2019 年本科生课程《交通信息融合与挖掘》期末大作业

（布置时间：2019 年 11 月 27 日；布置教师：唐克双、沈煜、王玲）

(一) 作业内容和要求

某城市干道包含 5 个信号控制交叉口和 6 个路段，限速为 60km/h，各交叉口交通流量和信号配时方案已知。该干道上每个路段安装有线圈检测器，干道两端安装有车辆自动识别（AVI）设备，干道车流中有 10%左右的出租车浮动车。现有线圈、浮动车和 AVI 检测数据估计得到的各路段 5 分钟平均行程速度（总计 14 个小时，168 个 5 分钟时间间隔）及真实行程速度数据。基于上述场景和数据，完成如下 6 个任务：

1. 以真实行程速度数据（Sheet 4）为基础，计算整条干线的平均行程速度（可以取各路段行程速度的加权平均值（按照路段长度加权）），并对计算得到的干线平均行程速度进行基本统计分析及可视化（10 分）：
   （1）计算算术平均值、中列数、中位数、标准差、变异系数、最大值、最小值、样本数；
   （2）计算五分位数并画出箱图；
   （3）计算区间频数（以 10km/h 为区间长度）和累计分布频率，并将两个分布图画在一张图上。
2. 针对线圈数据（Sheet 1）中的平均速度、流量和占有率数据进行预处理（15 分）：
   （1）自选方法剔除 Sheet 1 中的异常数据；
   （2）自选方法补全 Sheet 1 中的缺失数据（包含作为异常数据被剔除的数据）。
3. 自选方法对路段 2~5 中的 1 个路段线圈数据进行聚类分析（提示：运用平均速度、流量和占有率三个变量；可考虑分成 3-5 类），分析聚类得到的各类簇数据的统计特征并评价聚类质量（提示：统计特征差异分析：可计算平均速度、流量和占有率三个变量的 5 个基本统计参数：均值、方差、最大值、最小值、样本数；聚类质量评价：可采用教材《数据挖掘概念与技术》10.6 节中的轮廓系数）（15 分）；
4. 选择路段 3 和路段 6，分析线圈和浮动车数据估计得到的行程速度与真实行程速度数据的相异性（欧几里德距离和 DTW 距离）和相关性（相关系数和协方差）（10 分）；
5. 分别计算线圈、浮动车和 AVI 三种检测方式的行程速度估计误差（平均绝对百分误差 MAPE 和均方根误差 RMSE）（10 分）；

提示：（1）MAPE 和 RMSE 的计算公式如下；（2）线圈和浮动车的估计误差可按照每个路段分别计算，然后进行平均；AVI 的估计误差可直接按照干线平均行程速度直接计算。

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{i=N}(x_i-\hat{x}_i)^2}{N}} \quad \mathrm{MAPE}=\frac{\sum_{i=1}^{i=N}\left|\frac{x_i-\hat{x}_i}{x_i}\right|}{N}$$

其中，x_i=真实值，$\hat{x}_i$=估计值；N=样本量。

6. 自选方法对线圈、浮动车和 AVI 检测数据进行融合，估计每个路段的行程速度，要求估计结果尽可能接近给定的真实行程速度（Sheet 4）（即 MAPE 越小越好）（40 分）。

提示：（1）AVI 可以提供较可靠的干道平均速度信息，线圈和浮动车则可以提供每个路段的行程速度信息，交叉口信号控制延误（计算公式见附录）对行程速度的影响较大；（2）可将 14 小时数据集划分为两个部分：前面 10 小时数据作为训练集（用于数据建模和分析），后面 4 小时数据作为测试集（用于模型精度测试），最后 4 小时的测试精度作为评分的主要依据，占本小题得分的 70%，建模合理性和分析逻辑性占 30%。

(二) 作业提交形式和时间

1. 提交形式：（1）程序源代码（编程语言不限，推荐 R 和 Python）；（2）各小题中计算得到的新数据集（提交最终数据集即可，中间过程数据集不需提交）；（3）数据建模和分析报告 1 份（不超过 20 页（小四字体、单倍行距、格式工整），含必要图表、文字解释和中间计算过程说明）。
2. **提交时间：2019 年 12 月 29 日（周日）晚上 12 点之前，请各位同学将上述三个文件打包（邮件标题和作业文件命名规则：2019 年交通信息融合与挖掘大作业-学号-姓名）发送至 1466185960@qq.com 和 tang@tongji.edu.cn（两个邮箱同时发一份）。**

附录：期末大作业数据及场景说明

(一) 数据集说明

数据集一共包含如下 4 个表格：

1. 线圈估计得到的各路段行程速度数据（Sheet 1）：14 个小时内，每 5min 时间间隔内由线圈检测数据估计得到的路段平均行程速度（单位为 km/h），以及每 5min 间隔的占有率和交通流量；
2. 浮动车估计得到的各路段行程速度数据（Sheet 2）：14 个小时内，每 5min 时间间隔内每条路段上由浮动车检测数据估计得到的路段平均行程速度（单位为 km/h）；
3. AVI 估计得到的干线平均行程速度数据（Sheet 3）：14 个小时内，每 5min 时间间隔由 AVI 检测数据估计得到的干道平均行程速度（单位为 km/h）；
4. 各路段真实行程速度数据（Sheet 4）：14 个小时内，每 5min 时间间隔内的真实行程速度（路段 1：从车辆进入 AVI 断面开始，到车辆离开交叉口 1 的停车线为止；路段 2-5：从车辆进入该路段开始，到车辆离开该断面的停车线为止；路段 6：从车辆进入该路段开始，到车辆离开 AVI 断面为止），单位为 km/h。

(二) 干道拓扑结构、交通流量和信号控制方案说明

本次大作业数据取自连云港市某干道的仿真数据（仿真时长为 14 小时），数据采集的行车方向从西向东，该路段一共包含 5 个交叉口和 6 个基本路段，如图 1 所示，其中深蓝色段表示线圈，红色方框表示 AVI 采集设备。图 2 为每个交叉口的基本交通量，表 1 为每个交叉口的信号配时方案（定时控制）。

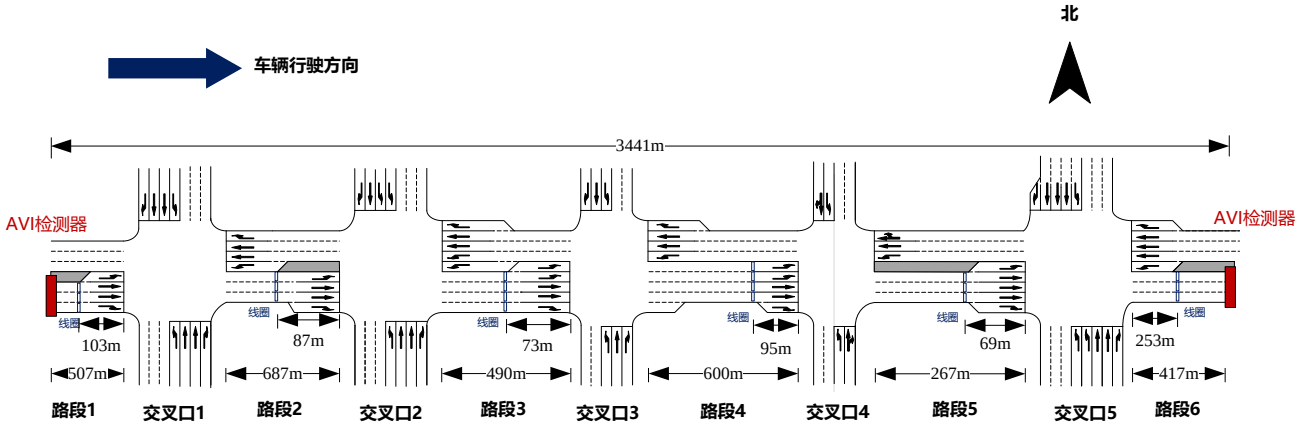


图 1 干道基本示意图

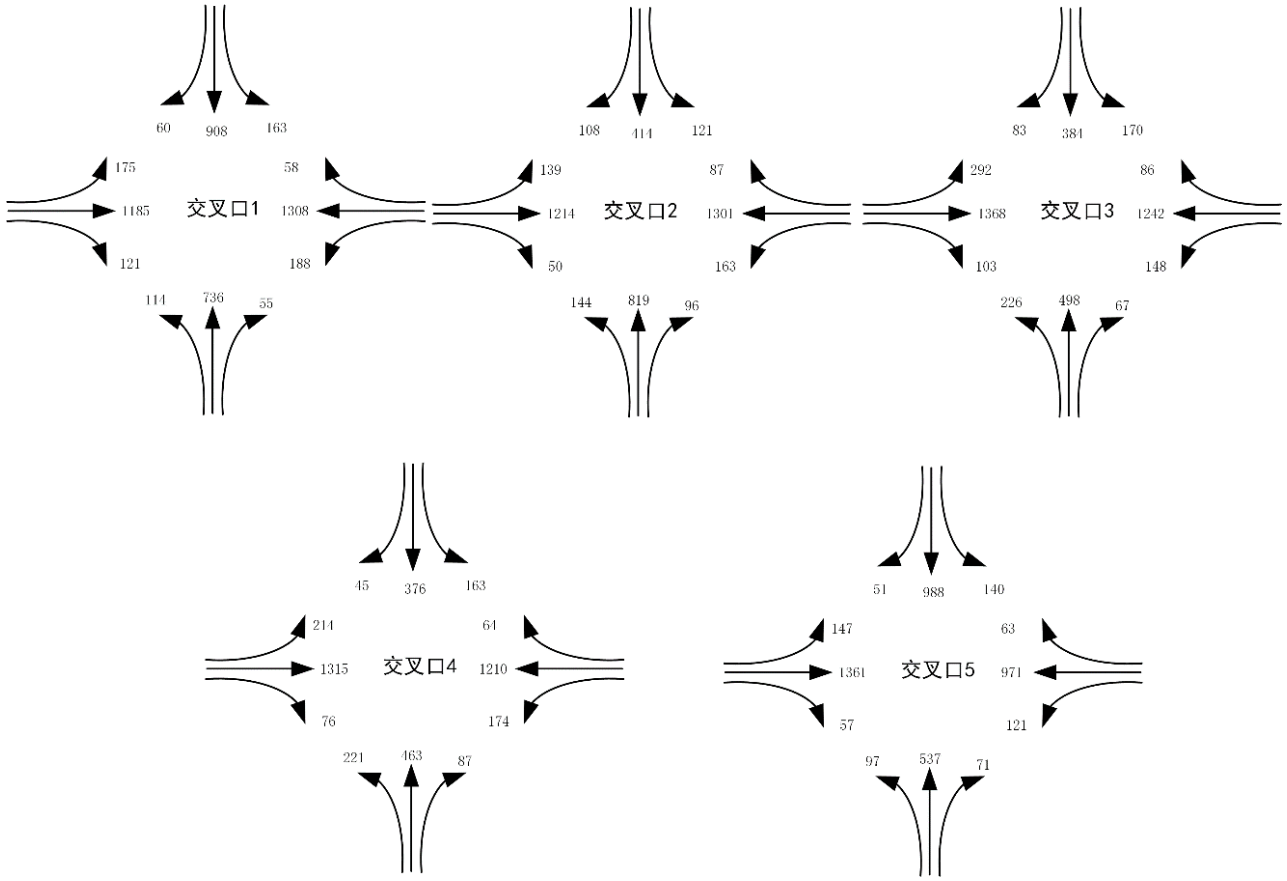
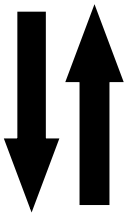
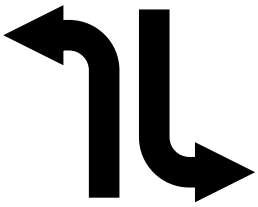
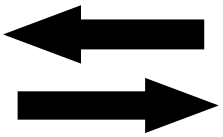
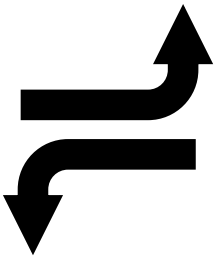


图 2 交叉口机动车交通量（pcu/h）

（注意：本次作业中不考虑行人和非机动车的流量及其影响；上述流量为仿真模型中设定的基本流量，实际仿真过程中，每个时段的流量会存在随机波动）

表 1 各交叉口信号配时方案（单位：秒）

	相位 1：南北直行	**相位 2：南北左转**	**相位 3：东西直行**	**相位 4：东西左转**	**周期**
					
交叉口 1	39s	12s	50s	14s	131s
交叉口 2	36s	11s	56s	12s	131s
交叉口 3	34s	15s	50s	12s	127s
交叉口 4	36s	14s	49s	12s	127s
交叉口 5	32s	18s	45s	19s	130s
注: 所有交叉口每个相位的黄灯均设为 3 秒、全红设为 1 秒，损失时间 4 秒/相位、16 秒/周期；所有交叉口的右转车辆都不受信号灯控制（即红灯期间可右转）。					

(三) 信号控制交叉口延误计算公式

单个已建成交叉口，可通过如下方法计算信控延误指标：

$$d = d_1 + d_2 + d_3$$

$$d_1 = d_s \frac{t_u}{T} + f_s d_u \frac{T - t_u}{T}$$

$$d_2 = 900T[(x-1) + \sqrt{(x-1)^2 + \frac{8ex}{CAP \cdot T}}]$$

$$d_3 = \begin{cases} 3600\frac{Q_b}{CAP} - 1800T[1 - \min[1, x]] & (t_u = T) \\ 1800\frac{Q_b t_u}{T \cdot CAP} & (t_u < T) \end{cases}$$

式中：d为平均信号控制延误(s/pcu)

d_1为均匀延误，即车辆均匀到达所产生的延误 (s/pcu)；

d_2为随机附加延误，即车辆随机到达并引起过饱和周期所产生的附加延误(s/pcu)

d_3为初始排队附加延误，及在延误分析期初停有上一时段留下积余车辆的初始排队使后续车辆经受的附加延误，s/pcu

d_s为饱和延误，s/pcu，可用下式进行计算：

$$d_s = 0.5C(1-\lambda)$$

d_u为不饱和延误，s/pcu，可用下式进行计算：

$$d_u = 0.5C\frac{(1-\lambda)^2}{1 - \min[1, x]\lambda}$$

t_u为在T中积余车辆的持续时间，h，可用下式表示：

$$t_u = \min[T, \frac{Q_b}{CAP[1 - \min[1, x]]}]$$

f_s为绿灯期车流到达率校正系数，按照下式进行计算：

$$f_s = \frac{1-P}{1-\lambda}$$

Q_b为分析期初始剩余车辆，辆，须实测；

P为绿灯期到达车辆占整周期到达量之比，可实地观测；

C为周期时长

λ为所计算车道绿信比

x为所计算车道饱和度；

CAP为所计算车道通行能力（pcu/h）；

T为分析时段持续时长，取0.25h；

e为单个交叉口信号控制类型校正系数，定时控制宜取0.5；感应控制e随饱和度与绿灯延长时间而变，取值范围宜为0.04 ~ 0.5。

※ **在本次作业中，只需计算d_1和d_2, d_3可不考虑。**