基于大数据分析的交通事故成因分析

Cause analysis of traffic accidents based on big data technologies

汇报人：贾鹏飞

单位：中国城市规划设计研究院

目录
CONTENTS

PART ONE

研究背景

研究背景

PART ONE

我国交通事故导致死亡人数达世界10%，而我国机动车拥有量只有世界数量2%

PART TWO

2001-2011年，我国交通事故频率有逐年下降趋势，但死亡人数超过10万人

PART THREE

道路交通安全成为关系人民群众生命安全、关系社会经济协调发展的大众社会问题

研究背景

研究方法

国内外交通事故研究可分为：
描述性研究和机理模型构建两类

统计建模

机理建模

研究背景

研究方法

国内外交通事故研究可分为：
描述性研究和机理模型构建两类

基本数据

事故记录数据主要涉及人、车、天气，道路，用于数据预处理、事故发生地定位

Driver

Car

Weather

Road

研究背景

研究方法

国内外交通事故研究可分为：
描述性研究和机理模型构建两类

基本数据

提供事故记录数据主要涉及人、车、天气状况，用于数据预处理、事故发生地定位

辅助数据

基于百度兴趣点(POI)确定事故所在道路状况，展开深入分析，分析交通事故成因

研究目标

01

Accident Perpetrator

分析肇事者特征，为交通部门提供驾车人行为习惯

02

Accident Models

分析各个车型分布特征，为交通管控车型提供帮助

03

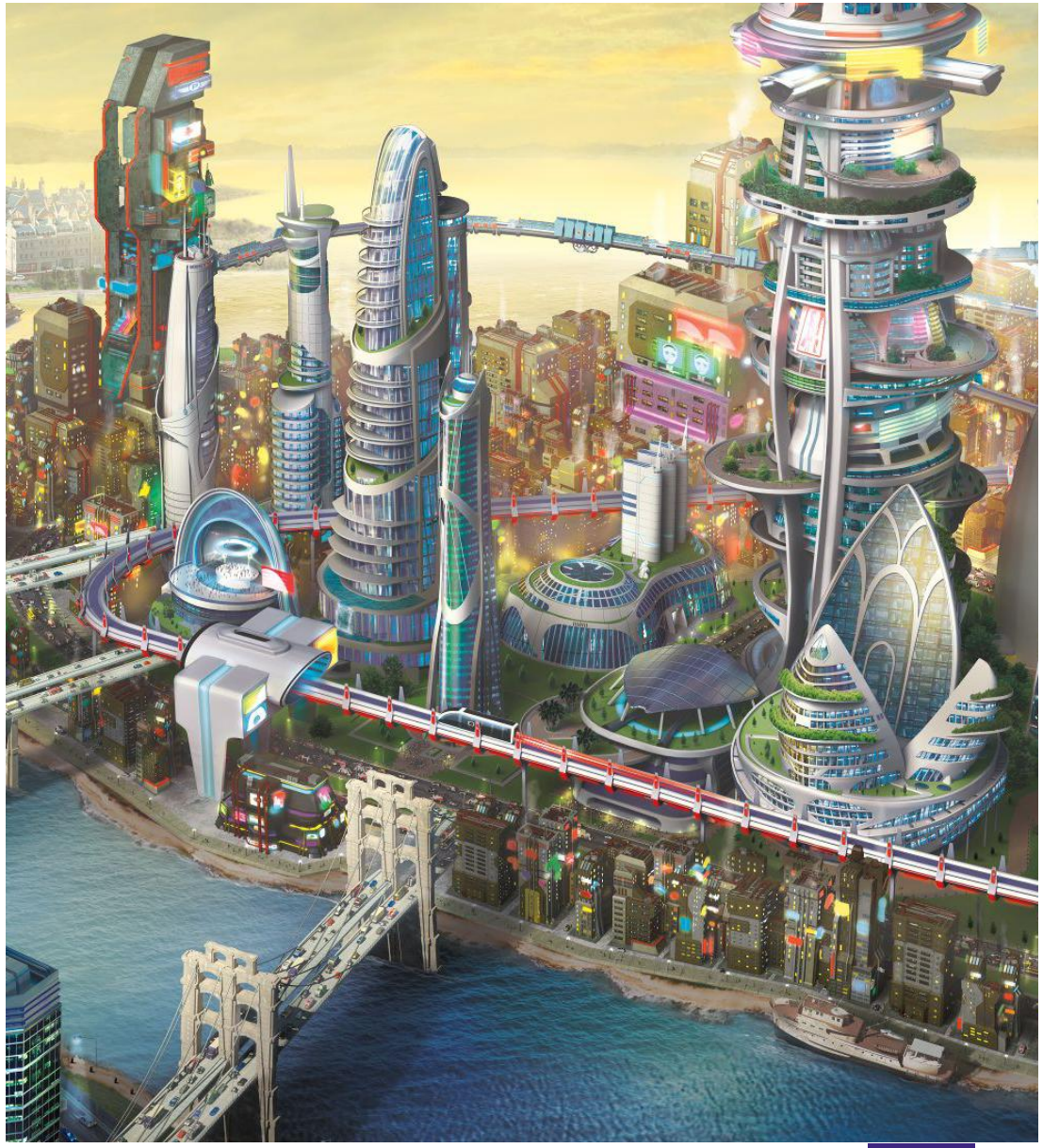
Accident Weather

分析事故发生天气状况，为交通管理提供决策依据

04

Accident Period & Location

分析事故时段及地点分布，解释事故发生综合原因



PART TWO

数据使用

数据使用

事故记录数据

百度POI数据

贵阳市GIS数据

数据记录表
天气记录表
交通违法记录表

22大类
122823兴趣点

道路数据
高程数据

数据使用

- 年龄
- 性别
- 驾驶年龄
- 责任认定
- 毕业驾校
- 违法记录次数

- 车牌照所属地
- 车辆颜色
- 车辆产地
- 车辆类型

- 变天前的天气
- 变天后的天气
- 一天最高温度
- 一天最低温度
- 毕业驾校
- 违法记录次数

- 事故发生时段 (month/day/hour)
- 事故发生路段
- 事故碰撞类型

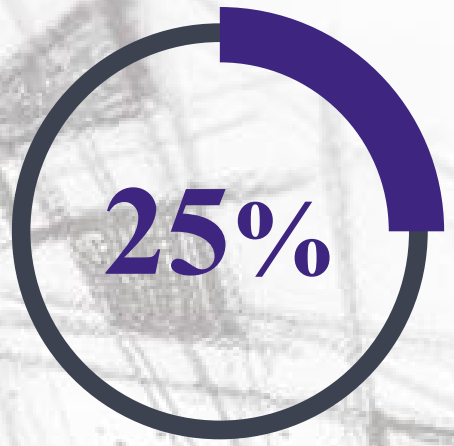
数据使用

POI数据介绍

PART THREE

技术使用

技术使用

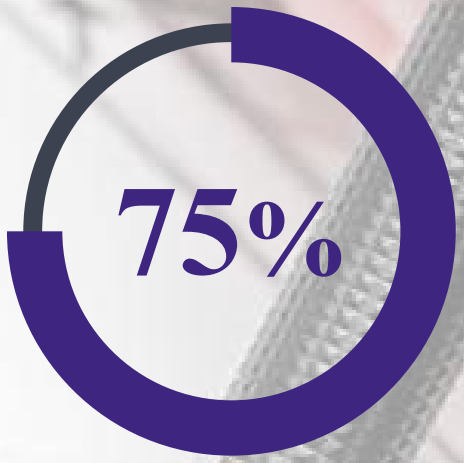


数据预处理

年龄、驾龄等简单变量计算(Matlab实现)

事故发生地点定位

事故记录表与百度POI地点匹配(Python)



描述性统计分析

探索交通事故发生频率与各类因素相关系

数据挖掘分析

广义线性模型、决策树回归、随机森林

模型介绍

广义线性模型

- **GLM**，广泛应用统计模型，经典线性模型推广，在交通、医学、金融学、保险和生物等统计领域中有非常应用背景。在使用GLM进行分析交通事故，随机成分多服从泊松分布/正态分布

具有广泛应用统计模型，为经典线性模型推广

决策树模型

- **Decision Tree**，通过归纳和提炼现有数据规律，并用于新数据分类预测的一种非参数方法；没有特定的函数形式，且不需要任何样本数据先验分布假设。算法以信息增益率为分类标准

本研究采用对数线性模型，因变量*Y*服从泊松分布

模型介绍

随机森林算法

- **Random Forest**，一种统计学习理论，采用bootstrap重采样方法从原始样本中抽取多样本，对每个bootstrap样本进行决策树建模，组合成多棵决策树进行预测，通过统计得预测结果

统计学习、bootstrap重采样、决策树

随机森林算法

- **Random Forest**，每一棵决策树就是一个精通某一个较窄领域“专家”，形成较多精通不同领域“专家”；对于一个新问题，可以用不同的角度去看待，最终由各个专家投票得到结果

决策领域“专家”，“专家”投票

PART FOUR

创意描述

创意描述

Step One

数据匹配

结合POI数据，根据地名匹配确定事故发生经纬坐标，确定两组数据完全匹配

Step Two

空间分析

结合道路矢量数据和高程数据，得到事故发生基本状况

Step Three

描述统计

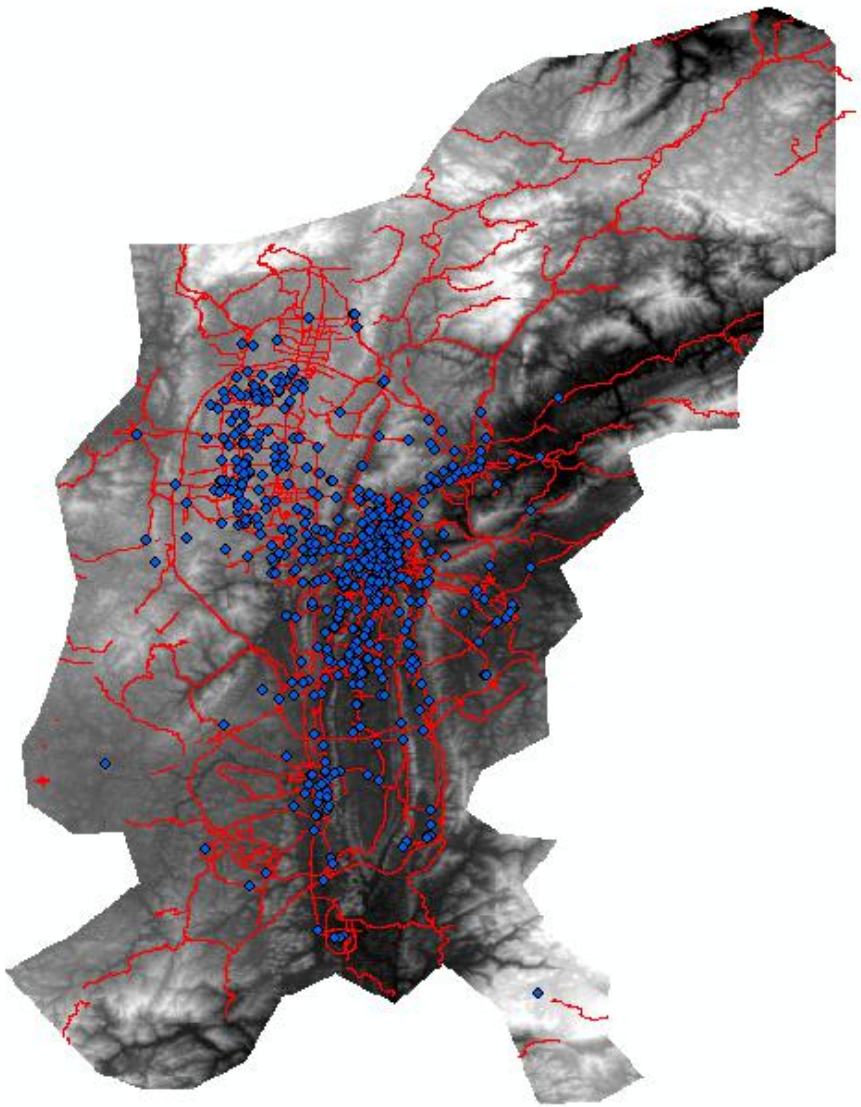
探索交通事故发生频率与时间、事故人、事故车辆状况以及天气状况之间的相关系

Step Four

数据挖掘

应用广义线性模型、决策树回归、随机森林回归建立道路发生追尾次数模型并验证

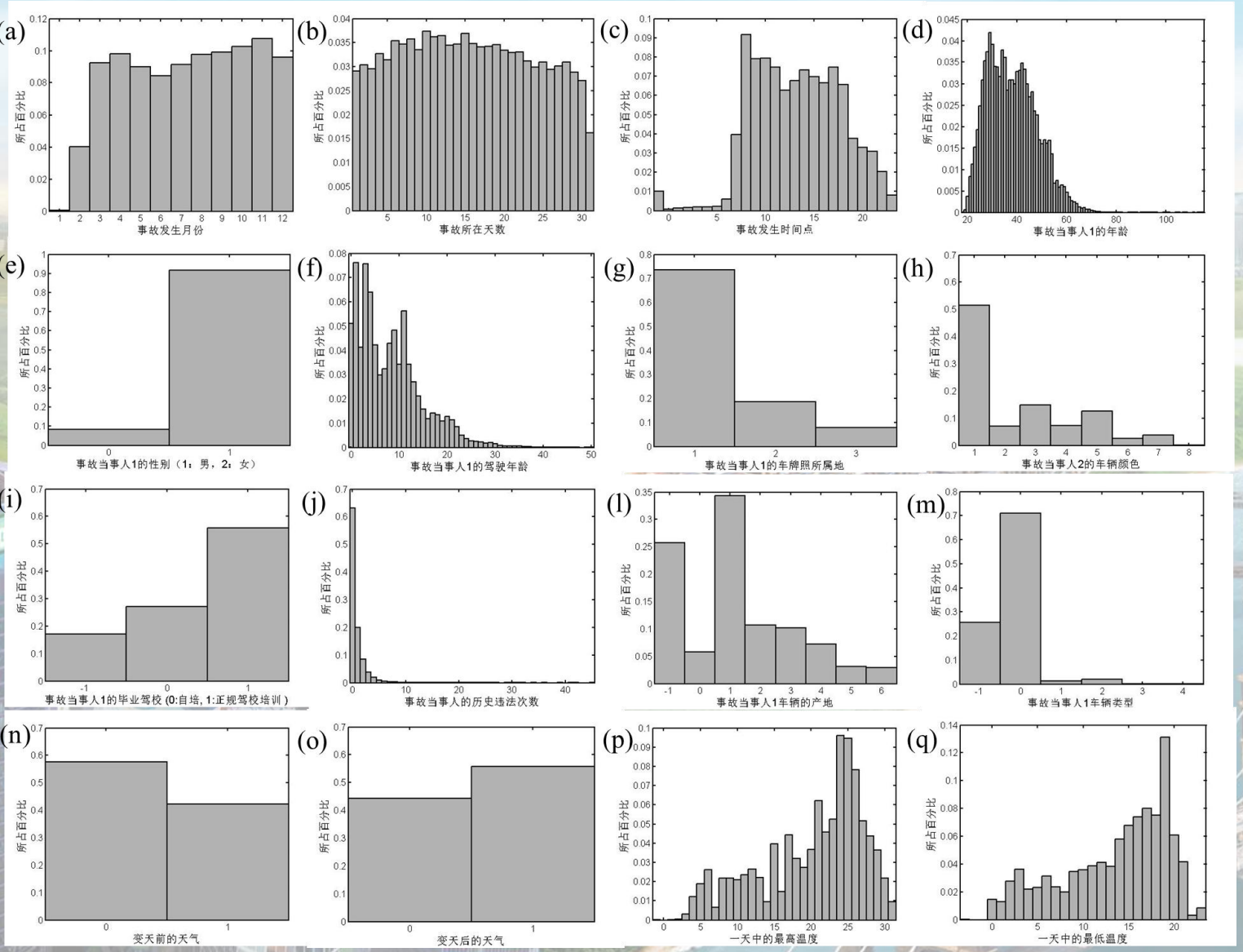
创意描述



数据匹配

图中蓝点表示定位的事故记录数据（记录数据地点描述与百度POI地名完全匹配），主要分布于市中心处，这可能与记录人员对这块区域最熟悉有关

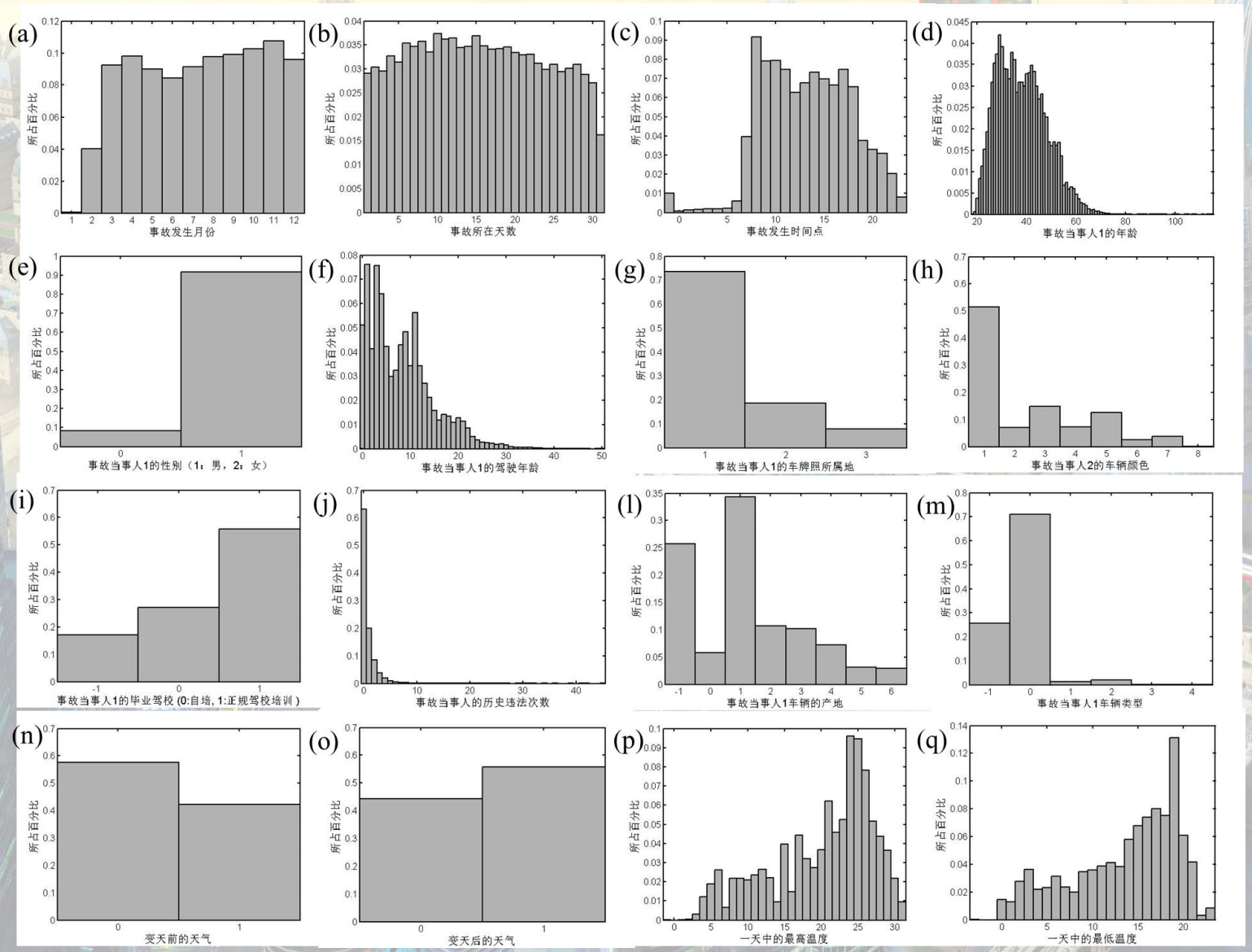
创意描述



描述统计

- 通过描述性统计分析探索交通事故发生频率与时间、事故人、事故车辆状况及天气状况之间的相关性，为分析交通事故成因提供可行的依据
- 根据竞赛方提供的原数据，我们进行了相关的统计分析，右图展示了在不同要素下交通事故（追尾事故）发生的频率

创意描述



描述统计

- 交通事故发生月份、所在天数、时间点(a-c)
- 事故责任人的年龄、性别、驾驶年俩(d-f)
- 事故责任人的车牌所属地、车辆颜色(g-h)
- 事故责任人毕业驾校、违法记录、车辆产地、车辆类型(i-m)
- 事故当天天气变化前状况、天气变化后状况、最高温度、最低温度(n-q)

创意描述

01

Monthly/daily

交通事故发生频率在一年中的不同月份及不同天数没有明显的差异，季节差异对交通影响不大；图1a显示1-2月交通事故发生频率较低，但由于1月份处在春节，导致大量的相关统计数据缺失

02

Diurnal

白天是人类活动主要时间，因此大部分交通事故发生在6~18点。此外，6~8及16~18时间段形成了两个事故高发时间段，在此时间段内，上下班高峰的来临，大大增加了交通事故发生的风险

03

Age/Drive Age

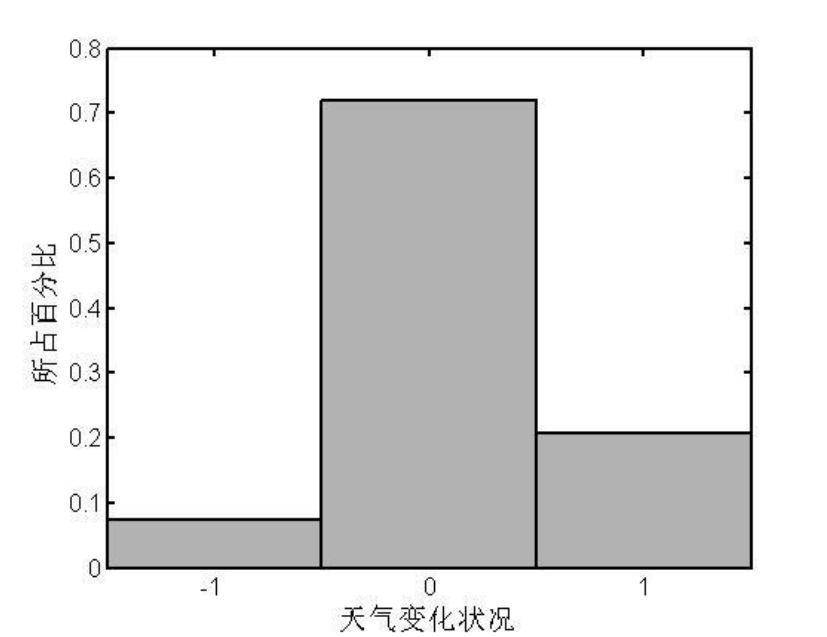
事故发生频率在事故责任人年龄要素上基本呈正态分布，主要为在青、中年阶段男性，这可能是因为青、中年男性是该地区车辆驾驶的主要人群；随着驾龄的增加，交通事故发生频率不断下降

04

Temperature

从图p及q中可以看出，随着当天温度的升高，交通事故发生的频率存在相应的增加，且在最高温度为23~25度的范围内，存在明显的增加；当温度相对较高时交通事故发生的频率呈现下降趋势

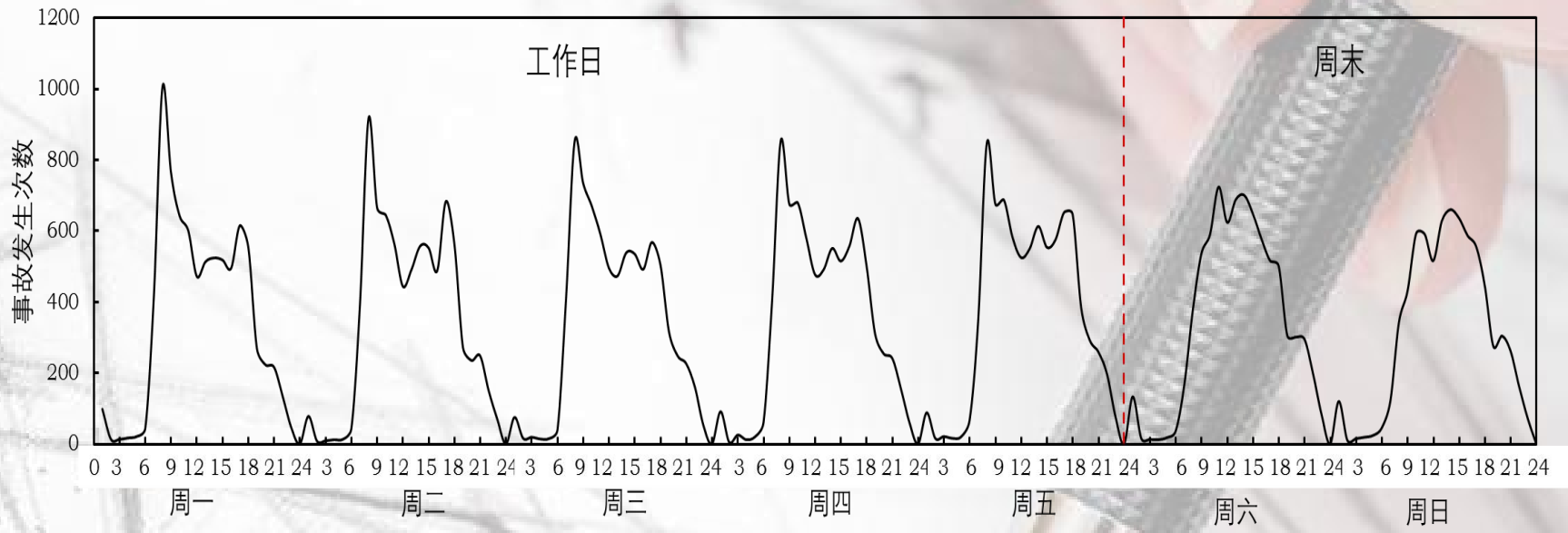
创意描述



事故当天天气变化状况（-1：天气变好了，0：天气未变化，1：天气变差了）

- 在事故当天天气变好的情况下，事故发生的频率要低于天气变差的情况下
- 这说明天气状况转好，可以一定程度上减少交通事故的发生，改善交通质量

创意描述



1周不同时间段交通事故发生次数统计图

- 工作日与周末的交通事故发生情况存在明显的差异
- 工作日，存在明显的早高峰与晚高峰事故发生段，这可能主要是受上下班车流量较大而引起的
- 休息日，事故发生的早高峰与晚高峰随之消失，有效缓解了交通压力
- 大量工作人员驱车前往工作地，会大大增加交通事故的风险

PART FIVE

数据挖掘

数据挖掘

01

因变量

每条道路发生追尾总次数

02

自变量

道路周围各类POI数量

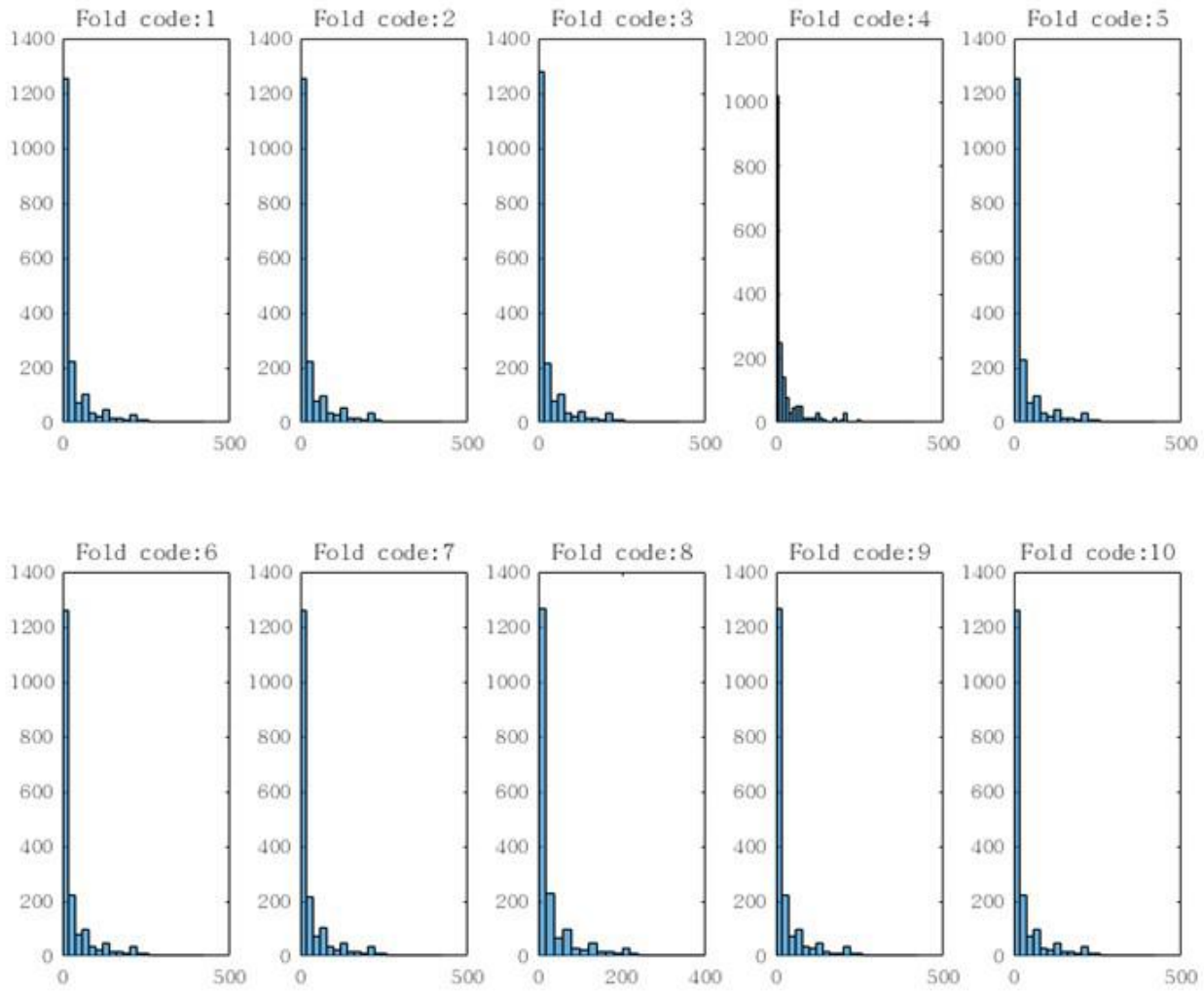
03

采用模型

广义线性模型
决策树回归
随机森林回归

对道路做500米buffer空间分析，统计每条道路周围500米内追尾发生次数，各类POI数量

数据挖掘

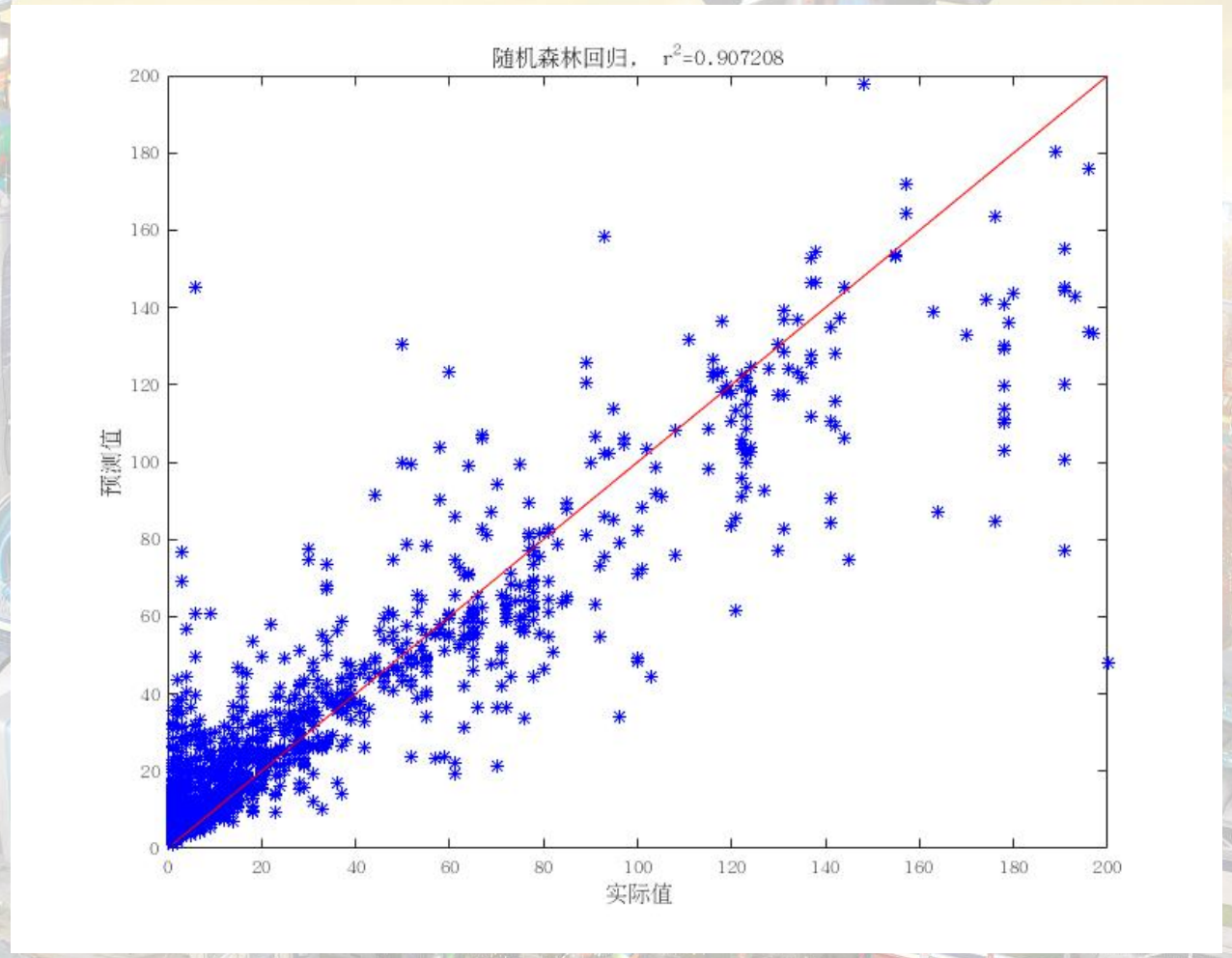


- K-fold交叉验证方法，将原始数据分成K组，每个子集数据分别做一次验证集
- K个模型，作为验证集预测的相关系数平均值作为K-CV模型性能指标，k=10
- K-CV可以有效避免学习及欠学习状态发生，最后得到结果也比较具有说服性

数据挖掘

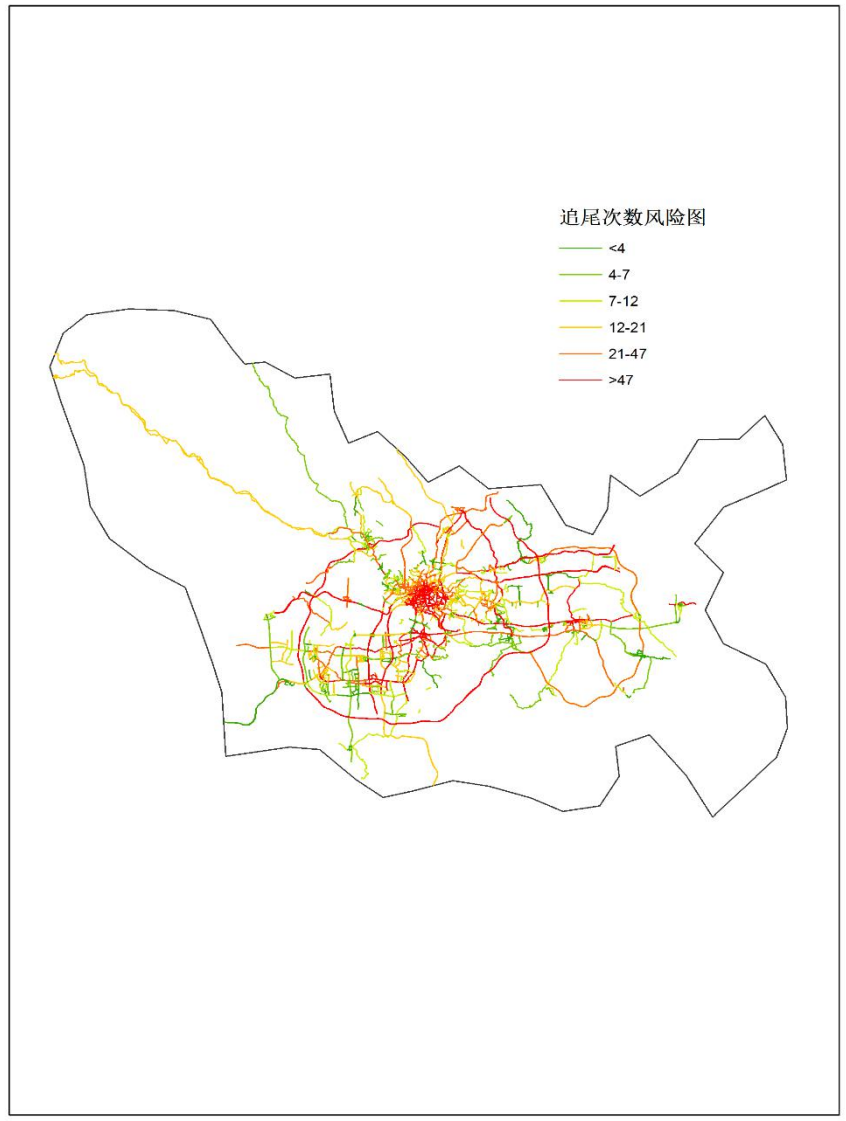
10次交叉验证结果：无论从模型的准确性，还是从模型的稳定性角度评价，均是随机森林回归的效果最佳，R^2在0.65以上，有些能到0.75以上

k	广义线性	决策树	随机森林
1	0.26	0.71	0.71
2	0.21	0.48	0.65
3	0.47	0.57	0.76
4	0.38	0.60	0.84
5	0.37	0.67	0.68
6	0.26	0.45	0.60
7	0.11	0.72	0.76
8	0.41	0.52	0.67
9	0.13	0.62	0.66
10	0.49	0.49	0.71
平均值	0.31	0.58	0.70
标准差	0.14	0.10	0.07



采用全部数据进行随机森林回归，模型 R^2高达0.91

数据挖掘



由于名称完全匹配的事故记录占少数，在建回归模型时仅为追尾次数≥1的道路作为样本，这里给出了用所建模型预测所有道路追尾次数的结果图，可以看作是发生追尾的风险力

PART SIX

未来展望

未来展望

事故匹配

事故记录时需要记录具体位置坐标，这样才能与地理数据匹配，确定事故发生地的道路、环境等状况。现在虽然采取与百度POI匹配定位，但处理中会存在误差

详细记录

目前事故损失相关的记录仅有是否发生事故。对于是否存在人员伤亡及经济损失均未知，若有这样数据，可以进行更详细的分析，并提出针对性的交管建议

道路建设

加强道路基础数据建设

数据使用

城市全信息视角

THANKS