾驶员缺少驾驶经验,导致各类事故的发生。加之当前一些驾校培训水平不高,对驾驶员的培训只是注重驾驶技术的培训,没有进行必要的安全教育,造成很多新驾驶员的安全意识淡薄,为道路交通事故的发生带来了隐患。机动三轮、农用车辆、摩托车无证驾驶、无牌上路、违章载人的现象突出,很多农民驾驶员驾驶技术生疏,不了解交通法规和安全常识就驾车上路,行车中对交通情况观察不够,预见性差,遇有情况时又不能采取有效措施,是造成农用车辆发生事故,以致于发生群死群伤事故的主要原因。因此,笔者建议:通过各种形式(主要道路事故多发段的可视性标志放置要醒目等)加大宣传力度;提高驾驶员、行人和交管部门人员的协调

配合力度;严格落实对交管部门人员的责任,强化路面交通管理;对“三超”和“三无”车辆加强治理、增加力度;抓住事故发生的突出特点,采取针对性的措施;加大对马路市场和事故黑点的整治力度;不断加强对驾驶人员的安全教育和培训工作。

表 10　交通事故的交通方式分布

发生交通事故的交通方式	大客车	大型货车	小型客车	小型货车	农用三轮、拖拉机、摩托车	全年总计
发生事故数/起	1	4	12	5	11	33
所占比例/%	3	12	36	15	34	100

三、结　语

笔者从影响交通事故发生的各个方面就 2004 年该县交通事故的基本数据做了一些比较深入的挖掘分析,希望这些工作能够对交管部门有所帮助。

参考文献:

[1] 秦利燕,邵春福. 基于 GIS 的道路交通安全管理系统的研究[J]. 中国安全科学学报,2004,14(2):34-36.

[2] 刘志强,王兆华,钱卫东. 基于速度的交通事故分析[J]. 中国安全科学学报,2005,15(11):35-38.

[3] 路　平,龚瑞庚. 道路交通事故的基本规律与对策[J]. 长沙交通学院学报,1999,15(4):77-81.

[4] 佚　名. 2004 年道路交通事故情况分析及对策[EB/OL]. (2006-03-08)[2008-05-10]. http://www.40a.com/bggw/sort03021/211241.html.

[5] 薛　薇. SPSS 统计分析方法及应用[M]. 北京:电子工业出版社,2004.

[6] 李武选. 分段拟合技术在长期趋势建模过程中的具体应用[J]. 世界科技研究与发展,2008,30(3):363-369.

[7] 古扎拉蒂. 计量经济学[M]. 3 版. 林少宫,译. 北京:中国人民大学出版社,1999.

[8] 赵卫亚. 计量经济学教程[M]. 上海:上海财经大学出版社,2003.

[9] 周仁郁. SPSS13.0 统计软件[M]. 成都:西南交通大学出版社,2005.

[10] 张文彤. SPSS11.0 统计分析教程[M]. 北京:北京希望电子出版社,2002.

[11] 沈　浩,李亦兰,王迎迎. Excel 高级应用与数据分析[M]. 北京:电子工业出版社,2008.

[12] 胡可云,田凤占,黄厚宽. 数据挖掘理论与应用[M]. 北京:清华大学出版社,2008.

Mining analysis for traffic accident data in a county in 2004

LI Wu-xuan[1], GUO Yan-hong[2], LI Yuan[3], LI Jun[4]

(1. School of Economics and Management, Chang'an University, Xi'an 710064, Shaanxi, China; 2. Office of Party Committee, Beijing Institute of Business Management, Beijing 100028, China; 3. Department of Software System Testing, ZTE Corporation, Xi'an 710065, Shaanxi, China; 4. Shaanxi Binchang Limited Mine Supply and Marketing Materials Ltd, Xianyang 712000, Shaanxi, China)

Abstract: In order to conduct quantitative analysis for traffic accidents, the SPSS software is used for the examination of the basic situations of traffic accidents of a county in 2004. Through charts, clustering, correlation, partial correlation and regression, it is shown that the main factors that lead to the direct econmic losses are not the number of death and accidents, but the number of the injured. The authors in the paper analyze all the possible causes of the accidents from different angles and have offered the following suggestions to solve the problem. They are: the supervision for traffic should be strengthened at weekends; more staff should be sent for channeling traffic and they should work for longer time; there should be more work for traffic supervision during fine days; more educational work should be offered to both the drivers and passers-by; and publicity work, the clearing up of free markets near roads and improvement of black spots should be strengthened.

Key words: traffic management; traffic accident; driving staff; traffic mode