系统工程课程设计

自 42 晏筱雯 2014011459

2017.6.3

目录

1 课程设计要求

- 实验名称：系统工程方法在交通数据处理的应用。
- 实验类型：综合设计型实验。
- 实验目的：利用系统工程方法对实际数据完成建模和分析，锻炼自学能力。
- 实验内容：实际交通数据的建模、预测、压缩与聚类分析。
- 实验方法：提供北京市路网交通流原始数据，要求学生基于课堂讲授方法并自学其他方法，完成数据的建模与分析。主要包括以下几点：

  1）基于课堂讲授的黑箱建模方法，对上述数据进行预处理后，建立交通流预测模型，以最后两天的数据为预测值，之前的数据为训练值，给出分时段（5分钟，10 分钟和 15 分钟）预测结果，并给出预测精度（平均绝对误差百分比，平均相对误差等指标）。

  2）基于课堂讲授的主成分分析法，选取不同主成分，对上述数据进行压缩和解压缩，并对比分析压缩比、压缩精度等参数。

  3）基于课堂讲授的 K-means 或系统聚类等聚类分析方法，选取早高峰时段（早 7:00-9:00）的数据，对相同时段各个路口的交通流量进行聚类分析（将路段进行聚类分析研究）；要求：若选择 K 均值聚类，则聚类数目可变化；如选择系统聚类，则要求绘制聚类谱系图。

  4）自学至少一种新的交通流预测方法，仍以最后两天的数据为预测值，之前的数据为训练值，给出分时段（5 分钟，10 分钟和 15 分钟）预测结果，与课堂讲授方法在预测精度方面进行对比分析。

  5）自学概率主成分分析、贝叶斯主成分分析、核主成分分析等方法中的一种或者多种，对上述数据进行压缩和解压缩。与课堂讲授方法在压缩比、压缩精度等参数上进行对比分析。

  6）自学至少一种新的聚类分析方法（可以是 SOM 聚类方法），对同一时段各个路口的交通流量进行聚类分析。

  - 其中 1）2）3）为必完成项目，4）5）6）选其中一项完成即可。可小组完成（每个小组人数不超过 3 人，人数越少，得分越高），满分 10 分。

2 预测模型

2.1 数据预处理

2.1.1 移动平均法

移动平均法（moving average method）是根据时间序列，逐项推移，依次计算包含一定项数的序时平均数，以此进行预测的方法。其中，简单移动平均中各元素的权重都相等，计算公式如下：

$$F_t = \frac{x_{t-1} + x_{t-2} + \cdots + x_{t-n}}{n}$$

其中，F_t是对 t 时刻的预测值，x_i是 i 时刻的历史数据，n 是所用于预测的历史数据的个数。

2.1.2 指数平滑法

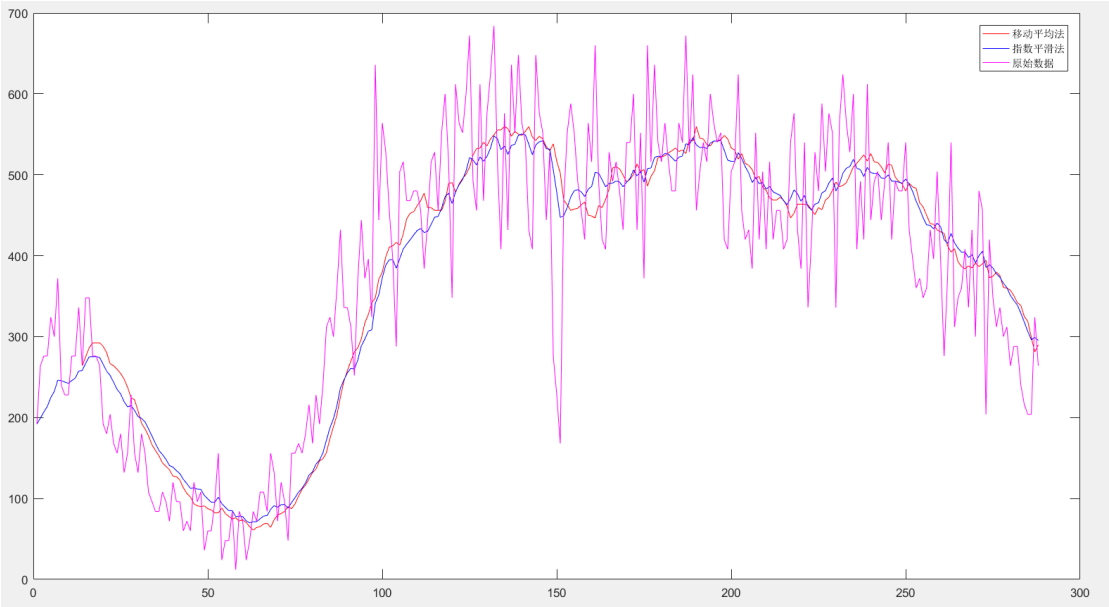
移动平均法则不考虑较远期的数据，并在加权移动平均法中给予近期资料更大的权重；而指数平滑法（Exponential Smoothing）则兼容了全期平均和移动平均所长，不舍弃过去的数据，但是仅给予逐渐减弱的影响程度，即随着数据的远离，赋予逐渐收敛为零的权数。其中，当时间数列无明显的趋势变化，可用一次指数平滑预测，其公式如下：

$$F_t = ax_t + (1-a)F_{t-1}$$

其中，F_t是对 t 时刻的预测值，x_t是 t 时刻的历史数据，a 是一个大于 0 小于 1 的数，其含义是上一时刻的实际值与预测值之差在当前时刻的体现。

2.1.3 处理效果

以第一天第一个检测器的数据为例，取 n=14，a=0.1，运用移动平均法和指数平滑法给出预处理后的数据，并用 plot 绘图如下：



其中紫色为原始数据，红色为移动平均法预处理的数据，蓝色为指数平滑法预处理的数据。可以看出，两种处理方式的效果差不多，因此后面用哪种方法进行预处理都可以。

2.2 黑箱建模

对于给定的 50 个检测器中的每一个检测器，我们用前 14 天预测时刻 t 之前的数据作为 X，其维度为每天 t 时刻之前的已知数据个数，样本数为 14；将前 14 天预测时刻 t 的数据值作为 Y。建立 X 和 Y 之间的关系，然后利用第 15 天和第 16 天 t 时刻之前的已知数据来预测 t 时刻的数据。

2.2.1 归一化

我们对数据作如下处理：

$$X = \begin{bmatrix} \frac{x_{11} - \bar{x}_1}{\sigma_{x_1}} & \cdots & \frac{x_{1n} - \bar{x}_1}{\sigma_{x_1}} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1} - \bar{x}_n}{\sigma_{x_n}} & \cdots & \frac{x_{nn} - \bar{x}_n}{\sigma_{x_n}} \end{bmatrix}$$

$$\bar{x}_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij}\,, \sigma_{x_i} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(x_{ij} - \bar{x}_i)^2}\,, i = 1,2,\dots\dots,n$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \, , \sigma_{\mathrm{y}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2} \, , i = 1,2, \dots \dots , n$$

其中 n 是维数，N 是样本的数量。

2.2.2 基函数

为了解决多项式逼近的缺陷，我们需要用限制基函数起作用的范围，用局部基函数代替全局基函数。我们用到的基函数主要有以下三种：

- x^k
- $Sigmoid$函数：$\frac{1}{1+e^{-x}}$
- $GaussRBF$函数：e^{-x^2}

我们首先将 X 进行归一化，再利用上述基函数构成一组新的 X。

2.2.3 病态问题处理

当矩阵的某些特征值远远小于其他特征值的时候，就会出现病态问题。处理方法如下：

对于矩阵 A：$A = X \times X^T$，通过特征值分解的方法求其特征值及其对应的特征向量，记为：$\lambda = [\lambda_1, \dots, \lambda_n]$和$Q = [q_1, \dots, q_n]$，其中$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$。

根据预先设定的特征值阈值 threshold（0<threshold<1）,寻找满足以下条件的最小的 m：

$$\sum_{i=1}^{m} \lambda_i \Big/ \sum_{i=1}^{n} \lambda_i \geq 1 - \text{threshold}$$

此时有$Q_m = [q_1, \dots, q_m]$.

取 threshold=0.05.

2.2.4 多元线性回归方程

根据以下公式进行回归系数的计算：

$$Z = Q_m^T X$$

$$\hat{d} = (ZZ^T)^{-1} Z Y^T$$

$$\hat{c} = Q_m \hat{d}$$

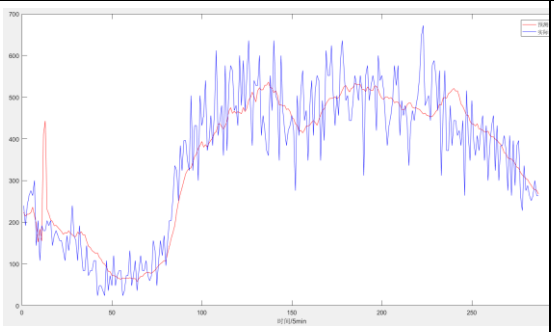
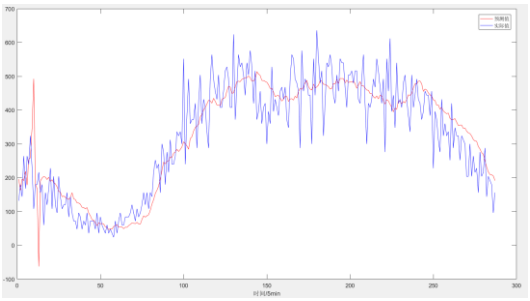
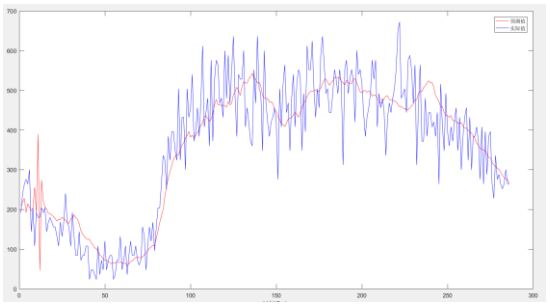
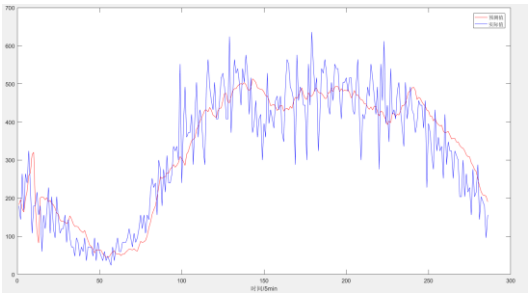
$$\hat{b} = \frac{\sigma_Y \hat{c}^T}{\sigma_X}$$

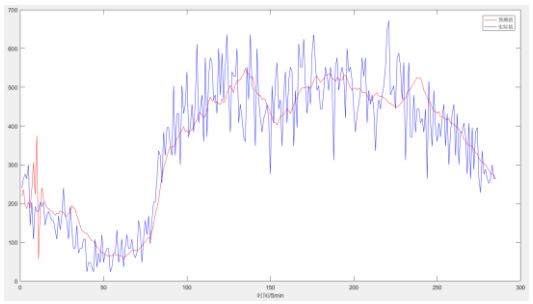
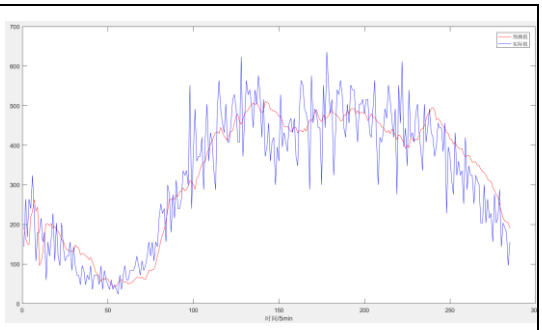
$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

有了回归系数，就可以写出回归方程对数据进行预测了。

2.3 预测结果

以 1 号检测器为例，画出时间间隔为 5min、10min 和 15min 的预测效果图：

<table>
<tr><td colspan="2">5min</td></tr>
<tr><td>Day 15 实际值与预测值对比</td><td>Day 16 实际值与预测值对比</td></tr>
<tr><td>  </td><td>  </td></tr>
<tr><td>绝对误差：56.7467
相对误差：24.29%</td><td>绝对误差：56.1885
相对误差：25.59%</td></tr>
<tr><td colspan="2">10min</td></tr>
<tr><td>Day 15 实际值与预测值对比</td><td>Day 16 实际值与预测值对比</td></tr>
<tr><td>  </td><td>  </td></tr>
<tr><td>绝对误差：56.7298
相对误差：24.37%</td><td>绝对误差：55.1940
相对误差：24.89%</td></tr>
<tr><td colspan="2">15min</td></tr>
<tr><td>Day 15 实际值与预测值对比</td><td>Day 16 实际值与预测值对比</td></tr>
</table>

	
绝对误差：57.2324 相对误差：24.77%	绝对误差：55.3821 相对误差：24.81%

从上述结果可以看出，整体预测效果不错，基本可以预测出数据的变化趋势。但是对于实际交通流曲线中的噪声部分，线性回归模型还是难以预测，因此相对误差较大。

下面对比 50 个检测器在不同时长下的平均误差变化：

相对误差	5min	10min	15min
Day 15	30.57%	30.54%	30.58%
Day 16	31.61%	31.64%	31.66%

由上表可以看出，随着预测时间间隔的增大，预测误差变大，但是变大的幅度并不大。故算法的鲁棒性较好。

以上数据基于步长为 15 的移动平均法，下面以第 15 天为例，测试步长取其他数字时的相对误差：

Step	0	5	10	15
Relevent error	31.57%	27.42%	28.24%	30.57%

以上数据说明，对数据进行预处理能够有效地减少相对误差，并且移动平均法的步长越小，相对误差减小得越显著。

3 交通流数据压缩

首先对数据归一化，步骤与 2.2.1 中相同。

3.1 PCA 压缩

当矩阵的某些行出现线性相关的情况时，可以对数据进行压缩，即找到一系列主

成分，用它们的线性组合来表示其他次要的成分。处理方法如下：

对于矩阵 A：$A = X \times X^T$，通过特征值分解的方法求其特征值及其对应的特征向量，记为：$\lambda = [\lambda_1, \dots, \lambda_n]$和$Q = [q_1, \dots, q_n]$，其中$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$。

根据预先设定的特征值阈值 threshold（0<threshold<1）,寻找满足以下条件的最小的 m：

$$\sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{n} \lambda_i \geq 1 - \text{threshold}$$

选择 threshold 的依据是应尽量使得绝对误差等距变化。

此时有$PCS_m = [q_1, \dots, q_m]$.

取 *threshold=0.05.*

压缩后的数据为：$CPRS = PCS_m' \times X$。

3.2 解压缩

$$X' = PCS_m \times CPRS$$

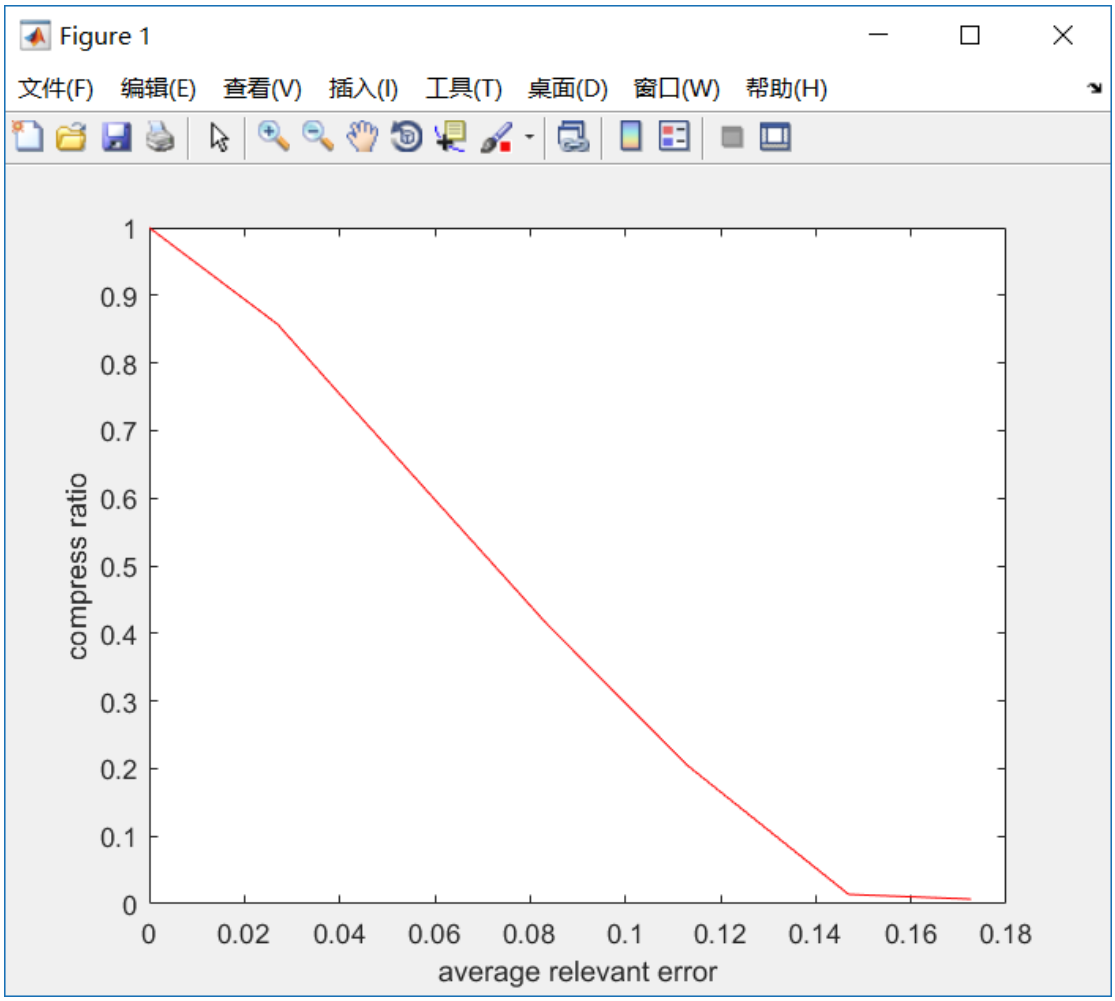
$$X = X' \times \sigma_{x_i} + \bar{x_i}$$

3.3 实验结果分析

用平均相对误差来表示压缩精度：

PCA 阈值	**绝对误差**	**相对误差**	**压缩后数据大小**	**压缩率**
0	0	0	288×800	1.0000
0.005	10.3293	2.69%	247×800	85.76%
0.010	14.6956	3.83%	221×800	76.74%
0.050	32.1206	8.36%	119×800	41.32%
0.100	43.3931	11.30%	7×800	20.49%
0.200	56.4103	14.69%	4×800	2.43%
0.300	66.2627	17.26%	2×800	0.69%

将相对误差和压缩率的关系绘制如下图：



可以看出，在压缩率为 0.1~0.8 区间内，平均相对误差和压缩率有着比较好的线性关系。于是我们可以以需要的压缩率作为依据，选择合适的平均相对误差阈值，进行数据压缩。

4 交通流聚类分析

4.1 K-means 聚类算法

步骤如下：

1）选中心点

根据确定的分类数目 k 在所有样本中挑选 k 个作为初始中心点。在这里我们选择了前 k 个数据作为初始中心点。

2）修改中心点

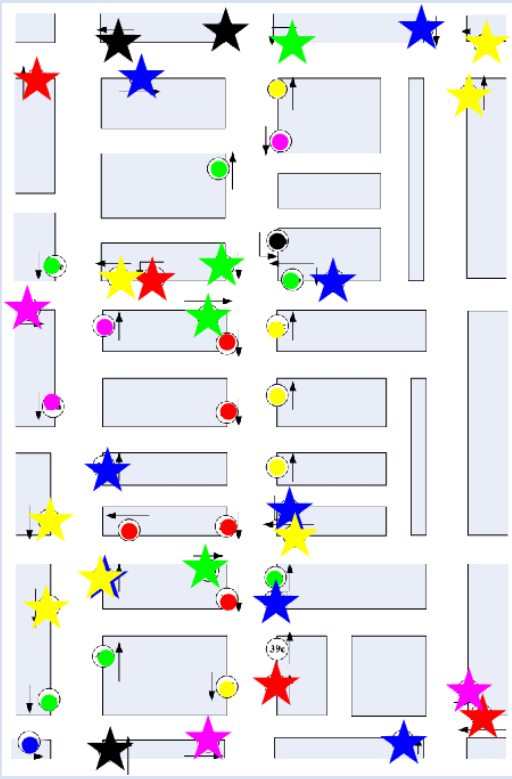
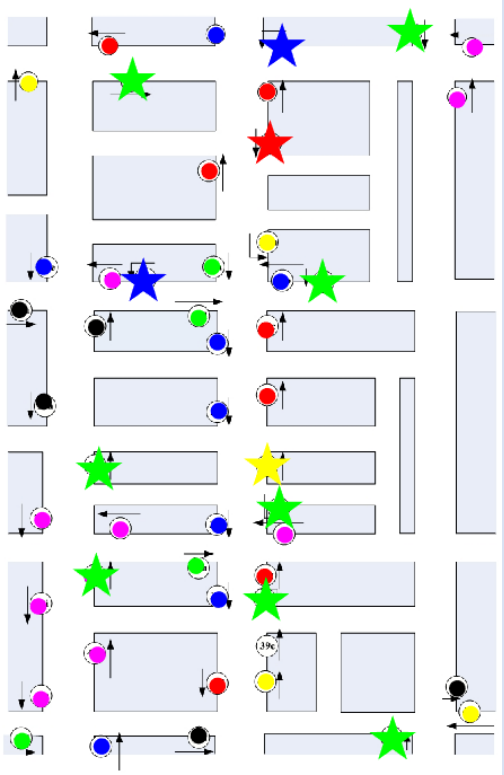
对所有的样本，按照以下步骤进行计算：

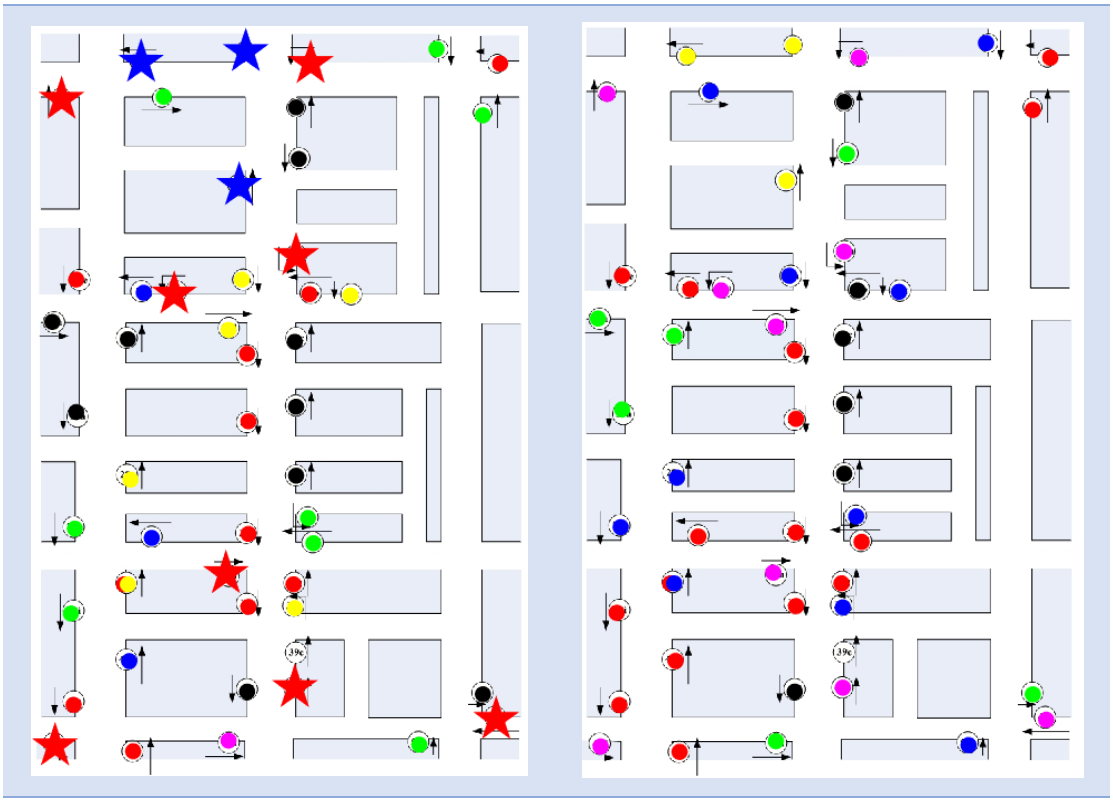
- 将其归入与其中心点坐在的类；
- 重新计算该类的中心并替换原中心。

3）停止准则

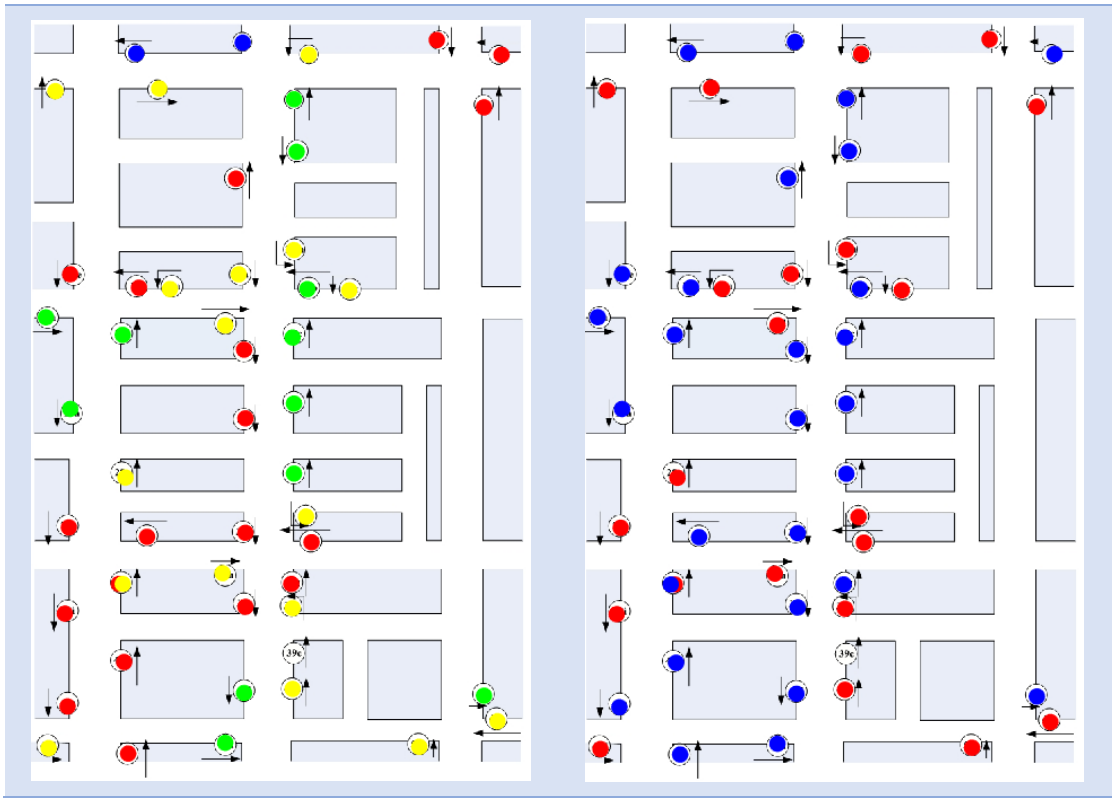
当开始和结束的中心点差别小于阈值，则停止分类。

4.2 结果展示

12 类	10 类
	
8 类	6 类



4 类 2 类



从聚类的结果我们可以看出，某一些被归为一类的点确实具有明显的相关性，但是大部分无法直接看出来，可能有待进一步的处理和分析。

另外，我们必须考虑到交通数据流中不同检测带你的数据存在带有延时的相关性，

但是这种相关性是无法通过 K-means 聚类算法体现出来。

5 总结

经过这次大作业，我对这一学期系统工程课上所学的知识进行了回顾，对这门课的整体框架有了更好的了解。虽然书写代码并不是考试所要求的，但是它帮助我理解了课上所介绍的算法，同时也帮助我梳理了各章之间的关系，收获颇丰。