

交通规划行业中的数据应用现状及思考

综合交通所

邹海翔

2019 年 2 月





大纲

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

再认识与展望

① 数据分析的“武器库”

② 再认识与展望



目录

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库” 3

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

① 数据分析的“武器库”

② 再认识与展望

数据分析的“武
器库”⁴

- 数据采集
- 数据预处理
- 业务分析
- 数据建模
- 数据库
- 分析算法
- 可视化
- 再认识与展望



谢逊取过手边的屠龙宝刀，拔刀出鞘，嚓的一声，在大树的树干上斜砍一刀，只听得砰的一响，大树的上半段向外跌落。

工欲善其事，必先利其器
—《论语·卫灵公》



数据分析的“七种武器”

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”
5

数据采集
数据预处理
业务分析
数据建模
数据库
分析算法
可视化

再认识与展望

- ① 数据采集
- ② 数据处理
- ③ 业务分析
- ④ 数据建模
- ⑤ 数据库
- ⑥ 分析算法
- ⑦ 可视化



数据采集

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

6

数据采集
数据预处理
业务分析
数据建模
数据库
分析算法
可视化
再认识与展望

- 利用服务器端提供的 API 接口调取数据
- 编写爬虫程序，从网页上抓取数据

13



数据采集

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

6

- 利用服务器端提供的 API 接口调取数据
- 编写爬虫程序，从网页上抓取数据

字段名称	字段含义	说明
alter	交通出行方式	推荐公交、巴士优先、骑车、小汽车、出租车等
O_lon & O_lat	出发地坐标	百度公司加密后的经纬度坐标
D_lon & D_lat	到达地坐标	百度公司加密后的经纬度坐标
distance	出行距离	包括步行在内的实际路网距离
duration	出行时间	包括步行在内
price	票价	
walkDistance	步行距离	
walkTime	步行时间	
transferNumber	换乘次数	

表：百度地图 API 返回的公交路径规划数据结果

13



数据采集

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

13

- 利用服务器端提供的 API 接口调取数据
- 编写爬虫程序，从网页上抓取数据



图：利用爬虫程序，可以从丰富的互联网资源中自动化采集数据



数据预处理

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

7

- 现实中的原始数据都是不完整、不一致的脏数据，需要编写程序对数据进行**清洗、集成、变换和归约**等自动化处理
- 示例：GPS 数据预处理

13



数据预处理

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

7

- 现实中的原始数据都是不完整、不一致的脏数据，需要编写程序对数据进行**清洗、集成、变换和归约**等自动化处理
- 示例：GPS 数据预处理

算法：GPS 原始数据清洗

```
Data: 原始车牌 GPS 文件:GPS_File, 每行 line=(ID,Postion,Time,Direction,Speed,State)
Result: 清洗后的 GPS 文件:GPS_File_P
输入: 时间容差 Δt, 距离容差 Δd, 原始 GPS 数据列表 G
      /* 对 G 按照时间排序
1 SortByTime(G);
2 bFirst ← true;
3 for i = 1 ← 1 to N do
4   /* 删除位置不在市域范围 Rect 的错误点
5   if G[i].position ∉ Rect then continue;
6   if bFirst = true then
7     GPS_File_P ← OutputResult(G[i]);
8     bFirst ← true; tt ← G[i].time; tp ← G[i].position;
9   else
10    /* 根据前后点时间容差和距离容差删除错误点
11    if (5 ≤ (G[i].time - tt) ≤ Δt) ∧ Distance(G[i].position, tp) ≤ Δd then
12      GPS_File_P ← OutputResult(G[i]);
13      tt ← G[i].time; tp ← G[i].position;
14    end
      end
      /* 输出结果 */
      /* 输出结果 */
      /* */
```

数据预处理

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

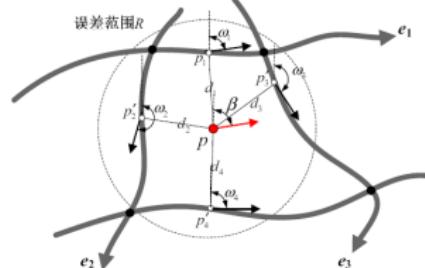
数据库

分析算法

可视化

再认识与展望

7



算法: GPS 与道路地图匹配

输入: 路网模型 G 和 GPS 轨迹 $T: p_1 \rightarrow p_2 \rightarrow \dots \rightarrow p_n$

输出: 处理后的 GPS 点序列 $P: c_1^{j1} \rightarrow c_2^{j2} \rightarrow \dots \rightarrow c_n^{jn}$

```

/* 清空列表 tList
1  tList ← empty;
2  for i = 1 to n do
3      /* 错误范围 R 内的候选路段集
4      s = GetCandidates(p_i, G, R);
5      tList.add(s);
6  end
7  G'_T = ConstructGraph(tList);           /* 构建图 G'_T */
8  return FindMatchedSequence(G'_T);
    */

```



业务分析

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集
数据预处理
业务分析
数据建模
数据库
分析算法
可视化
再认识与展望



图：轨道刷卡数据分析的业务流程



数据建模

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

9

13

- 根据业务需求对数据进行抽象，形成计算机能够理解的逻辑关系和物理结构
- 示例：车辆轨迹模型



数据建模

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

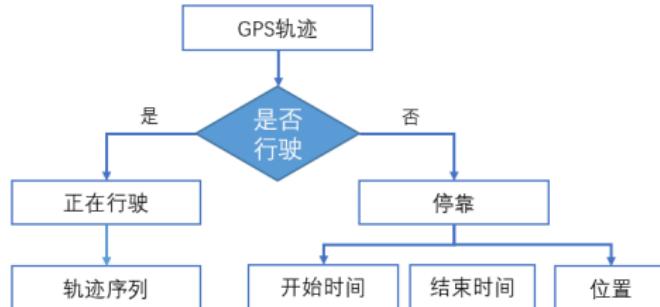
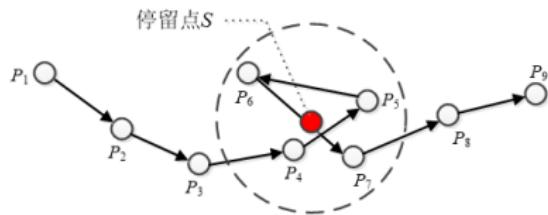
可视化

再认识与展望

9

- 根据业务需求对数据进行抽象，形成计算机能够理解的逻辑关系和物理结构
- 示例：车辆轨迹模型

	经度	纬度	时间
P_1 :	Lat_1	Lon_1	T_1
P_2 :	Lat_2	Lon_2	T_2
P_n :	Lat_n	Lon_n	T_n



图：车辆轨迹模型



数据建模

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

- 根据业务需求对数据进行抽象，形成计算机能够理解的逻辑关系和物理结构
- 示例：车辆轨迹模型

算法：停留点识别

输入：GPS 点序列 P , 距离阈值 $distThreh$, 时间阈值 $timeThreh$

输出：停留点集合 $SP = \{S\}$

```
1 i = 0; pointNum = |P|;
2 while i < pointNum do
3     j ← i + 1; Label = 0;
4     while j < pointNum do
5         dist = Distance( $p_i, p_j$ ); /* 计算两点距离 */
6         if dist > distThreh then
7             ΔT ←  $p_j.T - p_i.T$ ; /* 计算两点时间差 */
8             if ΔT > timeThreh then
9                 S.coord ← ComputeMeanCoord( $p_k | i \leq k \leq j$ ); /* 计算静止点集的中心坐标 */
10                S.avrT ←  $p_i.T$ ; S.levT ←  $p_j.T$ ;
11                SP.insert(S);
12                i ← j; Label ← 1;
13            end
14            break;
15        end
16        j = j + 1;
17    end
18    if Label ≠ 1 then i = i + 1;
19 end
20 return SP;
```



数据建模

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

9



图：车辆轨迹地图

13



数据库

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

10



图：数据库技术的核心是对数据进行高效组织存储，以及多终端并发查询访问

13



数据库

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

10

- 通过高效的组织和存储，实现数据“增删改查”操作
- 根据数据量级、数据内容、应用场景选择最适合的数据库技术
- 数据库是大数据最核心的技术之一，目前主流采用分布式存储方案来解决



图：根据实际需求选择最合适的数据库

13

数据库

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

10

- 通过高效的组织和存储，实现数据“增删改查”操作
- 根据数据量级、数据内容、应用场景选择最适合的数据库技术
- 数据库是大数据最核心的技术之一，目前主流采用**分布式存储方案**来解决



图：目前主流的大数据数据库产品



分析算法

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

11

- 算法不是简单的统计，而要能挖掘数据的规律和关系
- 交通规划中最常用的分析算法是空间分析算法
- 示例：核密度分析

13



分析算法

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

11

- 算法不是简单的统计，而要能挖掘数据的规律和关系
- 交通规划中最常用的分析算法是**空间分析算法**
- 示例：核密度分析

13

分析算法

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

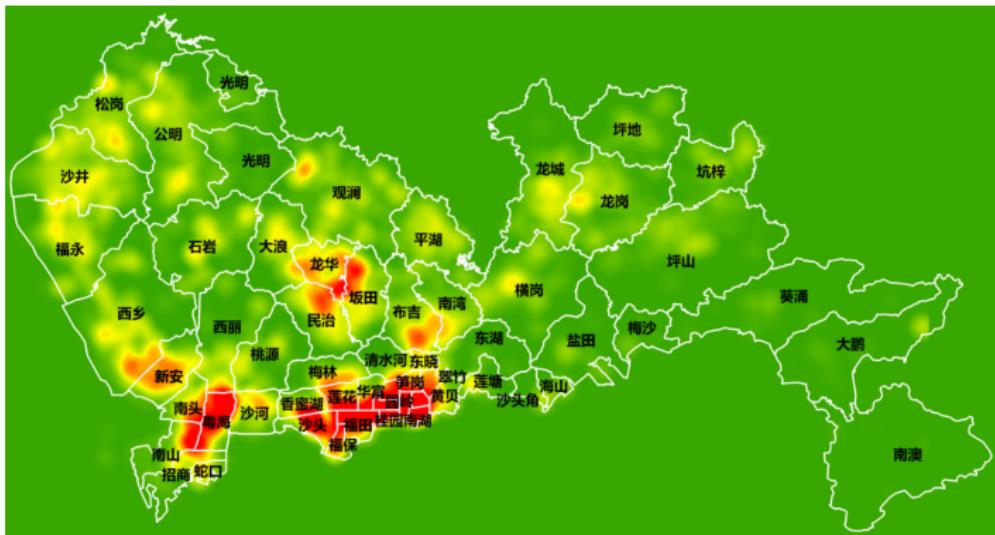
分析算法

可视化

再认识与展望

11

- 算法不是简单的统计，而要能挖掘数据的规律和关系
- 交通规划中最常用的分析算法是**空间分析算法**
- 示例：核密度分析



图：核密度分析

13

分析算法

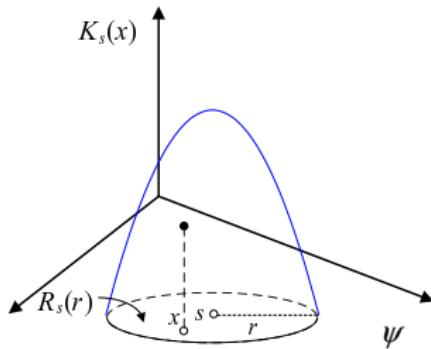
交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集
数据预处理
业务分析
数据建模
数据库
分析算法
可视化

再认识与展望

11



带权重的核密度估计^{1,2}

$$\hat{\lambda}(s) = \frac{1}{n\pi r^2} \sum_{i=1}^n w_i K_s \left(\frac{d_{sx}}{r} \right)$$

其中, $d_{sx} = \|s - x_i\|$ 表示位置 s 和第 i 个样本 x_i 之间的欧式距离; r 被称为带宽, 表示区域 $R_s(r)$ 的面积大小。

- KDE 的核心思想就是通过计算单位面积内样本的核密度平滑处理来估计位置的属性值
- KDE 结果表示区域 $R_s(r)$ 内不同样本对估计 s 位置属性值的权重
- 核函数是一个与距离有关的函数, 而且其赋予区域 $R_s(r)$ 内每个样本点的值是不同的, 离中心位置 s 越远的样本点, 其值应该越小, 而离中心位置 s 越近的样本点, 其值应该越大
- 对 KDE 结果产生影响的两个因素包括核函数的形式以及带宽的选择

¹ Silverman B W. 1986. Density Estimation for Statistics and Data Analysis[M]. London: Chapman Hall.

² Wand M P, Jones M C. 1995. Kernel Smoothing[M]. Boca Raton, Florida: Chapman & Hall/CRC.

分析算法

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

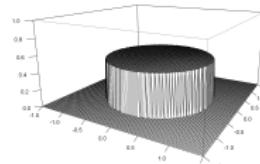
分析算法

可视化

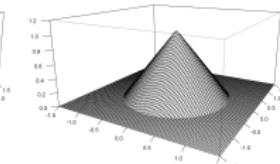
再认识与展望

11

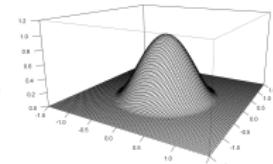
- 算法不是简单的统计，而要能挖掘数据的规律和关系
- 交通规划中最常用的分析算法是**空间分析算法**
- 示例：核密度分析



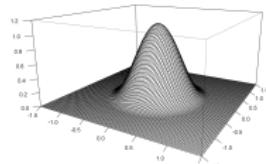
(a) Uniform



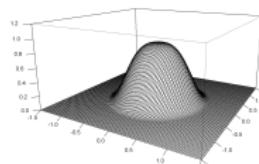
(b) Triangular



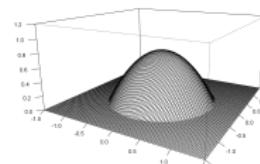
(c) Quartic



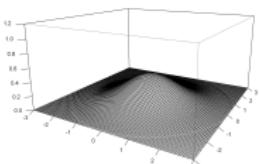
(d) Triweight



(e) Tricube



(f) Epanechnikov



(g) Gaussian

图：常见的核函数

13

可视化

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

分析算法

可视化

再认识与展望

12



图：Edward Tufte(1942-)，美国统计学家，数据可视化理论的先驱者和领军人物，人称“数据达芬奇”

13

- 计算机技术与艺术的结合
- 将绘图与数据分离，数据相关绘图与数据无关绘图分离
- 示例：交通流量可视化

The commonality between science and art is in trying to see profoundly - to develop strategies of seeing and showing.

—Edward Tufte



可视化

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

数据库

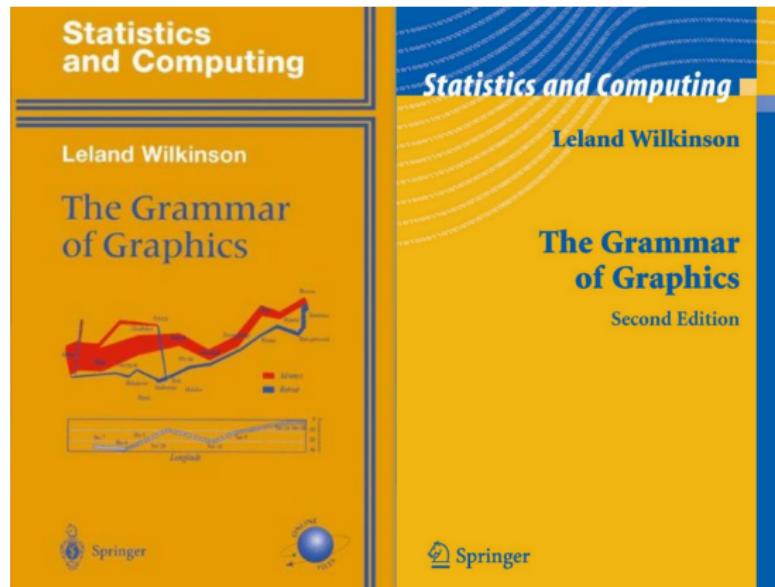
分析算法

可视化

再认识与展望

12

- 计算机技术与艺术的结合
- 将绘图与数据分离，数据相关绘图与数据无关绘图分离
- 示例：交通流量可视化



图：可视化领域经典著作《The Grammar of Graphics》的第一版（1999）和第二版（2005）。

13

可视化

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

数据采集

数据预处理

业务分析

数据建模

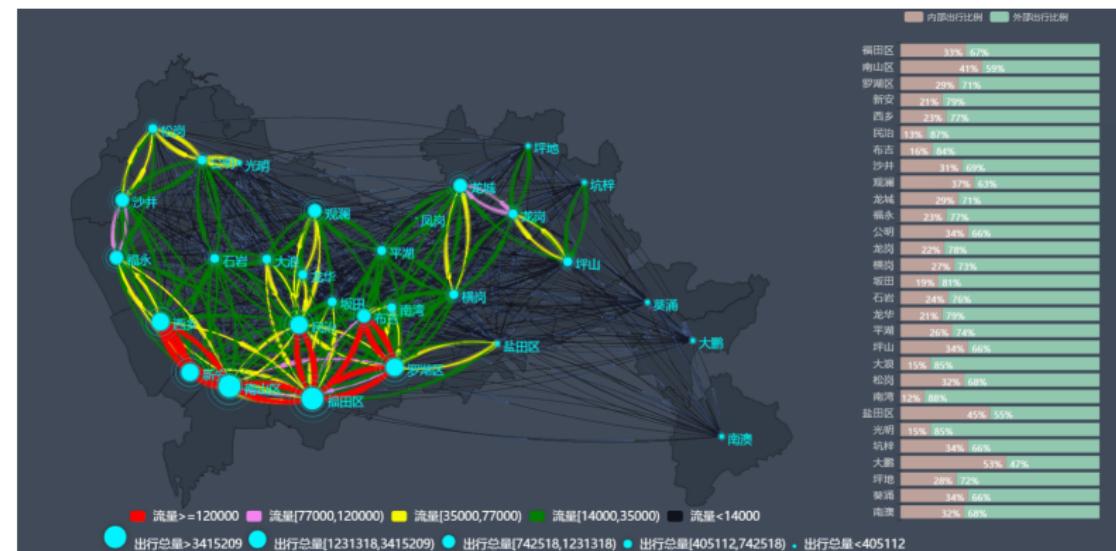
数据库

分析算法

可视化

再认识与展望

12



图：在一张图中同时展现总流量、OD量和内外部出行比例，数据在外部单独存储，与样式无关

13



目录

交通规划行业中
的数据应用现状
及思考

数据分析的“武
器库”

再认识与展望

13

① 数据分析的“武器库”

② 再认识与展望

13

汇报结束
谢谢!

