DOI:10.13986/j.cnki.jote.2015.05.006

基于数据可视化的短时小数值交通事故的描述及成因推理

陈永胜
(公路交通安全技术行业重点实验室,交通运输部公路科学研究院, 北京 100044)
(加拿大艾尔伯塔省埃德蒙顿市交通安全办公室,9304 41 Avenue Edmonton T6J 4L8)

摘 要:交通事故是小概率随机事件. 在特定时间空间内某些类型的交通事故指标通常是相对较小的数字. 应用传统的针对连续变量的方法(例如广义线性模型)对其进行分析和预测,通常由于数值小,随机性及起伏波动大而无法获得统计显著的结果;而采用传统的针对离散变量的方法(例如罗基模型)进行分析和预测,则又由于其分类数值太多而难以实现. 本论文探讨以数据可视化方法来解决这种短时、小数值交通事故数据的描述及推理分析的问题. 基于加拿大某城市的交通事故与天气数据,本论文探索使用一系列的"数据可视化"方法,例如数据分解、彩色散点矩阵图、三维散点图等,描述事故相关要素及其互动规律. 在此基础上,进一步应用图形模型作为成因推理的手段,以完成推理性的数据可视化分析,藉此分析造成交通事故的成因要素、各要素的关系结构以及要素对事故的数量化影响程度. 这一研究解决了在短时、小数值背景下对于交通安全状况进行精确描述及成因分析的问题,其成果可直接应用于交通安全管理、交通执法、道路养护等多个领域中的实时安全管控、安全治理措施的预案制订与效果评估等实际工作之中.

关键词:数据可视化;短时事故;成因推理
中图分类号:U 268.6 **文献标志码**:A **文章编号**:1008-2522(2015)05-27-09

Description and Causal Inference of Short-term Small-number Collisionsby Data Visualizations

CHEN Yong-sheng
(Key Laboratory of Road Safety Ministry of Transport, Research Institute of Highway Ministry of Transport, Beijing 100088, China)
(Office of Traffic Safety, Transportation Services, City of Edmonton, Alberta, Canada)

Abstract: Collisions are rare random events. Particular collision dataitems within specific temporal or spatial units, e.g., daily fatal and injury collisions of a small or medium sized city, are generally small numbers (say, 0 – 25). These small-numbered collisions are inadaptable to be analyzed and predicted by conventional approaches. For methods with continuous variables, such as generalized linear model (GLM), this type of data has limited value range, too high randomness and variation, so that statistically significant (SS) results are unlikely to be obtained. On another hand, for methods with discrete variables, e.g., the Logit Model, this type of datahas too many classificationsand therefore it is hard to be properly fitted. This paperworks on a solution to unravel this dilemma through newly developed data visualization approaches. Based on the sample data from a Canadian city, a series of data visualization methods, including data decomposition, colored scatter-plot matrix, 3D plots, were employed to describe collision patterns, and to identify its impact factors and figure out the interactions among the factors.

收稿日期:2014-12-30.
作者简介:陈永胜(1970—),男,首席研究员,加拿大安省注册职业工程师,研究方向为交通安全. E-mail:ys.chen@rioh.cn.

Then, the graphic model, as a particular causal inference method, was introduced in order to establish intrinsic connections from collisions to causal factors and to draw causal structure among factors. Moreover, the causal effects between each particular factor and the collision were quantitatively estimated. This study, combined with descriptive and inferential methods, fills the methodological vacancy for the short-time small-numbered collision data and outcomes of this study can be directly utilized to support real-time safety management and control, pre-scheduling and effect evaluation for safety countermeasures across multiple disciplines such as traffic management, enforcement, and road maintenance.

Key words: data visualization; short-term collision; causal inference

0 引言

交通事故是小概率随机事件. 在传统的交通事故分析中,必须在较长时间、较大空间两个维度上对事故数据进行累计,以满足最基本的统计分析有效性的要求. 与之相对的,在较短时间、较小空间内特定类型的交通事故数据一般是较小的数值. 这类事故数据包括了一些短时事故指标,例如一个城市的每日伤亡交通事故数;也包括了一些局部空间内的事故指标,例如一个小区的每年事故数据、某一特定路段的"驶出路外"事故等. 上述的交通事故指标一般较小,例如一个中小城市的每日伤亡交通事故数据,根据观测均介于 0 ~ 25 之间. 应用传统的针对连续变量的方法(例如广义线性模型)对其进行分析和预测,通常由于数值小,随机性及起伏波动大而无法获得统计显著的结果;而采用传统的针对离散变量的方法(例如罗基模型)进行分析和预测,则又由于其分类数值太多而难以实现. 因此,不论是针对连续还是离散变量的传统方法,均无法完成对于这类特定取值空间的事故数据的数据分析和建模,需要探索新的思路.

一般而言,对一个数据序列进行的分析可分为"描述性分析"(Descriptive Analysis)及"推理性分析"(Inferential Analysis)两大类[1],前者只能解决数据"看起来"呈现什么形态的问题,后者可解决结论和外推的问题. 对于上述小数值交通事故的数据而言,这两类分析都无法采用传统和简单的方法,必须探究一些新型的方法来完成. 传统的交通安全建模方法,包括获得最成熟应用、由 Ezra Hauer 教授所开创的广义线性模型(Generalized Linear Model, GLM)型式[2-3]及一些变化型式,都只适用于较长期(以年计),但不适用于短时(例如以天或小时计)的事故数据. 甚至于一些最近的作为 GLM 扩展形式而建立起来的广义线性估计方法(Generalized Estimation Equation, GEE)[4],尽管其有着更复杂和更灵活的模型结构,但依然需要观测数值必须积累到一定程度,才能满足模型分析的有效性.

另外一种传统的建模分析方法,罗基模型(Logit Model),只适用于数值非常有限的离散变量,例如一个城市的每日死亡交通事故数,特定公路区段上的每日驶出路外交通事故等. 通常而言,它们或者是二元变量,即 0/1 取值,或者是分类变量但取值仅为 0、1、2、3 等. 罗基模型典型的用途是用以辨别二元变量或者分类变量的致因并定量地预测该二元或者分类变量出现特定数值的概率[5].

综上所述,当前体系中尚未发现适用于短时、小数值交通事故数据的描述和推理建模分析的手段,需要开拓一种新的途径. 本论文的基本思路是尝试使用一些交通工程领域刚刚涌现的数据分析方法,例如数据可视化,来描述该种数据的形态、并挖掘其形态规律,进而辨别其相关要素和各要素间的关系结构,从而明确该类交通事故的成因.

1 数据样本

本论文研究所用的数据包括了样本城市的交通事故数据及其可能影响因素的数据,其中,交通事故数据选择为该市的"每日 FI 事故数",而影响因素则涵盖了每日的日历特征数据(年、月、日、周日、节假日等)、天气观测数据、天气预报数据等(以下有关数据的描述省略"每日"字眼,除非特别说明,所有数据均表示每日数据).

图 1 描述了上述数据的来源. 其中,FI 事故数来源于样本城市的"机动车交通事故信息库"[6];天气的历史观测数据由"加拿大环境部"网站下载[7];定量(本论文中天气对交通事故数的影响分析,必须用定量的天气指标,简单采用"阴、晴、雨、雪"等分类天气指标过于粗略,无法精确反映对交通事故数波动的影响)的天气预报由艾尔伯塔大学采用专

门针对样本城市进行了标定的“天气研究与预报模型”来计算获取，并持续提供更新的数据.

最终，上述 4 个数据源通过日期为关键指标而整合为一体，形成一个整体的数据序列，这一数据序列包括了每日事故及各种潜在的相关影响因素数据，以日期为标准排列.

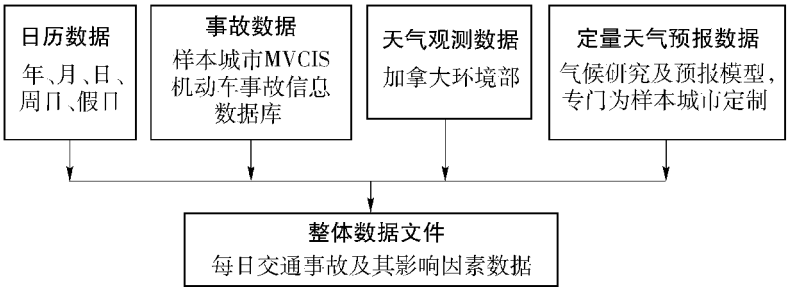


图 1　4 种交通事故及其相关因素数据源及其结构

2　技术路线

一种新型的数据分析思路、数据挖掘，可被用于处理复杂、大型的数据并从中发现规律. 在计算机科学领域中所称的数据挖掘[8]，在数据库领域中又被称为“知识发现”，是一个在大容量数据中发现模式、规律及关系的过程. 数据挖掘是一个集合，它组合了统计工具和人工智能分析工具（例如神经元网络和机器学习等），同时与数据库管理相结合，分析大型数字组合或者数据序列.

本论文以样本城市中 FI 事故为实例，探讨应用数据挖掘方法来分析交通事故自身的变化规律及其成因. 由上所述，FI 事故数据的特点是数值分布范围狭窄但波动幅度大，同时相关因素众多，FI 事故数与相关因素之间、相关因素彼此之间的关系难以辨别. 根据这些特点，本论文有针对性地提出采用“数据可视化”方法实施数据分析，原因是数据可视化有利于展现变量的变化趋势、众多相关因素之间的互动规律以及复杂的相关关系结构. 适用于众多变量的数据可视化方法归属于“高维数据挖掘”，在本论文研究中，具体的数据分析过程采用开源统计分析软件 R 中一系列的数据可视化工具来完成[9].

根据数据分析阶段的不同，又将这些数据可视化分析过程归类为描述性分析、推理性分析前后 2 个阶段. 其中，描述性分析为第 1 阶段，探索事故的规律和各因素彼此关联的基本形态，为第 2 阶段的分析提供初始的备选变量；第 2 阶段的推理性分析将更为深入、全面和系统地分析事故与各相关因素间的因果关系及其脉络结构，并获得和成因对 FI 事故数的定量化影响程度.

3　应用数据可视化方法的交通事故描述性分析

3.1　交通事故自身的一元数据可视化分析

每日交通事故自身的变化趋势，即其作为一元数据的可视化，可藉由“时间序列”[10]的“数据分解”来进行，即将其分解成为“趋势”、“周期”、“随机”3 个成份[11]. 图 2 描绘了事故数据拆分为 3 个部分后各自的变化趋势，其中“周期”部分展示了事故数据自身是一个带有周期重复性的时间序列数据.

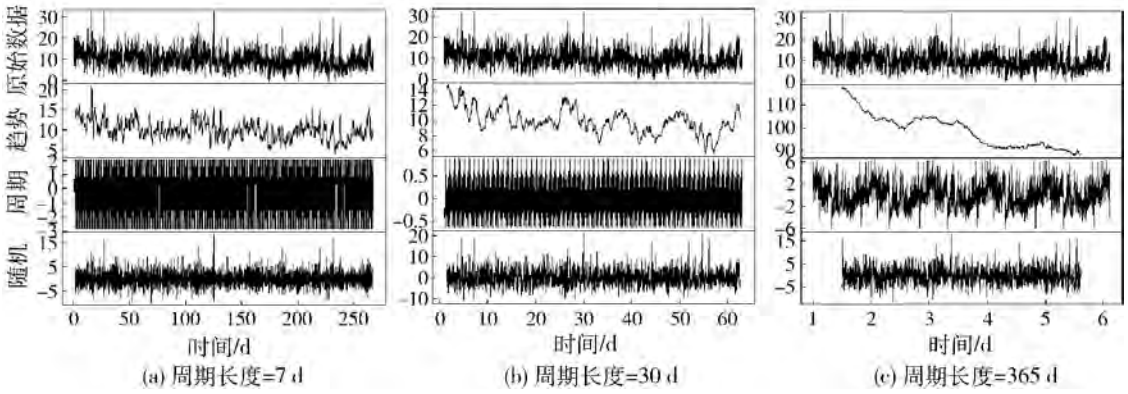
数据分解还可以更进一步地揭示事故的深层次规律特征，尤其是对于“周期”成份，如图 3 所示，可以进一步揭示事故数据周期变化的长度. 从图 3 中可以得出该事故数据最为明显的重复变化是以 1 周为周期的变化，按月变化的趋势不明显，而按年变化体现出一定的四季变化规律，但也有一定的波动性. 综合以上分析获得的结论是：事故与周日分布直接相关，其首要的周期变化长度为一周.

具体到该样本数据，则可见周一至周五，事故数逐步上升，到周五则达到峰值，然后在周末回落，其中周日的交通事故数最低.

3.2　交通事故及其相关因素的二元数据可视化分析

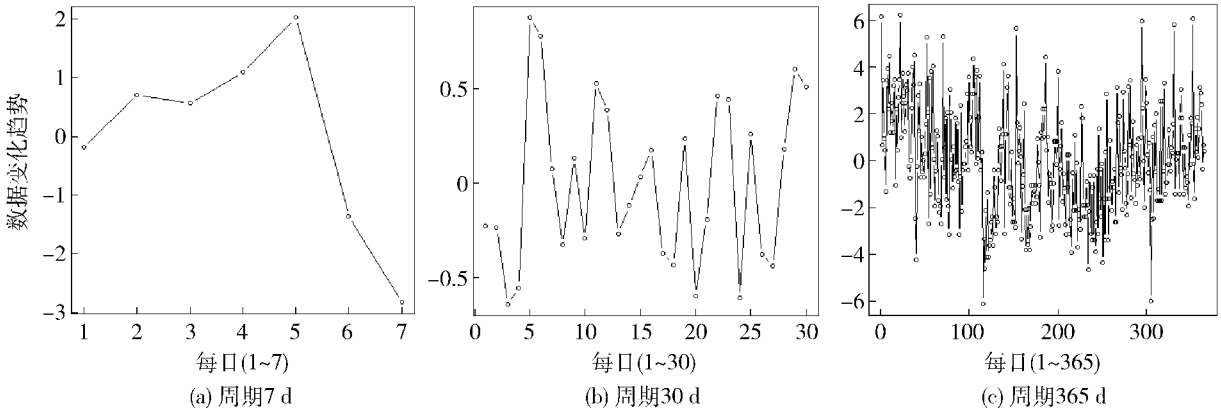
交通事故数据自身的变化趋势，即上述一元数据的可视化，揭示了事故的时间序列规律、周期变化特征以及周期长度，从中可以推断出事故与“周日”这一因素具有较强相关性，但事故可能的相关因素有许多，如图 1 所示，可能包括其他的日历参数，或者天气因素等，若要进一步厘清事故的相关因素，以及各要素间相互的关联特征，则必须进行二元数据的分析，即将两个变量进行相关分析.

相关分析既可用于识别某一个结果变量，在此



(a) 周期长度=7 d (b) 周期长度=30 d (c) 周期长度=365 d

图 2 交通事故的数据分解



(a) 周期7 d (b) 周期30 d (c) 周期365 d

图 3 交通事故数据的周期分析

研究中为 FI 事故数,和它的潜在解释性因素或称为相关因素之间的相关特性,也用于识别两个相关因素之间的相关特性. 对于前者,相关分析可初步筛选与结果变量有较好相关性的因素,为下一步的推理分析寻找潜在预测变量;对于后者而言,其根本目的在于判别彼此具有强相关性的两个相关因素,然后在未来的建模过程中避免将这两个因素同时纳入一个方程. 具有强相关的两个因素间的关系,称为"共线性". 共线性对模型具有负面干扰,它将导致回归模型的系数不能正确代表变量与结果变量之间的关系,因此在建模中应避免共线性现象[12].

由于 FI 事故数的潜在影响因素众多,对它与影响因素之间的相关性、相关因素彼此之间的共线性的分析,宜采用擅长高维变量的数据可视化的方法进行. 在研究采用散点图矩阵与相关系数表联合使用的技术路线完成二元因素间的相关分析. 它们均为典型的高维数据可视化工具,擅长一次性展示众多变量之间的两两相关关系. 其中,本论文采用"彩色加强版"的散点图矩阵,以色彩配合相关系数表中的数值,表征 1 对因素之间的相关性强弱程度.

采用 R 软件中的分析模块,针对样本城市的 FI 事故数及其潜在影响因素的彩色散点图矩阵见图 4 所示,而它们的数量化相关系数表详见表 1[13].

散点图矩阵一次性展示多维数据序列中任意 1 对数据项之间的相关关系. 在图 4 中,各单元中的"颜色代码"代表着这 1 对变量的相关程度. 这个颜色代码与表 1 中的"相关系数"取值彼此呼应. 表 1 中相关系数大于 0.6 或者小于 -0.6 的,代表着强相关性,以加粗字体表示;相关系数处于 -0.6 ~ 0.6,代表着相关性不强. 另外,相关系数为正值的,意味着两个变量之间存在的是正相关性关系,也即当一个变量增加,另一变量也增加,例如表中的最高气温和最低气温;相关系数为负值,意味着两个变量之间存在负相关性,即当一个变量增加,另一变量会减小,例如最高气温和地面积雪. 不论相关系数为正或负,只要其绝对值大于 0.6,并且在图 4 中的相应单元为红色,即代表两变量间存在强相关性. 例如,最高气温、最低气温和平均气温这 3 个变量两两之间存在着非常强的相关关系. 此外,降雨量和降水量之间有较强的相关性,而地面积雪、季节与上述

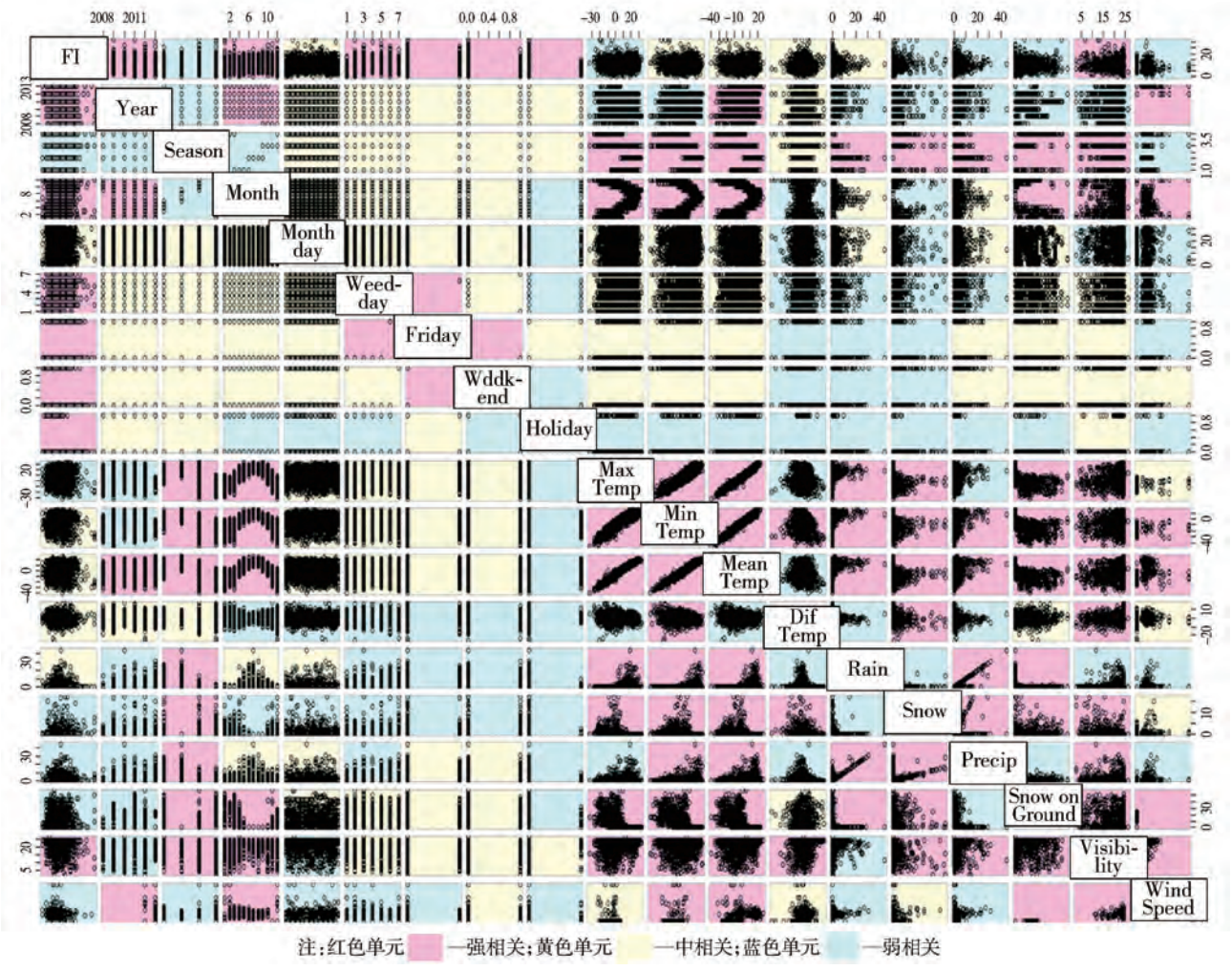


图 4　彩色散点图矩阵

3 个气温变量之间均存在着负相关的关系.

除了上述的相关变量外，从图 4 中可以看出，其他各变量之间没有存在着显著的相关趋势，代表着它们之间具有相对独立性[14,13,15].

3.3　交通事故及其相关因素的多元数据可视化分析

上述相关分析揭示的是两个变量之间的关联和变化规律，数据可视化中还有一些工具可用于多个数据项之间的互动特征的分析，图 5 为带有“竖向垂线”和“回归板”（即三维数据间的回归趋势平面）的“三维散点图”，它揭示出 FI 事故数与“降水量”、“周日”3 个变量间的相互关系和变化趋势. 由散点在三维空间上的分布规律，加上以垂线强化视觉效果，尤其是以回归板突显出相关趋势，最终可算出 FI 事故数随着降水量的增加而提高的规律，同时也可得到 FI 事故数自周一至周五逐渐上升，周五达到高峰，周末降低的总体趋势.

数据可视化描述性分析，能够从事故自身的一元化数据，事故及其相关因素的二元相关性、三维相关性等多个角度进行数据形态、分布和互动规律的探索，但其具有两方面的局限性：①无法系统性地展示所有相关因素与事故指标间内在的规律，尤其是其中复杂的关系结构；②描述性分析更多地揭示样本数据的视觉特性，而不是结论性的规律，即一般无法获得具有高度抽象和概括性的，具备趋势外推和预测功能的结果. 也就是说，它更多地是“展示”，而不是“推理”.

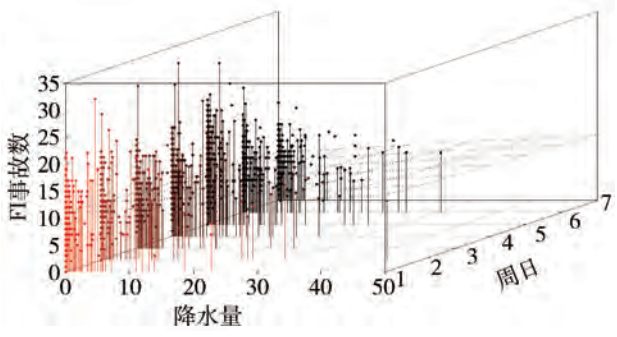
4　交通事故的成因推理分析

以上的描述性分析将 FI 事故及其相关因素数据的初步形态及相关规律进行了详尽的描绘，在此基础上，还应进一步采用“推理性分析”的方法，揭示所有相关因素与事故之间的内存关系，建立关系结构图，并定量化分析成因对结果变量的影响程度，从而完成对事故成因的结论性分析.

本论文采用基于数据可视化思维模式的推理性分析方法，例如“图形模型”来完成事故的“成因推理”.

表 1 相关系数分析表

	伤亡事故数	年	季节	月	月日	周日	周五	周末	假日	气温/度			温度差/度	降雨量/mm	降雪量/mm	降水量/mm	地面积雪/cm	能见度/km	风速/(m·s^{-1})
										最高	最低	平均							
伤亡事故数	**1.0**	−0.2	0.0	0.2	0.0	0.2	0.2	−0.3	−0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	−0.1	−0.1
年	−0.2	**1.0**	−0.1	−0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.1	0.1	0.3
季节	0.0	−0.1	**1.0**	0.0	0.0	0.0	0.0	0.0	0.0	**−0.6**	**−0.6**	**−0.6**	0.0	−0.2	0.1	−0.1	0.5	−0.2	−0.1
月	0.2	−0.2	0.0	**1.0**	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.2	−0.1	0.0	−0.1	0.0	−0.4	0.1	−0.2
月日	0.0	0.0	0.0	0.0	**1.0**	0.0	0.0	0.0	−0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
周日	0.2	0.0	0.0	0.0	0.0	**1.0**	0.4	0.0	−0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−0.1
周五	0.2	0.0	0.0	0.0	0.0	0.4	**1.0**	−0.3	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	−0.1
周末	−0.3	0.0	0.0	0.0	0.0	0.0	−0.3	**1.0**	−0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
假日	−0.1	0.0	0.0	0.0	−0.1	−0.1	0.0	−0.1	**1.0**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
最高气温/度	0.0	0.0	**−0.6**	0.1	0.0	0.0	0.0	0.0	0.0	**1.0**	**0.9**	**1.0**	−0.1	0.2	−0.3	0.1	**−0.6**	0.4	0.0
最低气温/度	0.0	0.0	**−0.6**	0.1	0.0	0.0	0.0	0.0	0.0	**0.9**	**1.0**	**1.0**	−0.2	0.3	−0.2	0.2	−0.6	0.2	0.1
平均气温/度	0.0	0.1	**−0.6**	0.2	0.0	0.0	0.0	0.0	0.0	**1.0**	**1.0**	**1.0**	−0.1	0.2	−0.2	0.1	**−0.6**	0.3	0.2
温度差/度	0.0	0.0	0.0	−0.1	0.0	0.0	0.1	0.0	0.0	−0.1	−0.2	−0.1	**1.0**	0.0	−0.1	−0.1	0.0	0.1	0.0
降雨量/mm	0.0	0.1	−0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.2	0.0	**1.0**	0.0	**0.9**	−0.2	−0.1	0.1
降雪量/mm	0.1	0.0	0.1	−0.1	0.0	0.0	0.0	0.1	0.0	−0.3	−0.2	−0.2	−0.1	0.0	**1.0**	0.4	0.2	−0.5	0.0
降水量/mm	0.1	0.1	−0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.1	−0.1	**0.9**	0.4	**1.0**	−0.1	−0.2	0.1
地面积雪/cm	0.0	0.1	0.5	−0.4	0.0	0.0	0.0	0.0	0.0	**−0.6**	**−0.6**	**−0.6**	0.0	−0.2	0.2	−0.1	**1.0**	−0.3	**NA**
能见度/km	−0.1	0.1	−0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.4	0.2	0.3	0.1	−0.1	−0.5	−0.2	−0.3	**1.0**	0.1
风速/(m·s^{-1})	−0.1	0.3	−0.1	−0.2	0.0	−0.1	−0.1	0.0	0.0	0.0	0.1	0.2	0.0	0.1	0.0	0.1	**NA**	0.1	**1.0**



注:图中“周日”为 1(星期一)至 7(星期日).

图 5 带有竖向垂线和“回归板”的三维散点图

4.1 成因推理的基本概念及方法

成因推理是现代数据挖掘中的一组方法集合的总称,在统计软件 R 中提供了成因推理的两项功能:“成因结构学习”以及“因果效应估计”[16].

本论文的研究首先厘清 FI 事故的影响因素及其相互关联的结构图. 具体而言,本次分析将应用统计软件 R 中的“PC”算法[17],以便发现哪些因素可能是、哪些因素可能不是事故的成因,这些成因关系以“有向非循环图”的形式来表达. 在 DAG 图中,1 个节点代表 1 个变量,1 个有向边代表 1 个“因果”关系. 在 R 中,这一算法的最终输出结果为“完成的部分定向非循环图”,用以描述数据中的“条件独立信息”. 在 CPDAG 图中,因果(即非独立)关系表现为有向边,而独立关系表现为无向边(在 R 的具体输出图中,以双向边来代替无向边).

上述“成因结构学习”的过程描述了因果关系,但不能回答因素间影响程度的问题,即定量化的成因效果水平. 这一问题需由“因果效应估计”来实现. 这一过程的实质为定量化两个变量间的因-果关系. 假设有 2 个随机变量 V_x 和 V_y,V_x 为“因”,V_y 为“果”;量化 V_x 与 V_y 之间的因果关系,其过程是强制 V_x 取值为 x,获取 V_y 的状态,并将之与强制 V_x 为 $x+1$ 或者 $x+\delta$ 时 V_y 的状态进行对比,以便分析随机变量 V_x 在被强制取值下对于另一个随机变量 V_y 分布的

影响[16]. 在这一“定制”过程之后随机变量的分布可表述为：

$$P[V_y \mid \mathrm{do}(V_x = x)] \tag{1}$$

这是一个与条件分布 $P[V_y \mid V_x = x]$ 不同的过程[18]. 通常情况，我们用“平均变化率”作为变量 V_x 作用在变量 V_y 之上的“因果效应”通用指标[18]：

$$\frac{\partial}{\partial x} E[V_y \mid \mathrm{do}(V_x = x)] \tag{2}$$

在统计软件 R 中，上述过程采用“IDA”方法[16]来实现，最终的输出为定量化的系数值，评价特定因素之间的因果效应的定量化程度.

4.2　事故及其相关因素的成因结构学习

在描述性分析的基础上选取潜在的事故相关因素，采用 PC 算法，可得到样本城市的 FI 事故的成因及其相互“因-果”关系流程的结构[16-17]，最终结果以 CPDAG 框图的形式显示，见图 6 所示.

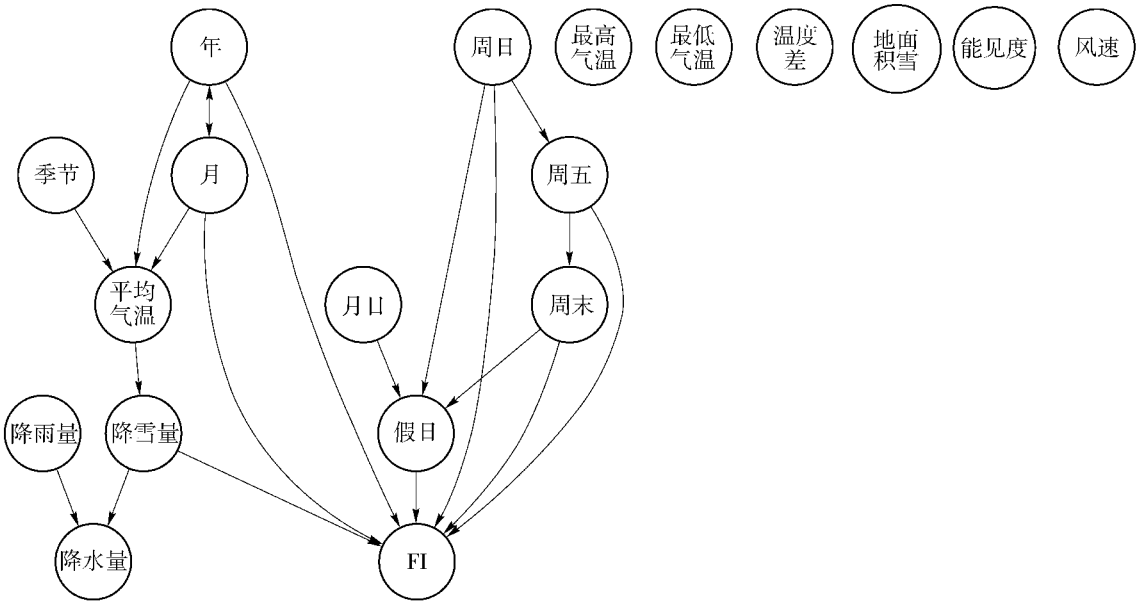


图 6　FI 事故的成因结构 CPDAG 图

由图 6 可看出，首先，可以剔除若干“可能不是”FI 事故成因的 6 个因素，包括最高气温、最低气温、温度差、能见度和风速. 在这个分析中，“成因结构图”是基于数据而获得的抽象的“因-果”关系结构图，因此也会体现出数据本身的一些前提条件或者局限性. 例如，“能见度”和“风速”是由于在样本数据中的有效样本量不足，无法满足统计有效性而被剔除在结构图之外的. 另外，由于和“平均温度”之间存在强相关性，“最高气温”、“最低气温”和“温度差”这 3 个变量最终也被排除在结构图之外.

在此基础上，图 6 中的 CPDAG 图描述了各变量间的因果关系，特别地，体现出了相关因素和目标变量-“FI 事故数”-之间的成因关系. “FI 事故数”共有 7 条直接“因果关联”的 CPDAG 图的有向边，而其起始节点分别为“降雪量”、“月”、“年”、“假日”、“周日”、“周末”和“周五”. 其中，“年”、“月”和“周日”3 个变量没有上游的“父母”成因变量，而其他的 4 个变量则同时又是其 1 个或者多个成因变量的“结果变量”，综合考虑这些上游的因果关联，最终“FI 事故”的成因除上述 7 个变量外，还可能间接地包括“平均温度”、“季节”、“月日”等. 这些变量共同构成了一个成因结构流线图，即图 6.

4.3　事故及其相关因素的因果效应估计

在“成因结构学习”的基础上，选取 FI 事故直接或间接的成因要素，进一步采用 R 统计软件中的“IDA”算法，定量化地计算这些要素与 FI 事故数的“因果效应”系数，结果见表 2 所示. 表 2 中每个成因要素所拥有的估计值的个数，取决于该要素与 FI 事故数之间的 DAG 图的个数，由于我们无法判断哪个 DAG 是“真的”成因 DAG，所以最终将所有可能的估计值都输出出来.

假设表 2 中的成因要素为 V_x，FI 事故数为 V_y，则表 2 中的因果效应值代表着下述回归公式中对应着要素 V_x 的系数：

$$\mathrm{lm}(V_y \sim V_x + \mathrm{Pa}(V_x)) \tag{3}$$

其中 lm 为线性回归模型；V_y 为 FI 事故数；V_x 为 FI 事故数的可能成因变量；$\mathrm{Pa}(V_x)$ 为成因变量 V_x 在 DAG 图中的“父母”变量（即有向边的上游节点）

表 2 FI 事故的因果效应估计值

成因要素	因果效应估计值 1	因果效应估计值 2
年	-0.699 203 6	-0.629 465 2
季节	0.162 699 3	
月	0.198 75 76	0.136 247 7
月日	0.007 222 22	
周日	0.330 304 1	
周五	1.913 913	
周末	-2.631 276	
假日	-3.785 783	
平均气温/度	0.020 510 89	
降雪量/mm	0.298 827 7	

由式(3)可知,表 2 中每个估计值,从数学角度代表着各成因要素发生了一个数量单位的变化时,所引发的 FI 事故数的波动.

以下逐一具体分析各要素之间与 FI 事故数的因果效应水平. 首先"年"有 2 个估计值,但考虑到 2 个值均远离零且彼此接近,因此可认为年与 FI 事故数之间有确切的因果效应,其效应值位于 -0.63 ~ -0.70 的区间;"季节"与 FI 事故数有一个正相关的因果效应值;"月"与 FI 事故数有 2 个效应估计值,但均显著大于零,因此也可认定它们之间存在着正相关的因果关系;"月日"与 FI 事故数之间的效应值接近零,可以认为不存在明显的因果关系;"周日"与 FI 事故数存在着相对较明显的正相关的因果效应;"周五"是 FI 事故数上升的显著成因,与之相对,"假日"与"周末"对应着 FI 事故数的下降;"平均温度"与 FI 事故数的因果关系较弱,而"降雪量"则较为显著地引发 FI 事故数的提高.

综上所述,在各种日历参数之中,假日与周末引发事故下降,而周五引发事故上升,不论其效应为正或者负效应,它们的绝对效应值均较高. 另外,事故数也呈现周一至周五递增的关系. 在天气要素之中,"降雪量"较明显地引起事故上升.

5 结论

本研究经历了"描述性分析"和"推理性分析"两个阶段,数据可视化和图形模型的方法贯穿着整个的研究过程. 其中,描述性分析完成初步工作,通过交通事故数一元、二元、多元等不同维度的数据可视化分析,突显其发展变化趋势及潜在相关特征,初步选定事故的相关要素.

在描述性分析成果的支持下,以其初选获得的相关因素和事故数相关联,着重进行了成因推理的分析. 这一过程通过"图形模型"中的 2 个相关算法来实现. 其中,"成因结构学习"算法构建事故及其要素之间的因果关系网络、建立关系结构框图. "因果效应评估"过程则更进一步地量化了各要素与事故之间的因果效应值,即要素对于事故数影响的数量化程度.

本研究直接建立了样本数据的结果变量"FI 事故数"的生成原因及各原因的数量化影响程度. 成果中的"因果效应估计值"在数学层面反映的是每个事故成因发生一个数量单位的变化时所能引起的事故数的变化数值. 通过研究,FI 事故数的成因主要在于日历参数和天气因素. 在各种日历参数中,FI 事故数最为显著的成因依次为假日、周末及周五. 某天为假日或者周末时,将分别带来约 3.8 和 2.6 的 FI 事故的降幅;而当某天为周五时,则带来约 1.9 的 FI 事故的上升. 在天气因素中,对于样本城市这样的北方城市而言,降雪量是最为显著的事故成因,日均降雪量每增加 1 cm,将增加约 0.3 的 FI 事故.

在理论层面,本研究突破了传统模型只能描绘长期事故变化趋势的局限,建立了短时事故成因推理方法. 在应用层面,本研究成果对于在较短时间维度上的交通执法、运营、养护等有着直接的指导作用. 交通执法部门的警力部署和物资储备、交通运行与实时信息发布(例如可变情报板)、针对瞬时事件的应对策略以及日常养护的预案制订等,均可以依据本研究所确定的事故成因结构和定量化的因果效应值而制订,从而为短时交通安全的执法、管理与控制提供数量化的信息与决策依据. 因此,本论文的研究在交通安全执法、管制和控制的实践将发挥积极作用,创造出良好的社会效益和经济效益.

致谢:作者感谢加拿大埃德蒙顿市政府交通安全办公室的 Stevanus Tjandra 博士及 Gerry Shimko 主任所提供的基础数据,以及对论文研究提供的建议. 同时,也感谢加拿大艾尔伯塔大学地球与空气科学系的教授 Gerhard Reuter 博士与 Clark Pennelly 先生所提供的天气预报数据.

参考文献:

[1] Wallace D. Descriptive versus inferential statistics, Lesson 1: Introduction. Lecture Note of Statistics for Psychology

[EB/OL]. Faytteville State University, North Carolina, United States [2014-08-04]. http://faculty.uncfsu.edu/dwallace/Lesson%201.pdf.

[2] Hauer E, Bamfo J. Two Tools for Finding What Function Links the Dependent Variable to the Explanatory Variables [C/OL]. Published in Proceedings of ICTCT 97 Conference, November 5-7 1997, Lund, Sweden. [2014-04-13]. http://www.oocities.org/hauer@rogers.com/Pubs/ICTCT97TwoTools.pdf.

[3] Hauer E. Observational Before-after Studies in Road Safety [M]. Bingley, United Kingdom: Emerald Group Publishing Limited, 2007.

[4] Chin H C, Huang H. Modeling multilevel data in traffic safety: a bayesian hierarchical approach, chapter 3 of transportation accident analysis and prevention[M]. New York, United States: Nova Science Publishers, Inc., 2008.

[5] Zhang H. Identifying and Quantifying Factors Affecting Traffic Crash Severity in Louisiana, Ph. D. Dissertation [D]. Baton Rouge, Louisiana, United States: Louisiana State University, 2010.

[6] Office of Traffic Safety. Motor Vehicle Collision 2012, Annual Report, City of Edmonton [EB/OL]. [2013-07-25]. http://www.edmonton.ca/transportation/OTS_Motor_Vehicle_Collisions_2012_Annual_Report.pdf.

[7] Environment Canada. Daily Data Reports, Environment Canada, Government Canada [EB/OL]. [2013-06-24]. http://climate.weather.gc.ca/climateData/dailydata_e.html?timeframe=2&Prov=ALTA&StationID=50149&dlyRange=2012-09-01%7C2012-11-08&cmdB1=Go&Year=2013&Month=6&cmdB1=Go.

[8] Clifton C. Encyclopædia Britannica: Definition of Data Mining [EB/OL]. [2014-08-04]. http://www.britannica.com/EBchecked/topic/1056150/data-mining.

[9] Torgo L. Data Mining with R-Learning with Case Studies [M]. Boca Raton, Florida, United States: Chapman & Hall/CRC, Taylor & Francis Group, 2011.

[10] Easton V J, McColl J H. Time Series Data. Statistics Glossary, v 1. 1 [EB/OL]. [2012-07-10]. http://www.stats.gla.ac.uk/steps/glossary/time_series.html.

[11] Coghlan A. Time Series 0. 2 Documentation [EB/OL]. [2012-07-03]. http://a-little-book-of-r-for-time-series.readthedocs.org/en/latest/index.html.

[12] Mela C F, Kopalle P K. The Impact of Collinearity on regression analysis: the asymmetric effect of negative and positive correlations [J]. Applied Economics, 2002, 34: 667-677.

[13] R Development Core Team. Correlation, Variance and Covariance (Matrices), R Documentation [EB/OL]. [2013-12-27]. http://stat.ethz.ch/R-manual/R-patched/library/stats/html/cor.html.

[14] King W B. R Tutorials: Simple Linear Correlation and Regression [EB/OL]. [2013-12-27]. http://ww2.coastal.edu/kingw/statistics/R-tutorials/simplelinear.html.

[15] Lund A, LundM. Laerd Statistics: Pearson Product-Moment Correlation [EB/OL]. [2013-12-27]. https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php.

[16] Kalisch M, Machler M, Colombo D, et al. Causal Inference Using Graphical Models with theR Package pcalg [J]. Journal of Statistical Software, 2012, 47 (11): 1-26.

[17] Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search [M]. 2nded. Cambridge, Massachusetts, United States: MIT Press, 2000.

[18] Pearl J. Causality [M]. Cambridge, United Kingdom: Cambridge University Press, 2000.