# 高速公路大数据与公路货运统计

中国国家统计局

服务业统计司
2014年10月28日

第一部分 研究背景

第二部分 高速公路大数据介绍及预处理

第三部分 基于大数据的超限车辆规律分析

第四部分 大数据应用启示及前景展望

# 第一部分 研究背景

一、公路运输统计的重要性及存在的困难

二、结合大数据的公路运输统计新方案简介

# 一、公路运输统计的重要性及存在的困难

## (一)重要性

交通运输 → 国民经济运行的大动脉

公路运输 → 交通运输的主要组成部分

公路运输统计 → 对于反映交通运输行业和国民经济形势具有重要意义

## （二）存在的困难

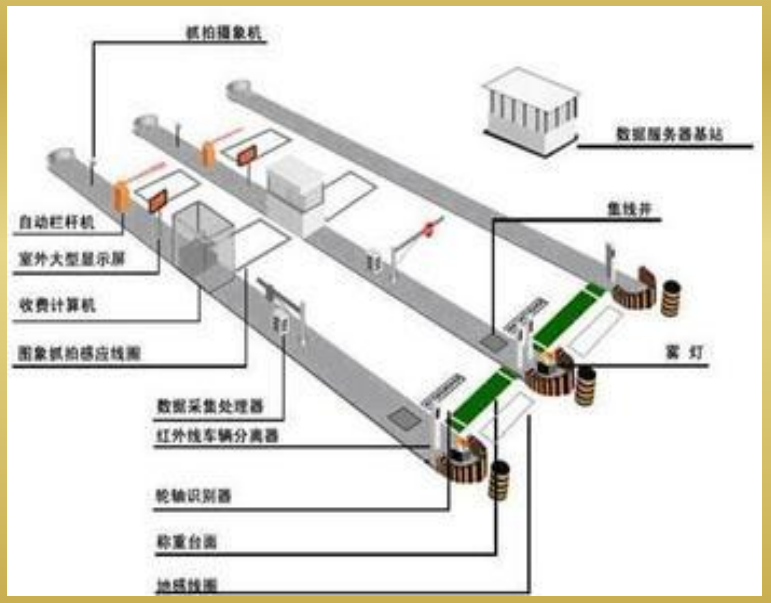
公路运输统计受行业自身特点影响，存在一些困难，主要集中在公路货运统计上：点多面广、流量流向繁杂、市场准入门槛较低、经营业户极不稳定。这使得公路运输统计对公路运输实际情况反映有限，影响了其服务社会的功效。

# 二、结合大数据的公路运输统计新方案简介

服务业统计司与交通运输部综合规划司结合大数据设计了公路运输量统计的新方法：通过传统抽样调查获得月度运量基数，采用大数据思路推算月度波动系数，月度运量基数乘以月度波动系数得到当月运量。

客运波动系数
——客运站售票记录

货运波动系数
——高速公路计重收费记录




新方案仍存在一些问题尤其是货运统计部分，因此我司对高速公路大数据展开了更深入的研究，希望能进一步完善公路货运统计以及探索传统统计与大数据结合的途径。

# 第二部分 高速公路大数据介绍及预处理

一、数据来源及情况

二、原始数据预处理

# 一、数据来源及情况

## （一）数据来源

高速公路联网监控系统的原始记录。

各类检测监控设备

| | |
|---|---|
| 环形线圈检测器 | 微波检测器 |
| 超声波检测器 | 视频检测器 |
| 计重收费系统 | 人工输入 |

对经过车辆实时识别、记录


,0,2013-05-12 08:47:55.0,,2301,2013-05-12 08:47:55.0,苏ASM113,1,0,0,0,,0,0
,0,2013-05-12 08:48:14.0,,2301,2013-05-12 08:48:14.0,苏A9NP86,1,0,0,0,,0,0
,0,2013-05-12 08:48:32.0,,2301,2013-05-12 08:48:32.0,苏AZ6571,1,0,0,0,,0,0
,0,2013-05-12 08:48:50.0,,2301,2013-05-12 08:48:50.0,苏AZ5812,1,0,0,0,,0,0

# （二）数据情况

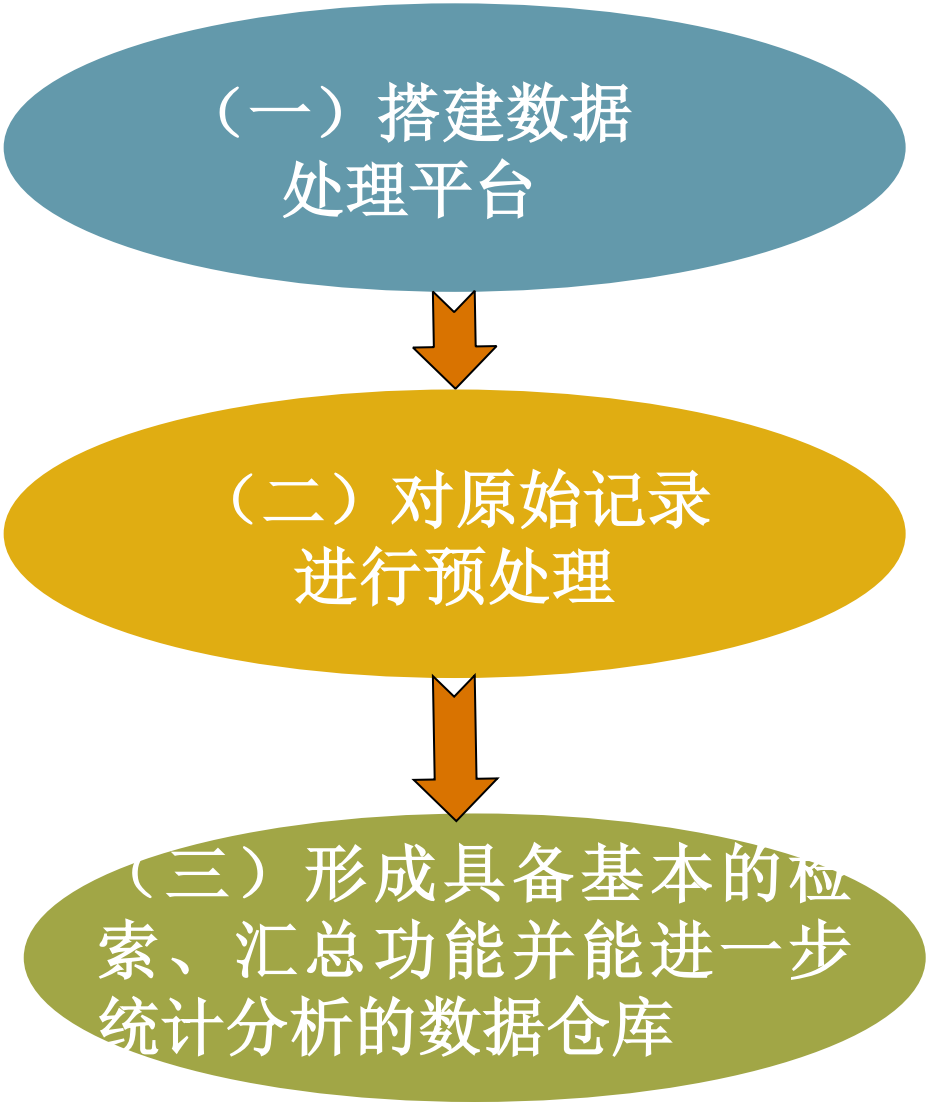
**原始数据情况**

- **原始数据**：2013年1月至2014年4月、14个省的原始数据，共计约50亿条记录。
- **解压前**：原始记录以txt格式存储，均被压缩后存放在共计2210个压缩包中，所有压缩包合计大小约为90GB。
- **解压后**：解压后大小达到1TB，解压就耗费近20个小时。

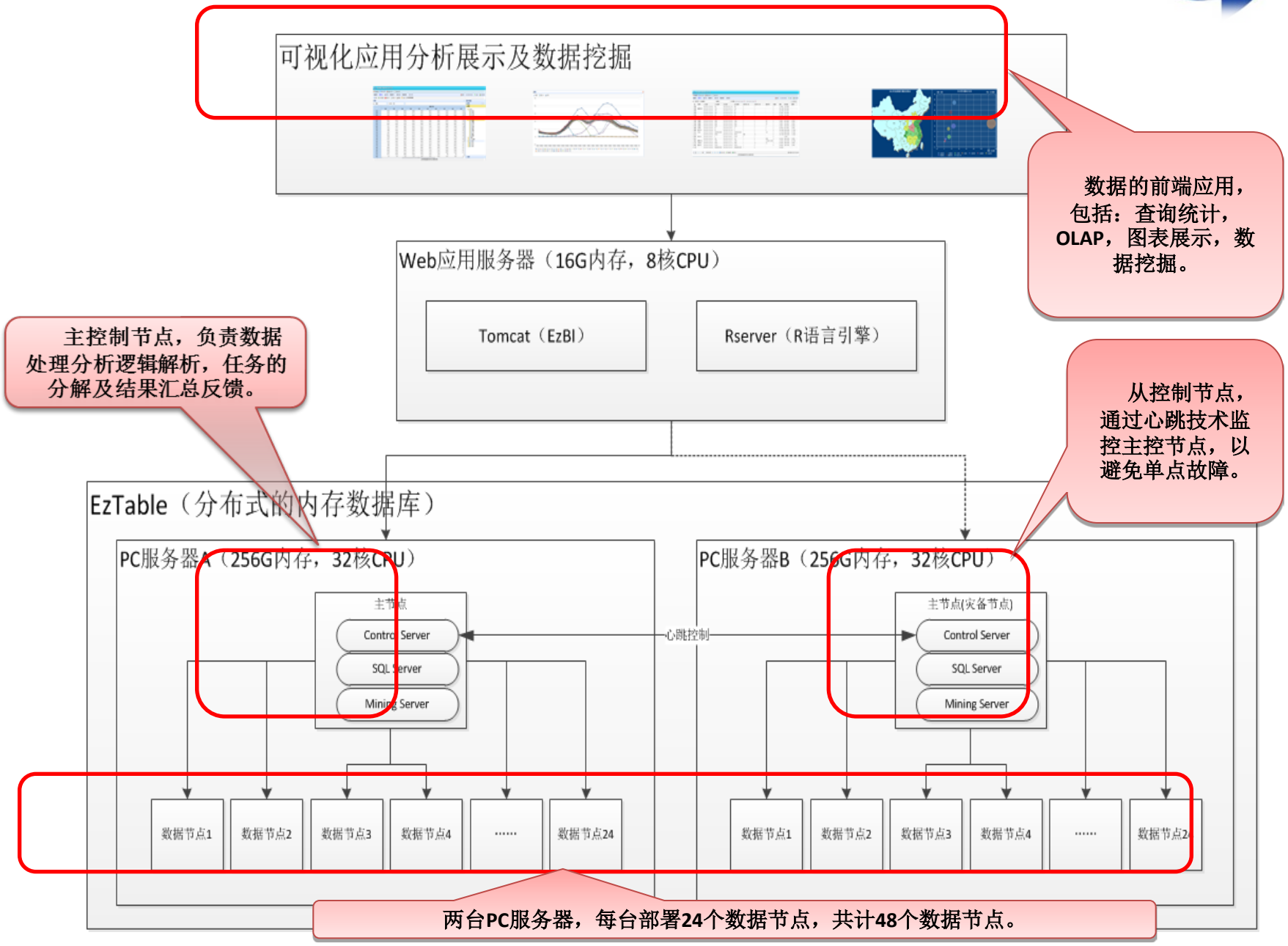
**符合大数据特性**

- Volume数据体量大
- Velocity高速性
- Variety多样性
- Value价值性高

# 二、原始数据预处理



# （一）搭建数据处理平台


可视化应用分析展示及数据挖掘
数据的前端应用，包括：查询统计，OLAP，图表展示，数据挖掘。
Web应用服务器（16G内存，8核CPU）
Tomcat（EzBI）
Rserver（R语言引擎）
主控制节点，负责数据处理分析逻辑解析，任务的分解及结果汇总反馈。
从控制节点，通过心跳技术监控主控节点，以避免单点故障。
EzTable（分布式的内存数据库）
PC服务器A（256G内存，32核CPU）
主节点
Control Server
SQL Server
Mining Server
心跳控制
PC服务器B（256G内存，32核CPU）
主节点(灾备节点)
Control Server
SQL Server
Mining Server
数据节点1
数据节点2
数据节点3
数据节点4
.....
数据节点24
两台PC服务器，每台部署24个数据节点，共计48个数据节点。

# （二）对原始记录进行预处理

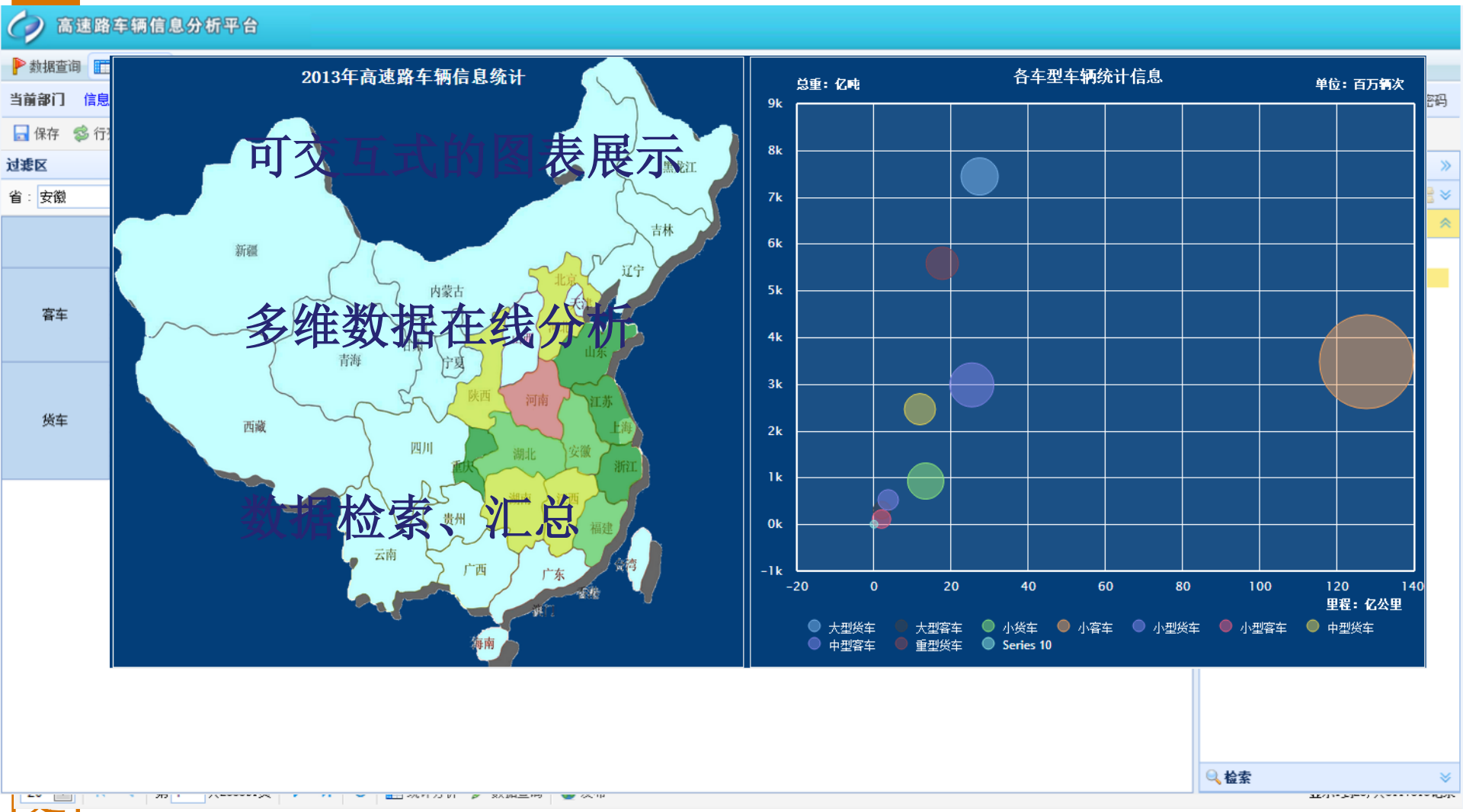
各省之间的数据格式、字段含义等存在一定差异。

通过预处理，去除或者调整了其中的一些无效数据、异常数据，补充了部分可以估计的不完整数据，对格式、字段含义、代码进行了转化统一，形成了可以进行初步检索、汇总、分析的数据仓库。

以车型代码为例：

| 省 | 样例 | 说明 |
|---|---|---|
| 安徽 | 1-4，11-15 | 标准代码 |
| 福建 | 1-4，1-5 | 根据车种代码处理，1-5转换为11-15 |
| 广东 | | 非标准代码，根据推测转换 |
| 河北 | 1-4，1-9 | 需要转换，1-5转换为11-15，6-9转换为15 |
| 河北省京沈片区 | 1-4，11-15 | 标准代码 |
| 河南 | | 非标准代码，根据推测转换 |
| 湖北 | | 非标准代码，根据推测转换 |
| 湖南 | 1-4，1-9 | 1-4转换11-14，5-9转换15 |
| 江苏 | 1-4，11-15 | 标准代码 |
| 江西 | 1-4，11-22 | 21-22转换15 |
| 山东 | 1-4，11-15 | 标准代码 |
| 陕西 | 1-4，1-5 | 根据车种代码处理，1-5转换为11-15 |
| 上海 | 1-4，5-11 | 5-9转换11-15，10-11转换为15 |
| 浙江 | 1-4，1-7 | 1-5转化11-15，6-7转换15 |
| 重庆 | 1-4，1-5 | 根据车种代码处理，1-5转换为11-15 |

# （三）形成具备基本的检索、汇总功能并能进一步统计分析的数据仓库



# 第三部分 基于大数据的超限车辆规律分析

一、目的

二、思路和方法

三、具体实施过程

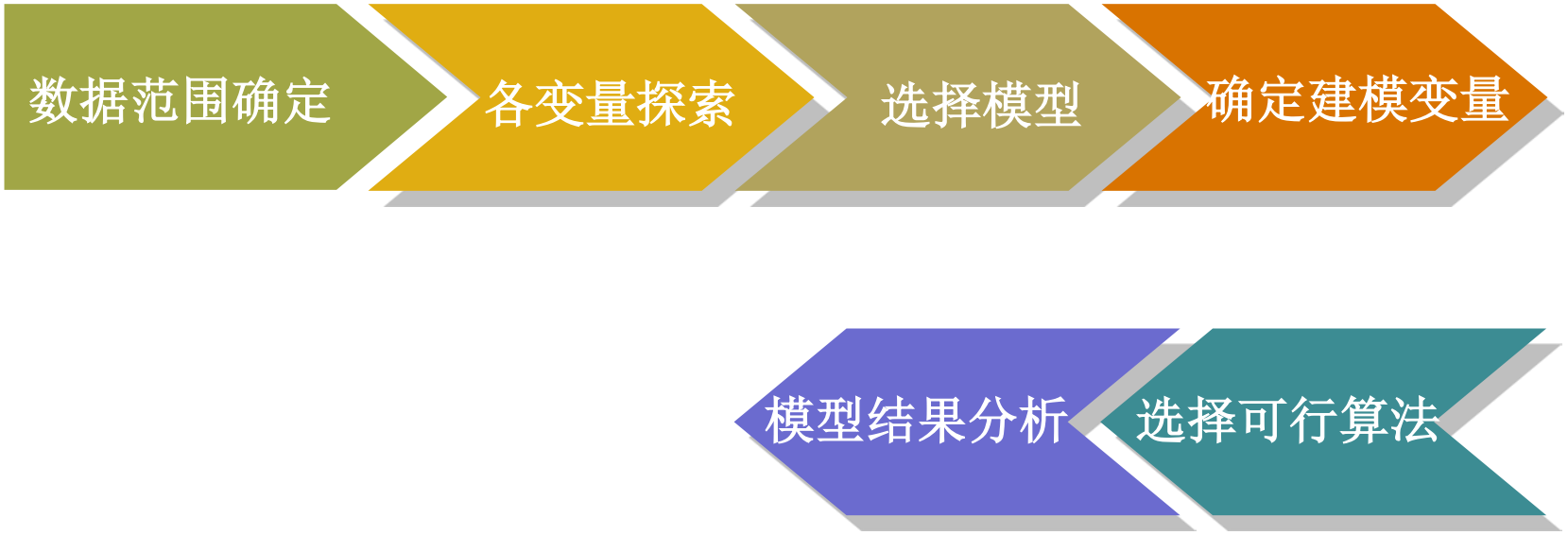
四、初步结论

# 一、目的

1. 熟悉大数据的挖掘分析方法
2. 进一步摸清各变量特点
3. 为完善公路货运统计打下基础
4. 积累统计中应用大数据的经验

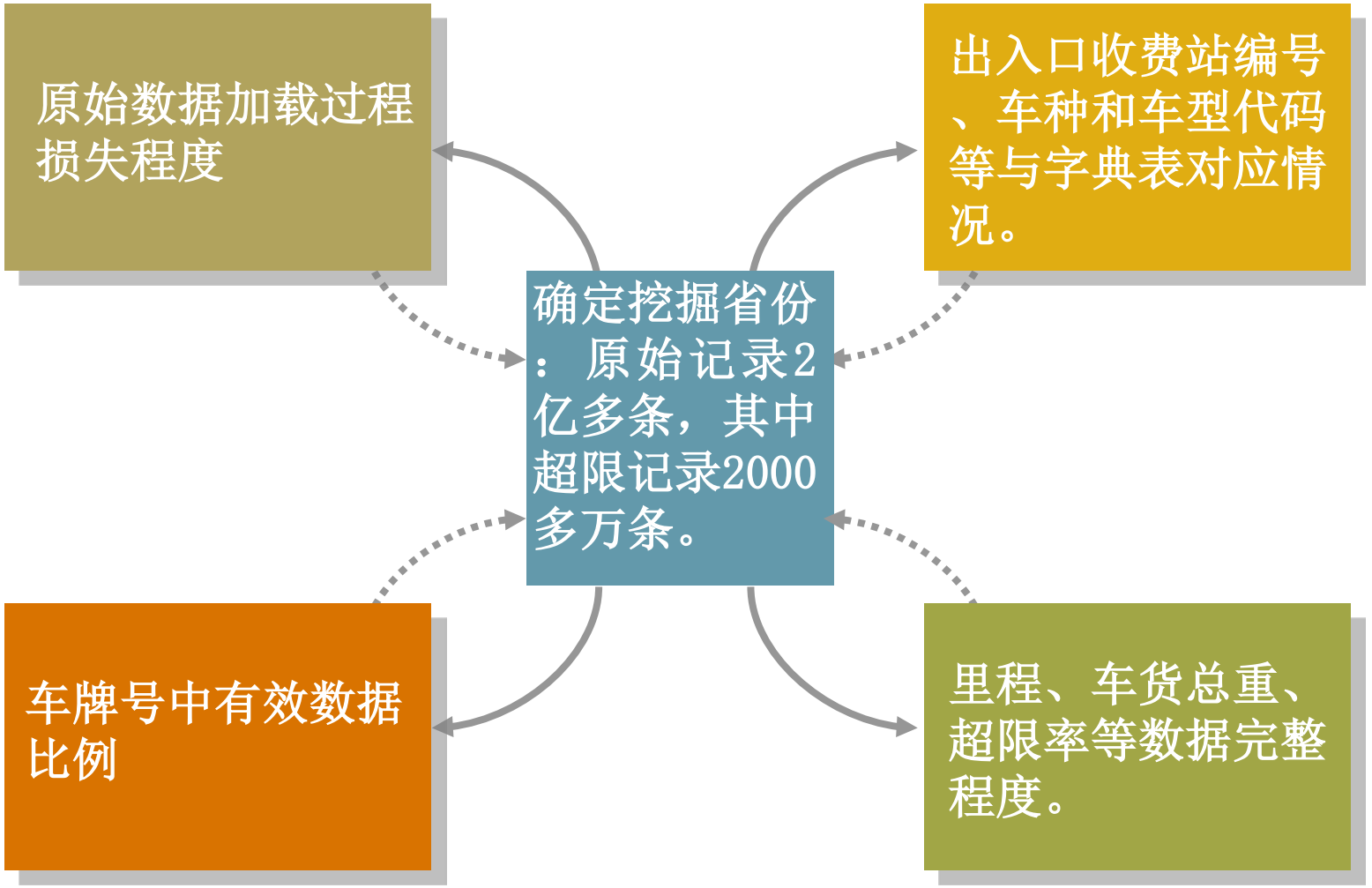
# 二、思路和方法

思路：从大数据本身出发，对已有各变量自身特征进行初步挖掘，基于发现的规律建立合理的模型，选择可行算法展开深层挖掘，分析模型挖掘结果，找出超限车规律。
根据思路确定方法步骤如下：



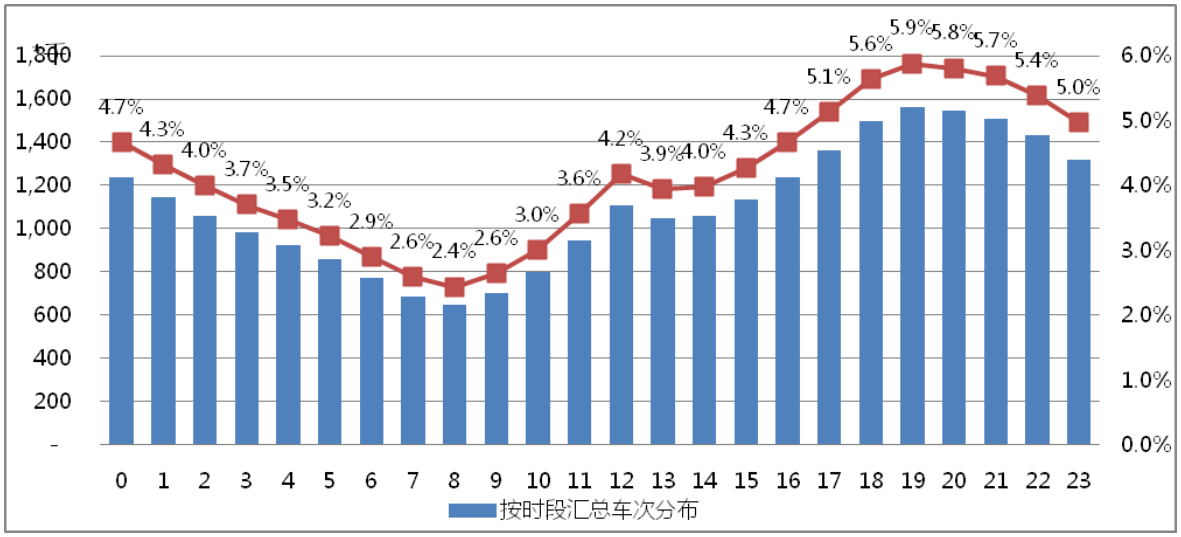
# 三、具体实施过程

## (一)数据范围确定：选择数据质量较好的某省初探挖掘

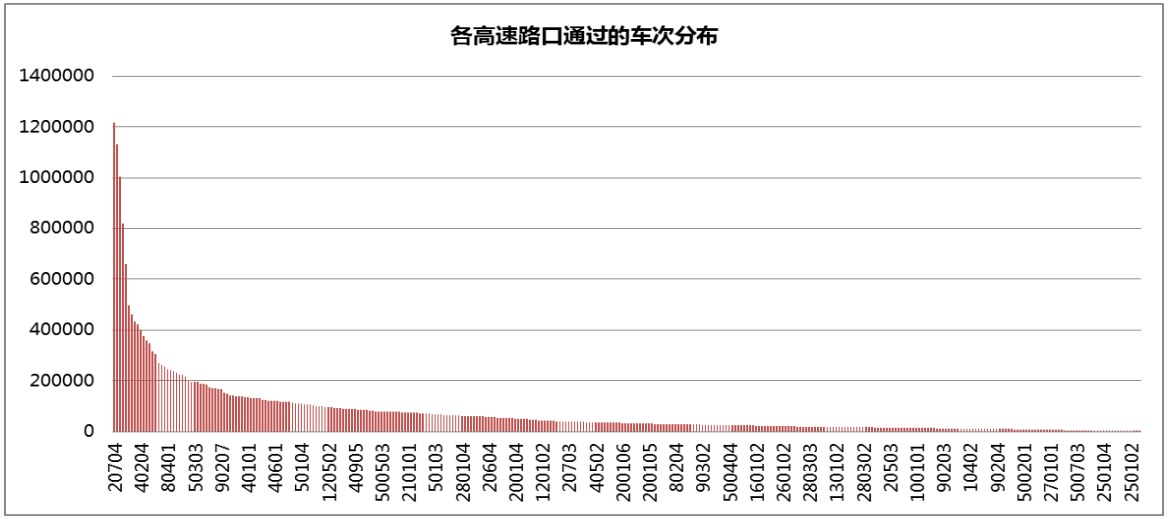


# （二）各变量探索

## 小时



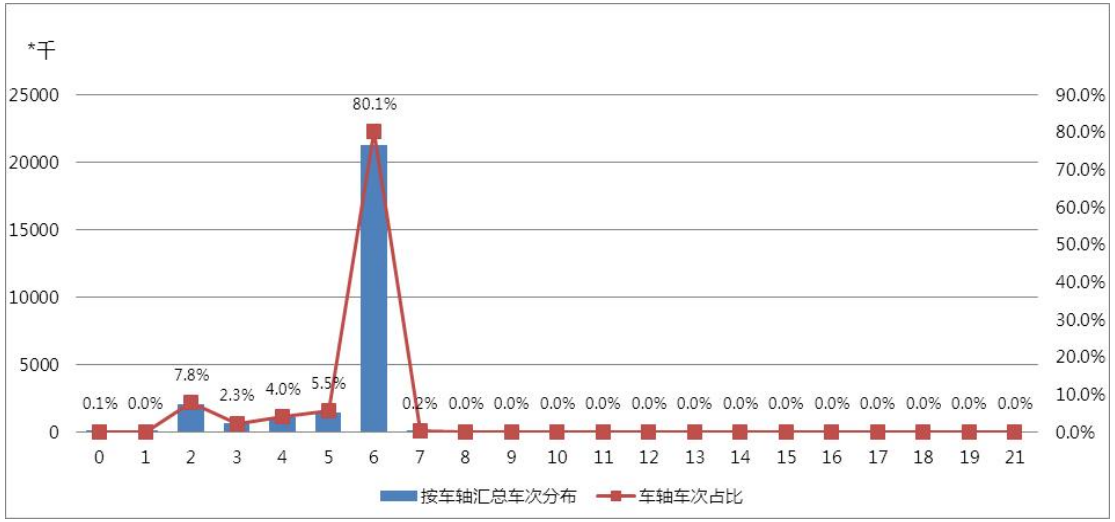

## 入口站点



第三部分

基于大数据的超限车辆规律分析

## 车轴




## 里程、车货总重、车次

| 分位数 | 里程 | 车货总重 | 车次 |
|---|---|---|---|
| 0% | 0 | 0 | 0 |
| 5% | 15 | 18,100 | 1 |
| 10% | 25 | 22,400 | 1 |
| 15% | 39 | 29,600 | 1 |
| 20% | 54 | 36,600 | 1 |
| 25% | 75 | 39,200 | 1 |
| 30% | 100 | 45,200 | 1 |
| 35% | 130 | 52,500 | 2 |
| 40% | 166 | 57,000 | 2 |
| 45% | 210 | 73,600 | 2 |
| 50% | 263 | 87,300 | 3 |
| 55% | 324 | 105,300 | 3 |
| 60% | 395 | 127,100 | 4 |
| 65% | 491 | 161,500 | 5 |
| 70% | 640 | 211,100 | 7 |
| 75% | 858 | 290,400 | 9 |
| 80% | 1,239 | 432,700 | 12 |
| 85% | 2,003 | 729,100 | 19 |
| 90% | 4,015 | 1,562,600 | 36 |
| 95% | 11,705 | 5,030,000 | 96 |
| 100% | 1,380,436 | 331,668,700 | 6,336 |

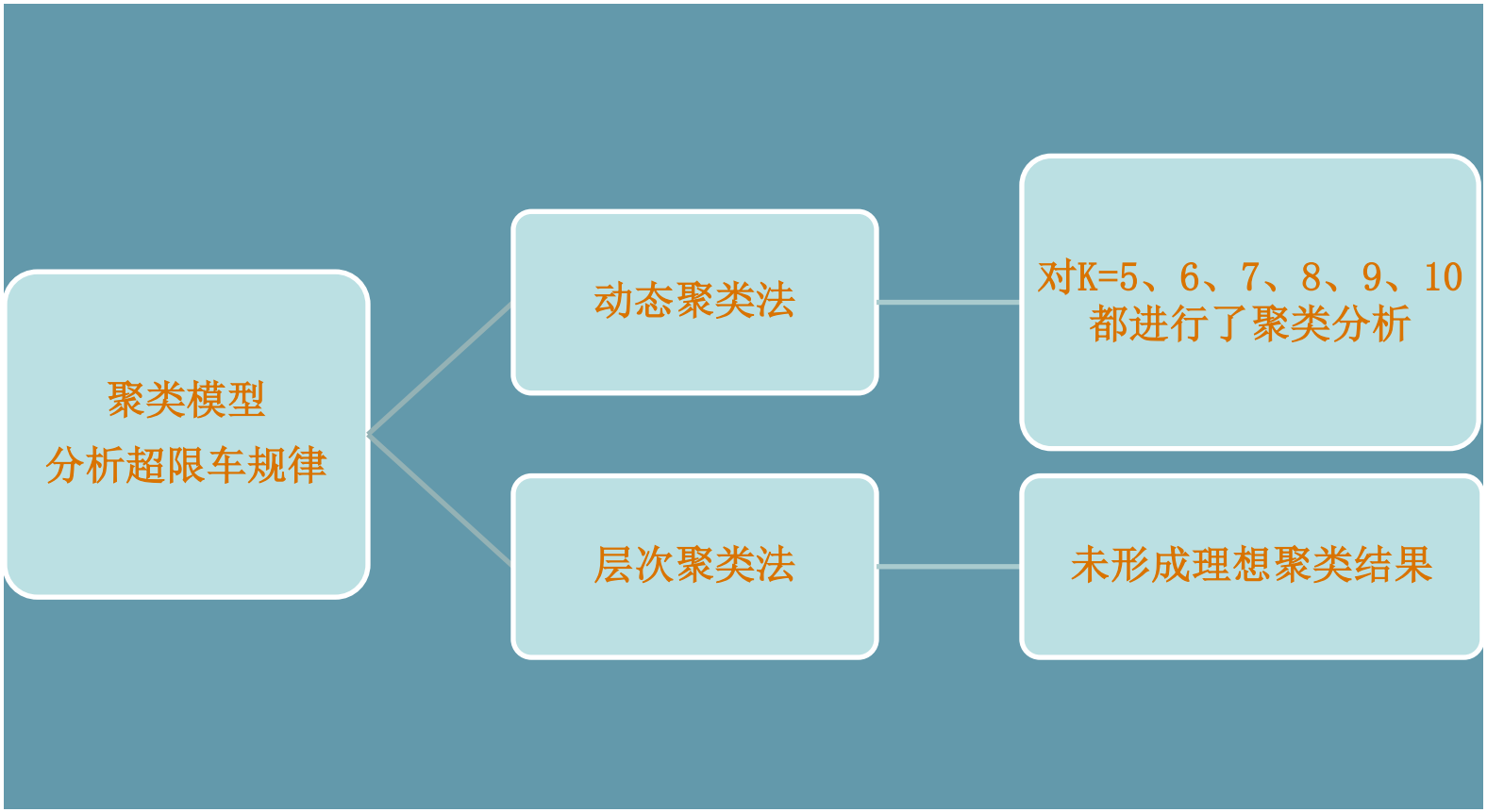
# （三）选择模型和确定建模变量

根据对小时、里程、车轴、车货总重等变量的初步探索，结合考虑需处理的数据量大小，比较多种模型的可行性和结果的有效性后，最终选择聚类分析模型。

按变量对于超限特征分析的解释能力和对于超限分析的重要程度进行筛选，确定放入到模型的变量为:月份、小时、里程、车货总重和车次。

| 变量 | 变量字段名 | 变量中文解释 | 是否放入模型 |
|---|---|---|---|
| V1 | province | 省份，目前只有山东 | 不放 |
| V2 | year | 进入高速路的年，目前是13年和14年 | 不放 |
| V3 | ENTRY_mon | 进入高速路时间的月份 | |
| V4 | ENTRY_hour | 进入高速路时间的小时，0-23 | |
| V5 | entry_station | 进入高速路的入口收费站编号 | 不放 |
| V6 | VEHICLE_CLASS_S | 车种 | 不放 |
| V7 | VEHICLE_TYPE_S | 车型 | 不放 |
| V8 | axis_num | 车轴数 | 不放 |
| V9 | sum(DISTANCE) | 里程 | |
| V10 | sum(TOTAL_WEIGHT) | 车货总重 | |
| V11 | count | 车次 | |

# （四）模型具体算法



# （五）模型结果分析

经过比较，k等于8的聚类效果优于k=5、6、7、9、10时的效果，故选择k=8的聚类结果：

| 变量列表 | 群1 | 群2 | 群3 | 群4 | 群5 | 群6 | 群7 | 群8 |
|---|---|---|---|---|---|---|---|---|
| 进入高速路的月份 | 6 | 7 | 2 | 3 | 10 | 10 | 3 | 6 |
| 进入高速路的小时 | 13 | 12 | 12 | 4 | 5 | 18 | 20 | 13 |
| 里程 | 586543 | 1616 | 1646 | 1489 | 1542 | 2044 | 2183 | 99034 |
| 车货总重 | 138981 | 618 | 578 | 704 | 728 | 711 | 738 | 38653 |
| 车次 | 2632 | 13 | 12 | 14 | 15 | 15 | 16 | 685 |
| 颗粒数量 | 1483 | 119328 | 170986 | 179667 | 157086 | 182685 | 205430 | 11625 |
| 颗粒数据占比 | 0.14% | 11.60% | 16.63% | 17.47% | 15.28% | 17.77% | 19.98% | 1.13% |

| | 群1 | 群2 | 群3 | 群4 | 群5 | 群6 | 群7 | 群8 |
|---|---|---|---|---|---|---|---|---|
| 总里程 | 869842656 | 192803531 | 281469518 | 267586532 | 242202921 | 373384933 | 448405833 | 1151266010 |
| 占总里程百分比 | 22.73% | 5.04% | 7.35% | 6.99% | 6.33% | 9.76% | 11.72% | 30.08% |
| 每辆车的平均里程 | 222 | 121 | 132 | 103 | 103 | 133 | 138 | 144 |
| 车货重量 | 206109379067 | 73690664829 | 98857784998 | 126413106809 | 114330606021 | 129899985706 | 151697784241 | 449339267947 |
| 占车货总重百分比 | 15.26% | 5.46% | 7.32% | 9.36% | 8.47% | 9.62% | 11.23% | 33.28% |
| 车次 | 3902859 | 1587032 | 2125577 | 2590700 | 2350762 | 2800574 | 3246403 | 7964479 |
| 占车次总数百分比 | 14.69% | 5.97% | 8.00% | 9.75% | 8.85% | 10.54% | 12.22% | 29.98% |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 群1 | 收费站编号 | 120106 | 20101 | 20704 | 50305 | 20801 | 50304 | 60108 | 230207 | 50205 | 240104 |
| | 颗粒数量 | 334 | 324 | 306 | 123 | 105 | 64 | 62 | 54 | 48 | 30 |
| | 位置 | 省界 | 省界 | 省界 | 省界 | 省界 | 省界 | 省界 | 省界 | 省内 | 省界 |
| 群8 | 收费站编号 | 20505 | 40204 | 120201 | 50304 | 230207 | 240104 | 80109 | 60108 | 503010 | 50305 |
| | 颗粒数量 | 364 | 348 | 327 | 310 | 306 | 303 | 297 | 284 | 260 | 259 |
| | 位置 | 省界 | 省界 | 省内 | 省界 | 省界 | 省界 | 港口 | 省界 | 港口 | 省界 |

# 四、初步结论

超限车规律

空间规律：

大部分超限车从11个省界站点、2个港口站点进入高速公路（全省共300多个站点）；剩下的小部分超限车辆多从2个省内站点进入高速公路。

时间规律：

超限车辆较多的月份是2、3、4月；较多的小时是夜里的7-9点和0点。

抽样分层基础：
各自具有明显特征的8个群，可作为在运量基数中考虑超限的基础。

# 第四部分 大数据应用启示及前景展望

一、大数据应用启示

二、前景展望

# 一、大数据应用启示

大数据应用需要在大量摸索尝试中，确定研究分析方向，并不断调整改进。

大数据挖掘中，简单算法时常比复杂算法更实用。

获得大数据及对大数据预处理形成可以进行分析操作的数据仓库至关重要。

# 二、前景展望

继续挖掘超限车规律，在统计运量基数时充分考虑超限造成的数据差距,设法加以完善。

深入研究高速公路运量变化与公路运量变化的关系,寻找完善波动系数的方法。

增长时间序列，研究高速公路运量与宏观经济的深层关系。

# 谢谢！