上海应用技术大学

随机模拟与数据分析

许建强

讲座内容

□ 随机模拟

1. 蒙特卡洛方法
2. 竞赛题1：眼科病床的合理安排
3. 竞赛题2：小区开放对周边道路的影响

□ 数据分析

1. 数据预处理
2. 数据建模的决策与评价方法
3. 数据建模的常用预测方法
4. 数据分析统计方法
5. 竞赛题1：电池剩余放电时间的预测
6. 竞赛题2：葡萄酒的评价

随机模拟的优势

- 对无法建模的一类问题进行模拟
- 对大量方案的比较和选优
- 对大型复杂系统进行模拟
- 对无法重复的现象进行模拟

蒙特卡洛方法的基本思想

- 基本思想：把各种随机事件的概率特征与数学分析的解联系起来，用试验的方法确定事件的相应概率与数学期望
- 特点：概率模型的解是由试验得到的，而不是计算出来的。
- 作用：可以解决其它方法无法解决的实际问题、对理论研究进行补充及辅助

MATLAB中部分随机数产生命令

命令	说明
rand(m,n)	[0, 1]上均匀分布
unifrnd(a,b,m,n)	[a, b]上均匀分布
unifrnd(N,m,n)	1，…, N的等概率分布
randn(m,n)	标准正态分布N(0, 1)
exprnd(λ,m,n)	均值为λ 的指数分布
poissrnd(λ,m,n)	均值为λ 的泊松分布
normrnd(μ,σ,m,n)	正态分布N(μ , σ ^{2})

注：以上都是产生不同分布m×n 阶随机矩阵

例：报童诀窍的简化版

（报童诀窍的简化版）报童每天清晨从报社购进报纸零售，晚上退回没有卖掉的报纸.若每份报纸的购进价为b=0.75元，售出价为a=1元，退回价为c=0.6元.每天需求量X是离散型随机变量，其分布为

X	500	510	520
P	0.34	0.36	0.30

问：如果报童每天购进报纸为n=510份，每天的平均利润是多少？

方法一：概率方法（略）

方法二：如果我们知道每天的需求量，可直接计算利润。而每天需求量可以按分布生成（随机模拟思想）

售出价a=1，购进价b=0.75，退回价c=0.6，

购进数量n=510份

需求量X	500	510	520
概率P	0.34	0.36	0.30

按照需求量的分布规律，随机生成N=20个数据:

510 520 500 510 520 500 500 500 500

510 500 500 500 520 510 520 510 510

代表20天的需求量，计算出报童在这20天的总利润和平均利润，用**平均利润**来近似报童的平均收入。

这就是Monte Carlo方法.

问题 •如何按分布规律产生随机数据？

•随机数据很多时，如何编程？

问题1：如何产生以下分布规律的随机数据？

需求量X	500	510	520
概率P	0.34	0.36	0.30

注:rand(m,n)可以生成[0,1]上均匀分布随机数

把 [0,1]分成长度为0.34、0.36、0.30的三个区间 [0,0.34]、(0.34,0.70]、(0.70,1]

用rand(1,1)产生1个[0,1]上均匀分布随机数，

如该数在[0,0.34]、 (0.34,0.70]或(0.70,1]内， 相当于该天的需求量相应为500、510、520

重复多次就可以若干天的需求量了

生成N=20天的需求量的matlab代码可以为

生成N=20天的需求量的matlab代码（定义函数）

```
function
y=randfun1(N)
y=zeros(1,N);
for i=1:N
      t=rand(1,1);
   if  t<=0.34
      X=500;
   elseif  t<=0.70
      X=510;
   else
      X=520;
   end
   y(i)=X;
end
```

文件名为randfun1.m

产生1行N列的全0矩阵，目的分配好矩阵大小，可以省略

根据随机数t的范围，确定需求量X值，并保存到数组的相应位置中（关键部分）

问题2：如何从需求量计算利润？售出价a=1，购进价b=0.75，退回价c=0.6，购进数n=510

函数文件名fun2.m

```
function  y=fun2(x)
a=1;        %售出价
b=0.75;  %购进价
c=0.6;    %退回价
n=510;    %购进数量
if(x>n)
        y=n*(a-b);
else
        y=x*(a-b)-(n-
x)*(b-c)
end
```

$$y=\begin{cases} n(a-b) & x>n \\ x(a-b)-(n-x)(b-c) & x\le n \end{cases}$$

模拟程序代码

```
N=1000;
x=randfun1(N);
y=0;
for  i=1:N
    y=y+fun2(x(i));
end
y/N
```

也可完整模拟程序：

```
a=1;b=0.75;c=0.6;n=510;
n=510; N=1000; y=0;
for i=1:N
    t=rand(1,1);
    if (t<=0.34) x=500;
    elseif (t<=0.70) x=510;
    else  x=520;
    end
    if (x>n) y=y+n*(a-b);
    else y=y+x*(a-b)-(n-x)*(b-c);
    end
end
fprintf(1,'平均利润=%.3f',y/N);
```

售出价a　购进价b　退回价c
购进数n　总利润y
模拟天数N

根据随机数t，
计算需求量x值

根据需求量x，
计算利润并累
加到y中。

显示平均利润

竞赛题1：眼科病床的合理安排

- 赛题介绍：B2009

数据分析与检验

- 在着手解决问题前首先应对所给数据进行分析，从中获得对解题有用的关键信息。
- **FCFS（First come, First serve）**规则安排住院，这样虽然对病人很公平，但缺乏效率。例如根据**FCFS**原则，白内障双眼的病人可能会在星期一入院，但医院规定“白内障双眼的患者在星期一做一只眼，在星期三做另一只眼”，所以该患者的术前准备时间就变为**7**天，而一般情况下，白内障患者的术前准备时间只需**1-2**天，这样的情况会导致不必要的术前准备，降低病床的有效利用率，延迟其它类病人的入院时间，进而使得病人队列越来越长。

建模案例：眼科病床的合理安排

指标定义

- 效率指标和公平性指标
- 效率指标——平均术前住院时间，或病床有效利用率。

非外伤病人入院第2日（白内障）或第3日（其他眼病）后等待手术的时间称为病床无效时间，病床有效利用率定义为：

病床有效利用率= 1 - 病床无效时间 / 该病人住院时间

- 公平性指标——FCFS规则是最公平的，公平度可以有多种定义方式，例如：以FCFS规则下所有非急症患者的入院等待时间的方差作为一个基准值，其他规则下的方差与之相比来定义公平度。

建模案例：眼科病床的合理安排

- 思路—对每一位等待入院病人，以该病人当日入院的公平性（等待时间）与病床使用效率（分类考虑）两方面综合排序（例如求两个指标的加权和），然后按排序结果安排当日入院病人，由此得到公平合理的住院方案。按此方案进行仿真，再统计各项评价指标值，并与FCFS方案作比较。

i	1	2	3	4	5
疾病类型	白内障单眼	白内障双眼	青光眼	视网膜类疾病	外伤

- 算法—

1. 建立术前等待时间矩阵

t_{ij}——为第i类疾病在周j日入院至手术时需等待时间，对于白内障双眼患者，其术前准备时间在这里认为是从住院到第二次手术的时间。

$$t=\begin{bmatrix} 2 & 1 & 5 & 4 & 3 & 2 & 1 \\ 7 & 6 & 7 & 6 & 5 & 4 & 3 \\ 3 & 2 & 2 & 2 & 2 & 3 & 2 \\ 3 & 2 & 2 & 2 & 2 & 3 & 2 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

建模案例：眼科病床的合理安排

2. 术后观察时间

通过观察数据发现术后住院时间是随机的，所以要做拟合检验。

拟合检验——各类病人术后住院时间分别服从正态分布 、Γ分布 或埃尔朗分布，由于检验方法或检验细节处理不相同，可能得到以上不同的分布。

t_{is} ——为第 *i* 类疾病术后平均观察时间。计算每种疾病最大可能住院时间。

$$t_{i\max} = \max\left\{t_{ij} + t_{is}\right\},\ j = 1, 2\cdots, 7$$

对每位排队的病人按下式计算优先级***R***

$$R = \left(1 + \frac{t_w}{t_{ij} + t_{is}}\right)\frac{t_{i\max}}{t_{ij}}$$

t_w ——为该病人已等候的时间。

由优先级的定义可知，等待时间越长，住院时间越短，则优先级越高。

注：对于不同时间住院的患者，他们的优先级是不同的。

建模案例：眼科病床的合理安排

3. 仿真（FCFS算法流程）

- Step1：初始化：时间为8月30日，待入院人数为93，空床位数为14；估计当日住院病人的出院时间（用前面拟合分布生产随机数）。
- Step2：判断待入院人数是否大于零，是则转下一步，否则结束；
- Step3：计算当天出院人数；
- Step4：判断剩余床位数是否大于零，是则进行下一步，否则时间加一天，所有待入院患者的等待时间加一天，跳回Step3；
- Step5：选出待入院病人中排在最前面的患者A（外伤优先），安排患者A入院，并确定A的手术时间和出院时间（用前面拟合分布生产随机数）；待入院人数-1；
- Step6：判断剩余床位数是否大于零，是则返回Step5，否则时间加一天，所有待入院患者的等待时间加一天，跳回Step2。
- Step8：统计效率指标和公平度指标

建模案例：眼科病床的合理安排

3. 仿真（改进算法1流程）

- Step1：初始化：时间为8月30日，待入院人数为93，空床位数为14；估计当日住院病人的出院时间（用前面拟合分布生产随机数）。
- Step2：判断待入院人数是否大于零，是则转下一步，否则结束；
- Step3：计算所有待入院患者的优先级；
- Step4：计算当天出院人数；
- Step5：判断剩余床位数是否大于零，是则进行下一步，否则时间加一天，所有待入院患者的等待时间加一天，跳回Step3；
- Step6：选出待入院病人中优先级最大的患者A（外伤特殊考虑），安排患者A入院，并确定A的手术时间和出院时间（用前面拟合分布生产随机数）；待入院人数-1；
- Step7：判断剩余床位数是否大于零，是则返回Step6，否则时间加一天，所有待入院患者的等待时间加一天，跳回Step2；
- Step8：统计效率指标和公平度指标；与FCFS的指标进行比较分析。

建模案例：眼科病床的合理安排

患者的入院时间的估计

根据排队论的知识，可以假定各类病人到达人数分别服从不同参数的Poisson分布，但是需要进行分布拟合检验及分布参数提取。（k-s检验）

3. 仿真（算法2）

- Step1：导入当前病床使用，病人排队情况，时间t_0初始化为9月12日；
- Step2：按对应泊松分布产生的随机数作为第t天各类患者的门诊人数并进行随机排序，共产生N周数据，输入每个病人的门诊日期d，和病人的疾病类型i；
- Step3:判断t是否大于N*7，是则统计不同类型患者在不同日等待时间的最小值和最大值（如果病人数足够多也可以用置信区间 （置信度为1-α）来估计等待时间）并结束，否则进行下一步；
- Step4：用算法1计算队列中患者的优先级对患者进行排序；计算当天出院人数；
- Step5：判断门诊时间为d的i类患者能否如愿，是则记录其等待时间和出院时间，继续考虑下一个 患者，记录最后一个如愿患者的门诊时间的d，t=d+1，所有待入院患者的等待时间加1，返回Step3。否则日期t加1且所有待入院患者的等待时间加1，然后返回Step3。

建模案例：眼科病床的合理安排

患者的入院时间的估计

结果：**从9月12日起**，周一到周日，各类病人的等待入院时间区间，结果如下表

	周一		周二		周三		周四		周五		周六		周日	
	min	max	min	max	min	max	min	max	min	max	min	max	min	max
1	6	8	5	9	3	6	3	8	7	8	6	7	6	7
2	3	4	3	4	2	3	2	3	6	7	5	6	4	5
3	10	15	9	16	13	19	14	15	13	13	12	13	11	16
4	16	22	16	21	15	20	14	19	13	18	17	22	11	18
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1

结果检验：利用之前的数据检查结果是否合理

建模案例：眼科病床的合理安排

其他问题—周六周日不安排手术

- 目前方案：周一、周三安排白内障手术
- 调整方案：周二、周四或周三、周五安排白内障手术
- 不同方案，各类患者的术前手术时间将会发生变化。
- 利用算法1进行仿真，对评价指标进行比较，选择最合理的方案
- 结果：周三、周五较合理

竞赛题2：小区开放对周边道路的影响

- 城市规划和交通管理部门希望你们建立数学模型，就小区开放对周边道路通行的影响进行研究，为科学决策提供定量依据，为此请你们尝试解决以下问题：
- 1. 请选取合适的**评价指标体系**，用以评价小区开放对周边道路通行的影响。
- 2. 请建立关于车辆通行的**数学模型**，用以研究小区开放对周边道路通行的影响。
- 3. 小区开放产生的效果，可能会与小区结构及周边道路结构、车流量有关，请选取或构建**不同类型的小区**，应用你们建立的模型，定量比较各类型小区开放前后对道路通行的影响。

建模案例：小区开放对周边道路的影响

- 评价指标体系
  - 1. **通行能力评价指标**：单位时间通行流量，平均车速，流量饱和度，……
  - 2. 安全性评价指标：统计路口冲突次数，作为安全性评价指标。
  - 3. 脆弱性评价指标：一种路网结构的度量指标，可以通过计算堵塞概率的变化来反映

建模案例：小区开放对周边道路的影响

- **数学模型（模拟仿真）**

没有小区开放后的直接数据可供使用，只能通过建立仿真模型来评价，好的仿真模型应该重点考虑以下几个方面：

1. 小区周边道路（含小区道路）的流量分配；
2. 交叉路口的通行模型（有无信号灯控制）；
3. 周边道路和小区道路上车辆的最大行驶速度的区别。

建模案例：小区开放对周边道路的影响

- **1. 小区周边道路（含小区道路）的流量分配**；
- **均匀分配：**将进入讨论的小区周边道路区域的车辆按车道数平均分配；**随机均衡分配：**流量分配不事先确定，在仿真中，车辆在每 一个路口根据一定概率随机选择下一条道路，此概率与路径长短或行驶时间长短有关；**Wardrop均衡分配：**每次分配选择通行时间最小的路径进行分配。

建模案例：小区开放对周边道路的影响

- **2.交叉路口的通行模型**
- **有信号灯控制路口的仿真**

  确定t_g/T是绿信比，主路绿灯时间可以设置稍长。
- **无信号小区路口的仿真**

  进入交叉路口区域，通行效率下降（车速下降）。

通行效率折算系数表

车速 V（km/h）	交叉口之间距离（m）					
	100	200	300	400	500	600
20	0.45	0.62	0.71	0.76	0.80	0.83
30	0.31	0.48	0.58	0.65	0.70	0.73
40	0.23	0.38	0.48	0.55	0.60	0.64
50	0.18	0.30	0.39	0.46	0.52	0.56

建模案例：小区开放对周边道路的影响

- **3.仿真模型**
- N-S模型：利用**元胞自动机**模拟，为了降低仿真难度可以作一些简化处理，例如：不考虑变道与超车，不考虑车辆不同，不考虑行人与非机动车，路口处流量分配可成批处理，冲突延误时间固定等
- 利用**VISSIM**软件，必须具体给出软件参数是如何设置的。

基于元胞自动机的模拟仿真

一、什么是元胞自动机

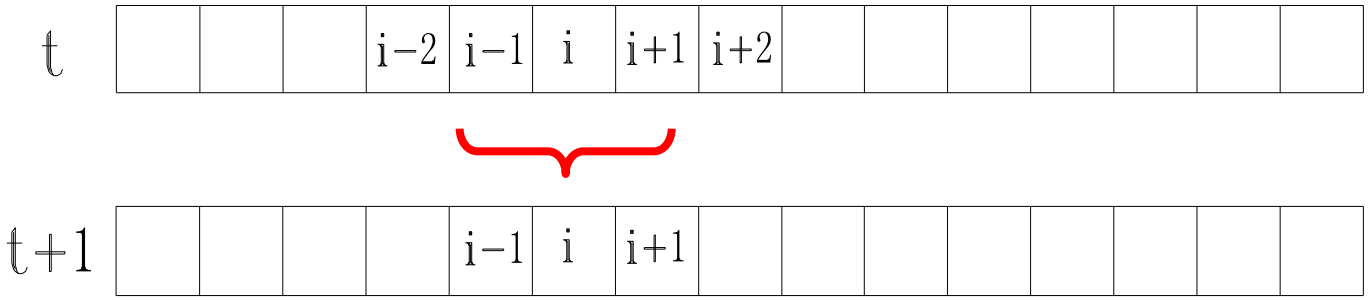
- 元胞自动机（**Cellular Automata，CA**）是一种时空离散的局部动力学模型，是研究复杂系统的一种典型方法，特别适合用于空间复杂系统的时空动态模拟研究。
- 元胞自动机不是由严格定义的物理方程或函数确定，而是用一系列模型构造的规则构成。凡是满足这些规则的模型都可以算作是元胞自动机模型。因此，元胞自动机是一类模型的总称，或者说是一个方法框架。

- 在CA模型中，散布在规则格网 (Lattice Grid)中的每一元胞(Cell)取有限的离散状态，遵循同样的作用规则，依据确定的局部规则作同步更新。大量元胞通过简单的相互作用而构成动态系统的演化。
- CA模型的特点：时间、空间、状态都离散，每个变量只取有限多个状态，且其状态改变的规则在时间和空间上都是局部的。

二、初等元胞自动机

- 初等元胞自动机是状态集S只有两个元素$\{s_1, s_2\}$，即状态个数k=2，邻居半径r=1的一维元胞自动机。由于在S中具体采用什么符号并不重要，它可取 {0，1}，{-1，1}，{静止，运动} 等等，重要的是S所含的符号个数，通常我们将其记为{0，1}。此时，邻居集N的个数$2 \cdot r=2$，局部映射f：$S_3 \rightarrow S$可记为:

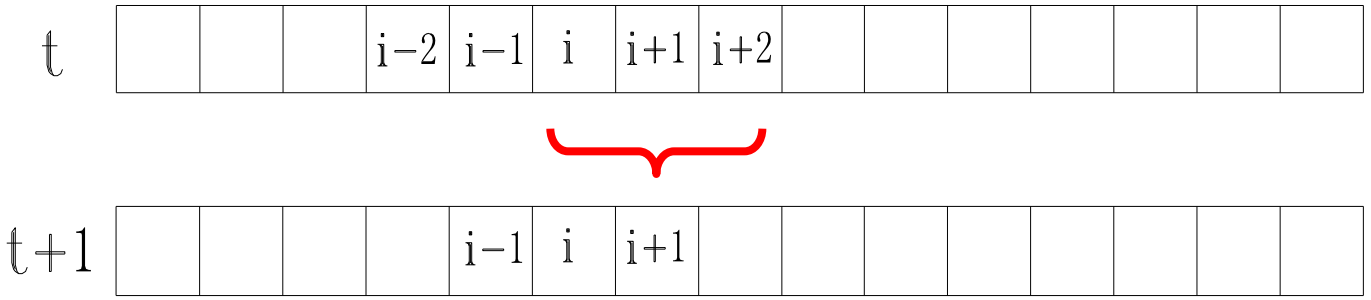
$$S_i^{t+1} = f(S_{i-1}^t, S_i^t, S_{i+1}^t)$$



二、初等元胞自动机

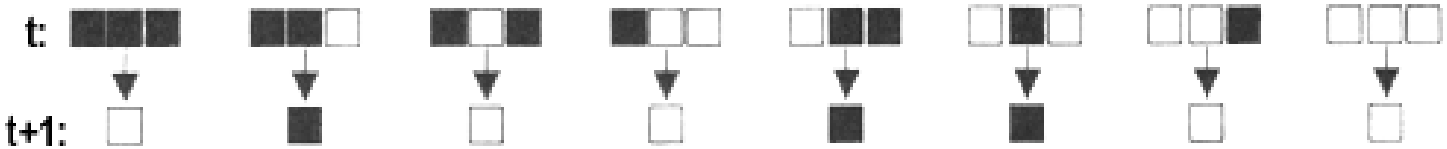
- 初等元胞自动机是状态集S只有两个元素$\{s_1, s_2\}$，即状态个数k=2，邻居半径r=1的一维元胞自动机。由于在S中具体采用什么符号并不重要，它可取 {0，1}，{-1，1}，{静止，运动} 等等，重要的是S所含的符号个数，通常我们将其记为{0，1}。此时，邻居集N的个数$2 \cdot r=2$，局部映射f：$S_3 \rightarrow S$可记为:

$$S_i^{t+1} = f(S_{i-1}^t, S_i^t, S_{i+1}^t)$$



S. Wolfram的初等元胞自动机

t	111	110	101	100	001	010	001	000
t+1	0	1	0	0	1	1	0	0

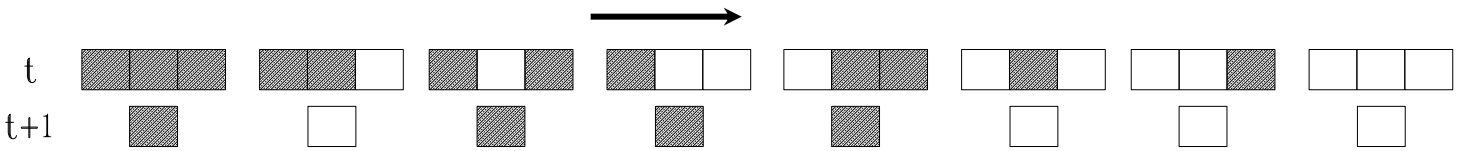


由于只有0、1两种状态，

所以函数*f*共有2^8=256种状态。

特别注意：第184号规则

车辆行驶规则为：黑色元胞表示被一辆车占据，白色表示无车，若前方格子有车，则停止。若前方为空，则前进一格。



t	111	110	101	100	011	010	001	000
t+1	1	0	1	1	1	0	0	0

1992年，德国学者Nagel和Schreckenberg在第184号规则的基础上提出了一维交通流CA模型，即，NS 模型（或NaSch模型）

三、NS 模型

- 在第184号规则的基础上，1992年，德国学者Nagel和Schreckenberg提出了一维交通流CA模型，即，NS 模型（或NaSch模型）
- Nagel and Schreckenberg. A Cellular automaton model for freeway traffie．Journal of Physics(France)，1992
- CA模型最基本的组成包括四个部分:元胞(cell)、元胞空间(lattice)、邻域(neighbor)及更新规则(rule)。

- **NS模型是一个随机CA交通流模型，每辆车的状态都由它的速度和位置所表示，其状态按照以下演化规则并行更新：**

- a）加速过程：$v_n \to \min(v_n + 1, v_{\max})$

- b）安全刹车过程：$v_n \to \min(v_n, d_n - 1)$

- c）随机慢化过程：$v_n \to \max(v_n - 1, 0)$
  （以随机慢化概率p）

- d）位置更新：$x_{n+1} \to x_n + v_n$

$d_n = x_{n+1} - x_n - L$　其中：L---车辆长度~7.5m

NS模型的演化规则：

1）加速：司机总是期望以最大的速度行驶

2）安全刹车：为避免与前车碰撞

3）随机慢化（以随机慢化概率p）：由于不确定因素

   a）过度刹车

   b）道路条件变化

   c）心理因素

   d）延迟加速

4）位置更新：车辆前进

5）边界条件：周期型，随机型

例：设 $v_{\max} = 2$

a）加速过程

b）安全刹车过程

c）随机慢化过程
（以随机慢化概率*p*）

d）位置更新

Configuration at time t:

a) Acceleration:

b) Braking:

c) Randomization ($p = 1/3$):

d) Driving (= configuration at time $t + 1$):

评价指标参数

- 1.车流密度：整个车道中每个元胞内某一瞬时平均存在的车辆数。

$$\rho = \frac{N_{\text{total}}}{N_L \cdot L}$$

其中，L 表示测试路段的元胞数；N_L代表车道数；N_{total} 表示所有车道中车辆总数。

评价指标参数

- 2.平均车速：在固定路段内不同车道上所有车辆的平均时速的平均值。

$$\bar{v} = \frac{1}{N_{\text{total}} \cdot T} \sum_{t=1}^{T} \sum_{j=1}^{N_{\text{total}}} v_j(t)$$

其中，$v_j(t)$表示第j辆车在时刻t的速度。

评价指标参数

- 3.交通流量：单位时间内通过某一固定点的车辆数。根据交通流的理论，交通流量可定义为车流密度与平均车速的乘积。

$$Q = \rho \bar{v}$$

条件：

- 随机慢化概率p；
- 密度ρ=13veh/km/lan（0.1）

  ρ=20veh/km/lan（0.17）

  ρ=27veh/km/lan（0.23）

  ρ=33veh/km/lan（0.28）
- 车辆长度~7.5m；道路长度L=7.5m×120=900m
- 速度：1 ~ 7.5m/s=27km/h；

  2 ~ 2×7.5m/s=54km/h；

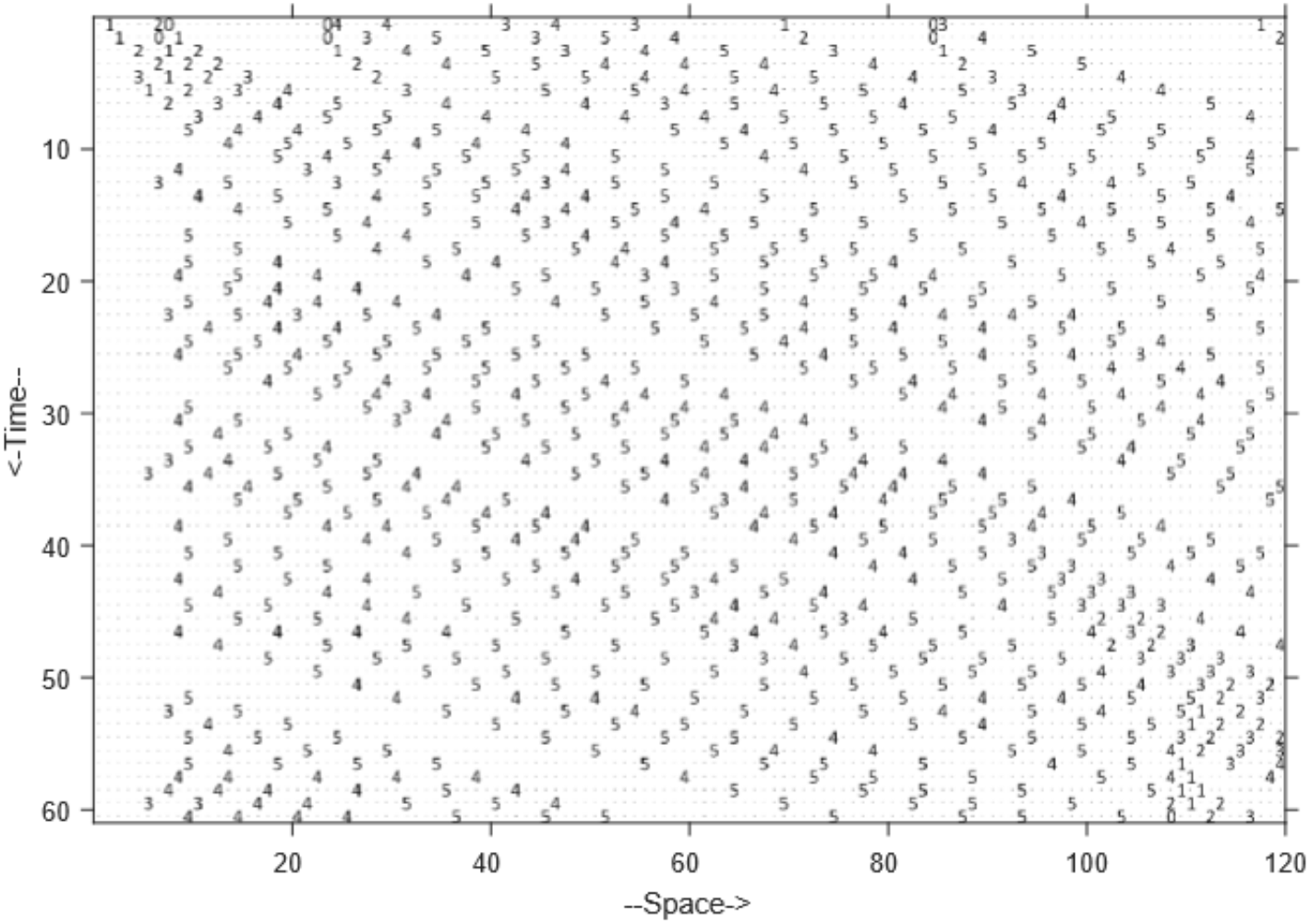
  3 ~ 3×7.5m/s=81km/h；

  4 ~ 4×7.5m/s=108km/h；

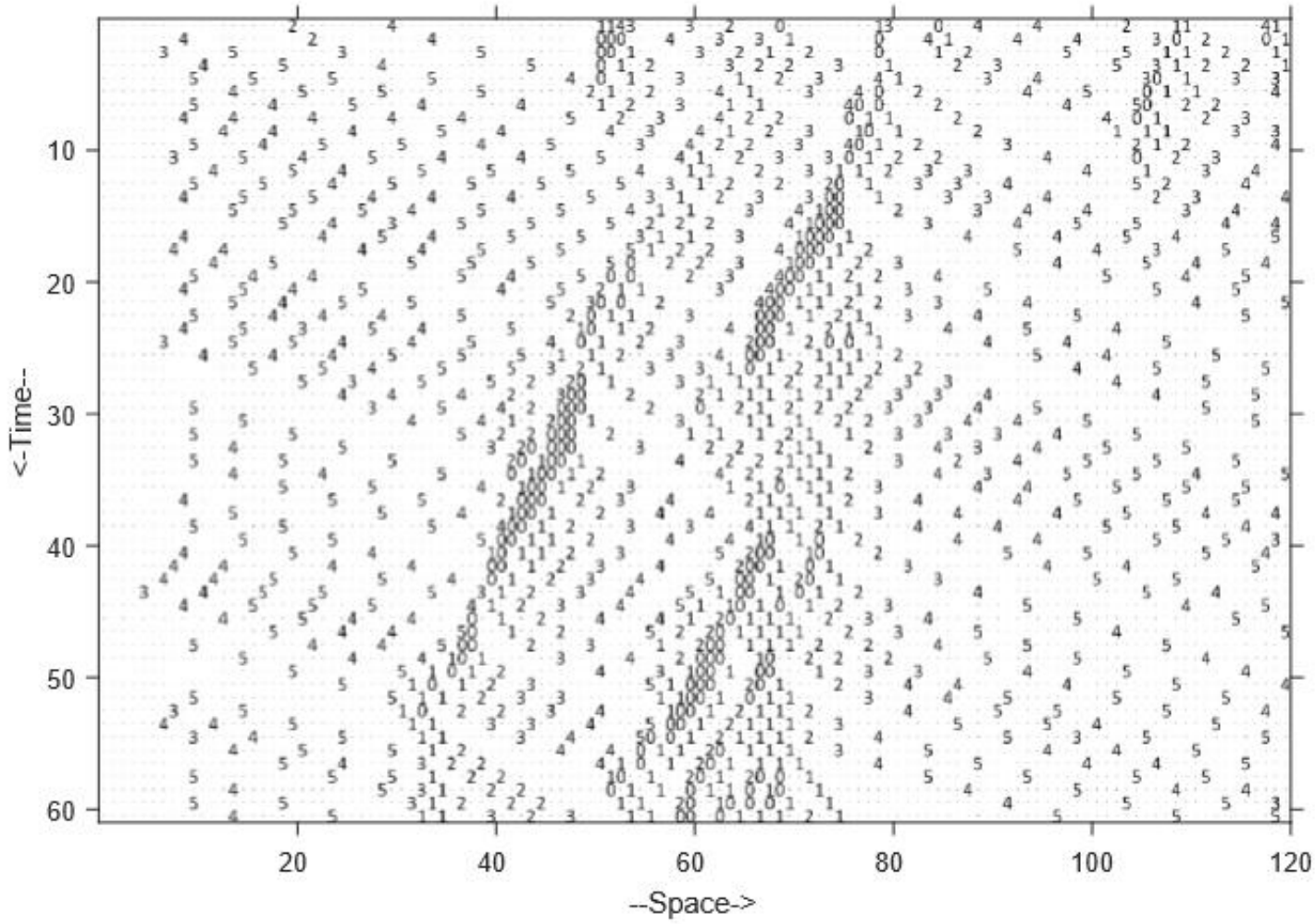
  5 ~ 5×7.5m/s=135km/h；

随机慢化概率p=0.2；密度ρ=13veh/km/lan（0.1）；

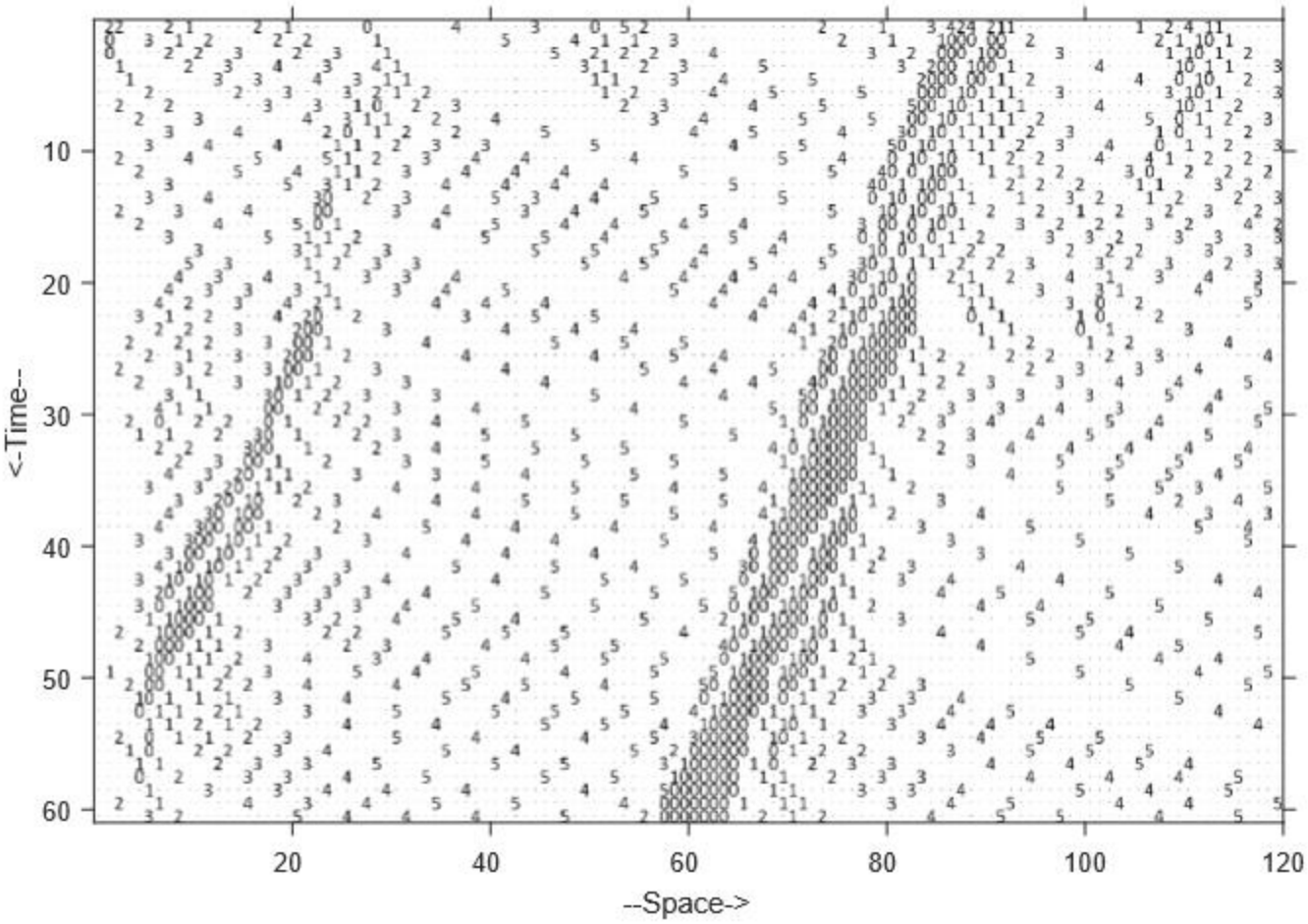
初始随机



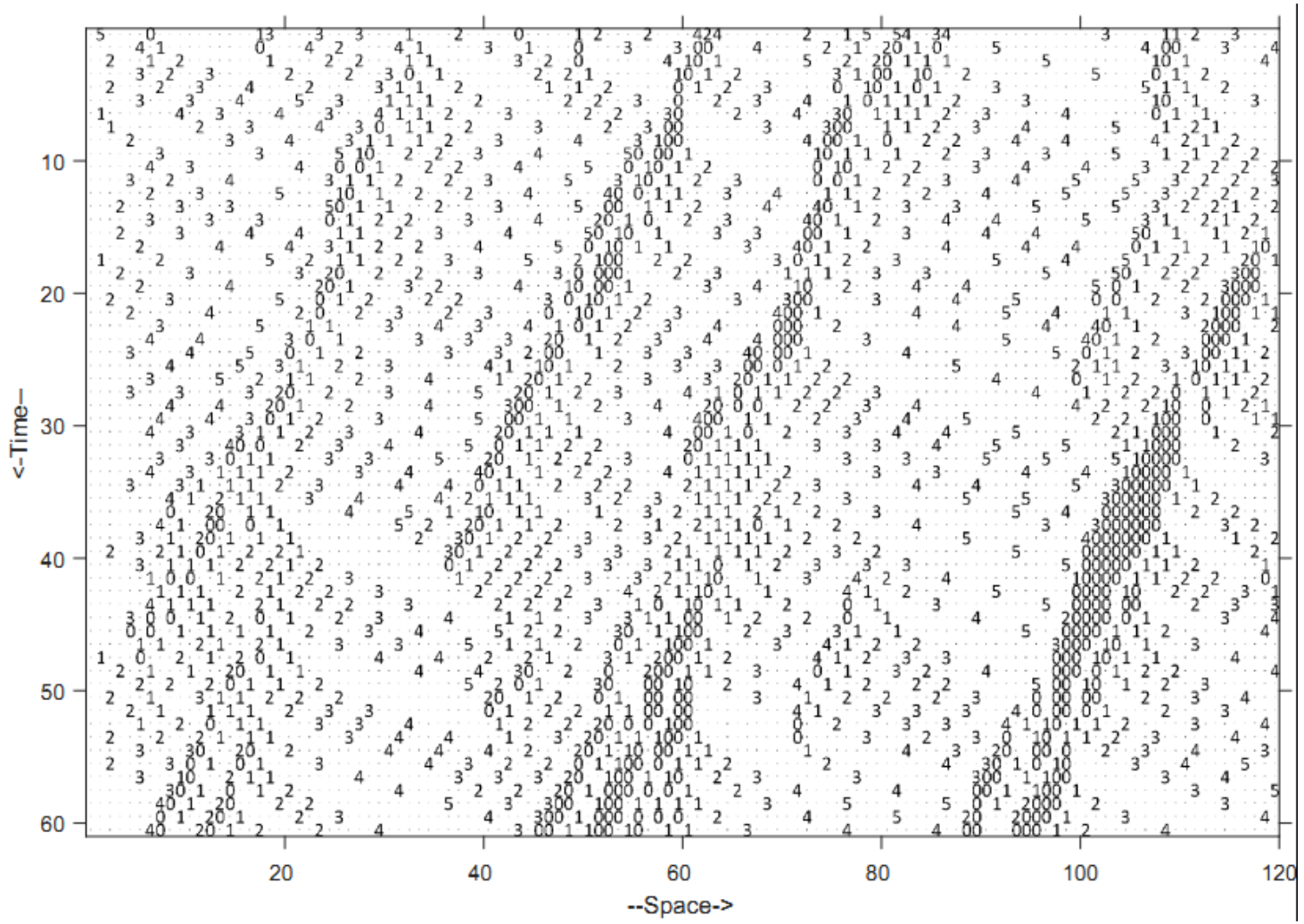
随机慢化概率p=0.2；密度ρ=20veh/km/lan（0.17）；



随机慢化概率p=0.2；密度ρ=27veh/km/lan（0.23）；



随机慢化概率p=0.2；密度ρ=33veh/km/lan（0.28）；



交通流CA模型的主要优点：

（1）模型简单，特别易于在计算机上实现。

（2）能够再现各种复杂的交通现象，反映交通流特性。在模拟过程中人们通过考察元胞状态的变化，不仅可以得到每一辆车在任意时刻的速度、位移以及车头时距等参数描述交通流的微观特性，还可以得到平均速度、密度、流量等参数，呈现交通流的宏观特性。

（3）能够再现单车道、多车道以及路网的交通流建模；机动车和非机动车交通流的建模

随机模拟方法总结

- 系统评价指标确定
- 收集、整理数据
- 分布拟合、检验
- 生成随机数
- 按照一定的算法进行仿真（根据问题的实际）
- 利用仿真数据对指标进行计算
- 结果合理性检验、分析

数据分析

数据预处理

- 收集数据
- 整理数据
  - 异常数据检测 盒子图
  - 缺省数据补充 预测、平均、补0（根据实际情况）
  - 单位、数量级
- 数据描述分析
- 散点图、直方图
- 数据分布的拟合与检验

数据预处理

数据类型的一致化处理方法

- 极大型:期望取值越大越好；
- 极小型:期望取值越小越好；
- 中间型:期望取值为适当的中间值最好；
- 区间型:期望取值落在某一个确定的区间为最好。

数据类型的一致化处理方法

（1）极小型：对某个极小型数据指标 x，

则 $x' = \dfrac{1}{x}(x > 0)$，或 $x' = M - x$.

（2）中间型：对某个中间型数据指标 x，则

$$x' = \begin{cases} \dfrac{2(x-m)}{M-m}, & m \le x \le \dfrac{1}{2}(M+m) \\ \dfrac{2(M-x)}{M-m}, & \dfrac{1}{2}(M+m) \le x \le M \end{cases}$$

数据类型的一致化处理方法

（3）区间型：对某个区间型数据指标 x ，则

$$x' = \begin{cases} 1 - \dfrac{a - x}{c}, & x < a \\ 1, & a \le x \le b \\ 1 - \dfrac{x - b}{c}, & x > b \end{cases}$$

其中 $[a,b]$ 为 x 的最佳稳定区间，$c = \max\{a - m, M - b\}$ ，M 和 m 分别为 x 可能取值的最大值和最小值。

数据指标的无量纲化处理方法

在实际数据指标之间，往往存在着不可公度性，会出现“大数吃小数”的错误,导致结果的不合理。

- 标准差法：

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$s_j = \left[\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2\right]^{1/2}$$

- 极值差法：

$$x'_{ij} = \frac{x_{ij} - m_j}{M_j - m_j}$$

$$M_j = \max_{1\le i\le n}\{x_{ij}\}$$

$$m_j = \min_{1\le i\le n}\{x_{ij}\}$$

$$x'_{ij} \in [0,1]$$

- 功效系数法：

$$x'_{ij} = c + \frac{x_{ij} - m_j}{M_j - m_j}\cdot d$$

$$x'_{ij} \in [c,d]$$

数据建模的决策与评价方法

- **指标权重的确定：信息熵法**

1. 指标的归一化

$$r_{ij} = \frac{x_{ij}}{\sum_i x_{ij}}$$

2. 计算各指标的熵

$$E_j = -k\sum_i r_{ij} \ln r_{ij} \qquad k = 1/\ln m, \quad j = 1,2,\cdots,n$$

3. 计算各指标的区分度

$$F_j = 1 - E_j, 0 \le F_j \le 1$$

4. 计算各指标的权重 $w_j = \dfrac{F_j}{\sum_{j=1}^{n} F_j}, \quad j = 1,2,\cdots,n$

数据建模的决策与评价方法

- 综合方法

1. 加权和法

$$v_i = \sum_{j=1}^{n} r_{ij} w_j, \quad i = 1, 2, \cdots, n$$

2. 加权积法

$$v_i = \prod_{j=1}^{n} r_{ij}^{w_j}, \quad i = 1, 2, \cdots, n$$

3. **TOPSIS法**：是通过检测评价对象与最优解、最劣解的距离来进行排序，若评价对象最靠近最优解同时又最远离最劣解，则为最好；否则不为最优。其中最优解的各指标值都达到各评价指标的最优值。最劣解的各指标值都达到各评价指标的最差值。
4. 层次分析法、模糊综合评价法、秩和比法等。

数据建模的常用预测方法

- 插值与拟合方法：小样本内部预测

应用案例—雨量预报方法的评价

- 回归模型方法：大样本的内部预测

应用案例—电力市场的输电阻塞管理

- 灰预测GM(1,1)：小样本的未来预测(外推）

应用案例—SARS的传播问题

- 时间序列方法：大样本随机因素或周期特征的未来预测；

应用案例—长江水质的评价与预测

- 神经网络方法：大样本的未来预测（外推）

数据建模的统计方法

- 常用统计模型：
  - 参数估计: 总体分布未知参数的估计
  - 假设检验和方差分析: 判别差异的显著性
  - 非参数统计：总体分布未知
  - 相关分析和列联表分析: 判断变量间是否相关
  - 主成分分析与因子分析: 降维
  - 聚类分析与判别分析: 分类
  - 回归分析与典型相关分析: 变量间的关系
- 统计软件：SPSS, R，SAS，Matlab，Excel
- 参考文献：姜启源等 数学模型（第五版）高等教育出版社 2018 第九章 统计模型

主成份分析

- 主成分分析(PCA)：对于原先提出的所有变量，建立尽可能少的新变量，并尽可能保持原有的信息。
- 主成分分析是通过降维技术用少数几个综合变量来代替原始多个变量的一种统计分析方法。

主成分分析的应用

- 削减回归分析的变量数目；
- 削减聚类分析的变量数目；
- 综合评分排序
  - 按第i主成分排序
  - 按总评分排序
    - 第i主成分=$a_i'*x$
    - a_i为特征向量，x为标准化向量
    - 总分= λ_1*第1主成分+ λ_2*第2主成分+…

聚类分析

- 聚类分析是根据“物以类聚”的原理，将本身没有类别的样本聚集成不同的组，这样的一组数据对象的集合叫作簇。
- 聚类分析也称无监督学习，聚类分析是研究如何在没有训练样本的条件下把样本划分为若干类。

两类聚类问题

1. 对样品的聚类: 统计指标是类与类之间的距离，它是把每一个样品看成高维空间中的一个点，类与类之间用某种原则规定它们的距离，将距离近的点聚合成一类，距离远的点聚合成另一类。
2. 对变量的聚类: 统计指标是变量间相似系数，根据这个统计指标将比较相似的变量归为一类，而把不怎么相似的变量归为另一类。

样品的聚类: 类间的距离

- 令G_p和G_q中分别有p和q个样品，它们的重心分别记为$\bar{x}_p$和$\bar{x}_q$。

- 最短距离：$D(p,q)=\min\left\{d_{jk}\middle| j\in G_p, k\in G_q\right\}$

  适合不规则类（条形，S形...）

- 最长距离：$D(p,q)=\max\left\{d_{jk}\middle| j\in G_p, k\in G_q\right\}$

  适合直径粗略相等的类

- 重心距离：$D(p,q)=(\bar{x}_p-\bar{x}_q)'(\bar{x}_p-\bar{x}_q)$

  适合处理异常值

- 类平均距离：$D(p,q)=\frac{1}{pq}\sum_{i\in G_p}\sum_{j\in G_q}d_{ij}$

  - 适合合并具有较小偏差的类

变量的聚类: 相似系数

- 设有n组样品，每组样品有m个变量，第i样品第k变量数据为x_{ik}，
- 夹角余弦

$$c_{ij} = \frac{\sum_{k=1}^{n} x_{ik} x_{jk}}{\left(\sum_{k=1}^{n} x_{ik}^{2} \sum_{k=1}^{n} x_{jk}^{2} \right)^{1/2}}$$

- 相关系数

$$r_{ij} = \frac{\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\left[\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{n} (x_{jk} - \bar{x}_j)^2 \right]^{\frac{1}{2}}}$$

$\bar{x}_i$ 表示第个指标的平均值。

系统聚类法

- 系统聚类法（Hierarchical clustering method）是目前使用最多的一种方法。
- 基本思想是首先将*n*个样品看成*n*类，然后规定样品之间的距离和类与类之间的距离。
- 将距离最近的两类合并为一个新类，再计算新类和其他类之间的距离，从中找出最近的两类合并，继续下去，最后所有的样品全在一类。将上述并类过程画成聚类图，便可以决定分多少类，每类各有什么样品。

k-means聚类法

- k-means聚类法也被称为k-平均或k-均值算法，是一种在数据挖掘中得到广泛使用的算法。
- 主要思想是通过迭代过程把数据集划分为不同的类别，使得评价聚类性能的准则函数（误差平方和准则函数）达到最优，从而使得生成的每个聚类内紧凑，类间独立。

k-means算法步骤

- 1.输入聚类的数目k和样本数据；
- 2.确定k个初始聚类中心；
- 3.将样本集中的样本按照最小距离原则分配到最邻近聚类；
- 4.使用每个聚类中的样本均值作为新的聚类中心；
- 5.重复步骤3，4直到聚类中心不再变化；
- 6.结束，得到k个聚类结果。

k-means聚类的优缺点

优点：

- 快速高效

缺点：

- 1.对符号属性的数据不适用；
- 2.必须事先给定k；
- 3.对初值敏感，对不同的初值，可能得到不同的结果；
- 4.对噪声和孤立数据敏感。

数学建模竞赛中的聚类问题

- 2012A 葡萄酒评价
- 2013D 公共自行车服务系统
- 2017B “拍照赚钱”的任务定价

回归分析

回归分析(Regression Analysis) 是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。

回归模型种类

- 多元线性回归：最常见，多个自变量；
- 多项式回归：引入高次项、交叉项；
- 0-1回归：含有分组自变量；
- 非线性回归：回归函数非线性；
- 引入自相关的回归模型：时间序列；
- 逐步回归：剔除无关变量；
- Logistic回归：因变量y为逻辑变量0-1。

多元线性回归的概念

- 模型： $y = b_0+b_1x_1+b_2x_2+\ldots+b_px_p+\varepsilon, \quad \varepsilon \sim N(0,\sigma^2)$
- 参数估计：最小二乘法

$$\hat{\beta} = (X'X)^{-1}X'Y, \qquad \hat{\sigma}^2 = MSE = \frac{SSE}{n-p}$$

- 模型显著性检验：R^2，F值，p值
- 参数显著性检验：参数b_i值置信区间是否包含零点？
- 模型诊断(残差分析)： ε是否正态？是否有规律可循？是否自相关？是否有重要的变量遗漏？

数学建模竞赛中的回归分析

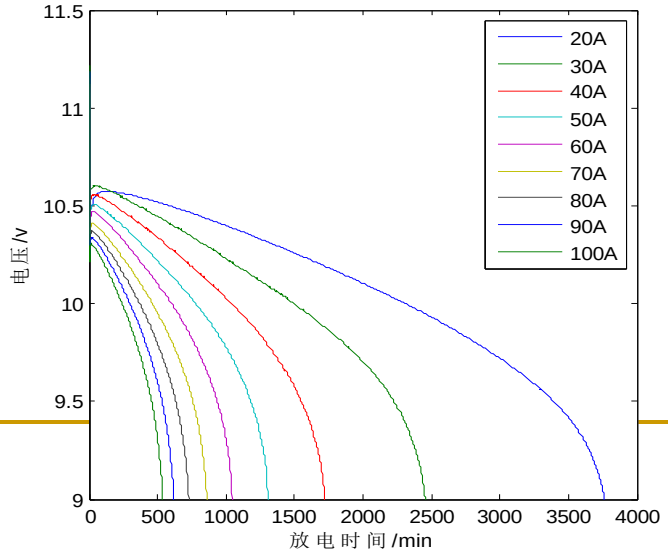
- 2005A 长江水质评价与预测
- 2006B 艾滋病疗法的评价及疗效的预测
- 2012A 葡萄酒的评价
- 2016C 电池剩余放电时间的预测

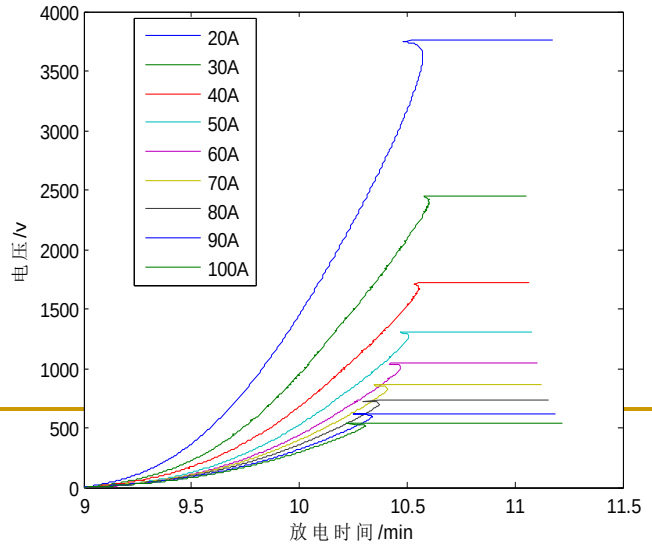
竞赛题1：电池剩余放电时间的预测

- 赛题介绍：C2016

数据处理

- **读取数据——利用MATLAB命令 x=xlsread(filename, sheet, range)**
- **例： x=xlsread('c:\MATLAB7\work\CUMCM2016-C-Appendix-Chinese.xls','附件1','a3:j1885');**
- **画出电池的放电曲线图和剩余放电时间曲线图**





竞赛题1：电池剩余放电时间的预测

建立模型

- 根据题意——建立剩余放电时间与电压和电流强度的整体表达式

$$t = f(U, I)$$

- 从图形看，并注意到当U=9v时，剩余放电时间为0，所以利用幂函数作拟合函数比较适合

$$t = c(U-9)^{a} I^{b}$$

其中a,b,c是待定常数。

竞赛题1：电池剩余放电时间的预测

模型求解

- 这是一个非线性最小二乘问题，利用MATLAB中nlinfit函数求解

```
clear
x=xlsread('c:\MATLAB7\work\CUMCM2016-C-Appendix-Chinese.xls','附件1','a3:j1885');
x(find(isnan(x)==1))=9;
y=[3764-x(:,1);2454-x(:,1);1724-x(:,1);1308-x(:,1);1044-x(:,1);862-x(:,1);730-x(:,1);620-x(:,1);538-x(:,1)];
y(find(y<0))=0;
beta0=[1 0 10];
dl=20:10:100;
x1=[x(:,2),dl(1)*ones(size(x,1),1)];
for i=2:length(dl)
  x1=[x1;x(:,i+1),dl(i)*ones(size(x,1),1)];
end
[beta,R,J]=nlinfit(x1,y,'dcsj1',beta0);
```

```
function yhat=dcsj1(beta,x)
yhat=beta(3)*(x(:,1)-9).^beta(1).*x(:,2).^beta(2);
```

- 求解结果：a=1.9347，b=-1.0473，c=33610

竞赛题1：电池剩余放电时间的预测

模型检验

- **MRE（平均相对误差）检验**

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{t_i - \hat{t}_i}{t_i}\right|$$

其中t_i是放电样本，$\hat{t}_i$ 是t_i的估计值

根据题意，在表中从**Um（Um=9v)**开始按不超过**0.005V**的最大间隔提取**231**个电压样本点。这些电压值对应的模型已放电时间与采样已放电时间的平均相对误差即为**MRE**

- 计算结果

电流/A	MRE/%	电流/A	MRE/%
20	0.3354	70	1.0031
30	0.7761	80	3.4490
40	1.3174	90	5.7217
50	2.6052	100	12.4465
60	2.1840		

竞赛题1：电池剩余放电时间的预测

模型应用

- 认为该模型通过检验，就可以利用该模型计算任意电压和电流强度下剩余放电时间的估计值
- 下表给出了在电压**9.8v**和不同电流强度下电池剩余时间的估计值和实际值

电流/A	实际值/min	估计值/min
30	594	619.53
40	430	458.37
50	326	362.85
60	277	299.78
70	254	255.08

- 下表给出了电流强度为**55A**的情况下，电池剩余放电时间的估计值

电压/v	10.3	10	9.8	9.5	9.0
剩余放电时间/min	840	506	328	132	0

竞赛题1：电池剩余放电时间的预测

- **问题3**

  问题3是考虑同一电池在不同衰减状态下以同一电流强度下的剩余放电时间。目标是补齐衰减状态3下的缺省数据

思路

1. 利用衰减状态3现有的数据进行拟合（幂函数，多项式），再利用拟合函数求得缺省数据，由于是外推，效果不一定好，可以通过在其它状态下检验来说明。
2. 将衰减状态3看成是另外三种状态的组合，利用回归的方法求解。

模型建立

- 线性回归模型

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

- 其中x1表示新电池状态，x2表示状态1，x3表示状态2，y表示状态3，ε 是残差，且满足$\varepsilon \sim N(0, \sigma^2)$

竞赛题1：电池剩余放电时间的预测

- 线性回归的MATLAB程序（其中数据用的是4个状态的前半段数据）

```
clear
x=xlsread('c:\MATLAB7\work\CUMCM2016-C-Appendix-Chinese.xls','附件2','a3:e150');
x1=x(:,1);
y=x(:,5);
x2=[ones(size(x,1),1),x(:,2:4)];
[b,bint,r,rint,stats]=regress(y,x2);%线性回归
b
bint
stats
```

- 运行结果

参数	估计值	置信区间
β_0	-0.6880	-1.9438 0.5679
β_1	-0.0928	-0.2978 0.1122
β_2	0.5660	0.3444 0.7877
β_3	0.2953	0.0454 0.5452

- 前面两个参数的置信区间都包含零点，说明这两个参数均不显著。

竞赛题1：电池剩余放电时间的预测

- 多项式回归+逐步回归

  增加二次项，然后通过逐步回归的方法，最终确定回归方程。

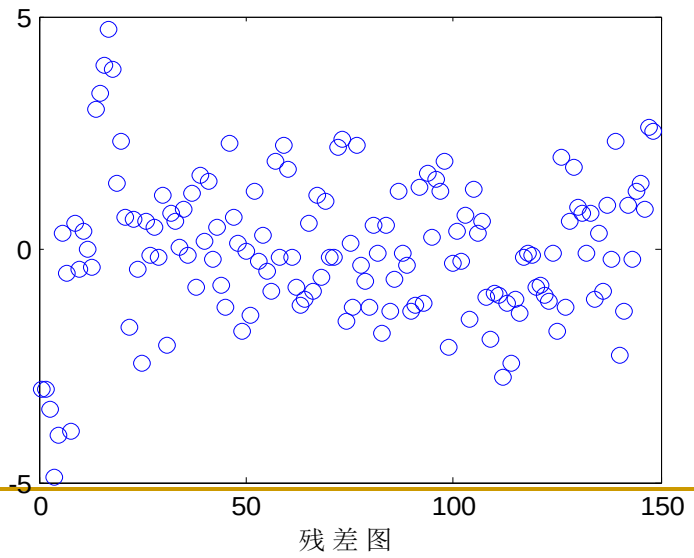
- 逐步回归MATLAB程序

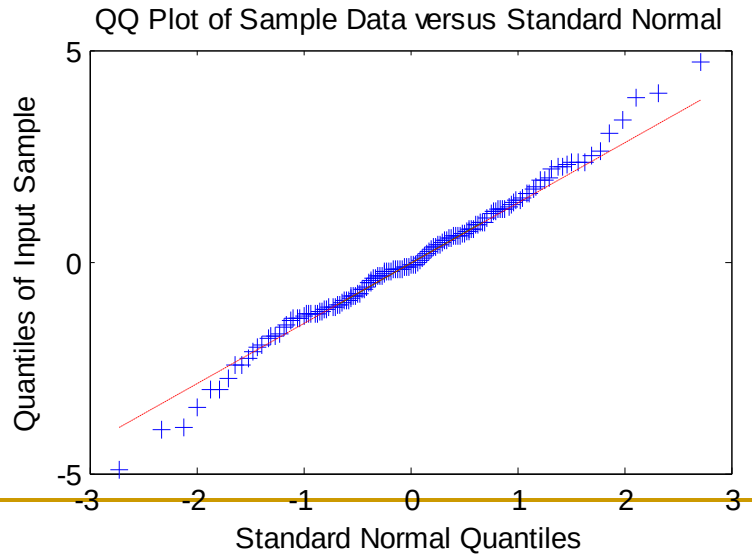
```
clear
x=xlsread('c:\MATLAB7\work\CUMCM2016-C-Appendix-Chinese.xls','附件2','a3:e150');
x1=x(:,1);
y=x(:,5);
x3=[x(:,2:4),x(:,2).^2,x(:,3).^2,x(:,4).^2];
stepwise(x3,y);%逐步回归
```

竞赛题1：电池剩余放电时间的预测

- 最终确定的回归模型不包含x1新电池状态，我们可以利用残差图和QQ图对该模型质量进行分析

```
x4=[ones(size(x,1),1),x(:,2:3),x(:,2).^2,x(:,3).^2];
[b,bint,r,rint,stats]=regress(y,x4);
figure(1);plot(r,'o');%画残差图
figure(2);qqplot(r);%画残差Q-Q图，可检验残差是否服从正态分布
```

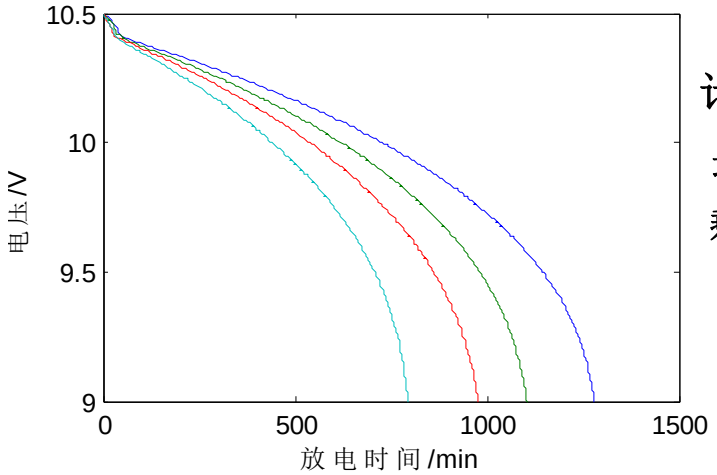




竞赛题1：电池剩余放电时间的预测

- 在完成检验后，就可以利用**x2衰减状态1**和**x3衰减状态2**的后半段数据来预测状态3的后半段数据

```
x5=xlsread('c:\MATLAB7\work\CUMCM2016-C-Appendix-Chinese.xls','附件2','a151 d303');
y=b(1)+b(2)*x5(:,2)+b(3)*x5(:,3)+b(4)*x5(:,2).^2+b(5)*x5(:,3).^2;
x5=[x5,y];
x6=[x;x5];
plot(x6(:,2:5),x6(:,1))
```



计算结果：

最终估计出来的总放电时间为：**805.73min**

剩余放电时间：**209.53min**

竞赛题2： 葡萄酒评价

- 每个评酒员在对葡萄酒进行品尝后对其分类指标打分，然后求和得到其总分，从而确定葡萄酒的质量。
- 酿酒葡萄的好坏与所酿葡萄酒的质量有直接的关系，葡萄酒和酿酒葡萄检测的理化指标会在一定程度上反映葡萄酒和葡萄的质量。
- 附件1给出了某一年份一些葡萄酒的评价结果，附件2和附件3分别给出了该年份这些葡萄酒的和酿酒葡萄的成分数据。

竞赛题2：葡萄酒评价

- 请尝试建立数学模型讨论下列问题：

1. 分析附件1中两组评酒员的评价结果有无显著性差异，哪一组结果更可信？
2. 根据酿酒葡萄的理化指标和葡萄酒的质量对这些酿酒葡萄进行分级。
3. 分析酿酒葡萄与葡萄酒的理化指标之间的联系。
4. 分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，并论证能否用葡萄和葡萄酒的理化指标来评价葡萄酒的质量？

竞赛题2： 葡萄酒评价

1. 分析附件1中两组评酒员的评价结果有无显著性差异，哪一组结果更可信？

- 数据预处理：检查数据，处理异常数据和缺失数据。针对红葡萄酒样品20评酒员4号对色调的评分缺失，利用同组评酒员对红葡萄酒样品20色调评分的平均值作为4号评酒员的评分值。
- 双因素（组别，酒品）方差分析：检验组别的显著性。结果：两组评酒员的评价结果存在着显著性差异（置信水平为95%）。
- 可信度指标=组内方差/品种方差。即认为同组品酒员之间差异尽可能小，且酒品之间差异尽可能大。结果：第二组评酒员的可信度较高。

竞赛题2：葡萄酒评价

2. 根据酿酒葡萄的理化指标和葡萄酒的质量对这些酿酒葡萄进行分级。

- 聚类分析：按照葡萄的理化指标xi聚类（可以先用主成分分析降维），相近的放在一类。然后计算各小类的酒质量平均分，分出等级。

A	葡萄样品	3	21	2	9	23		平均分
	综合评价指标	10.074	9.669	10.201	10.138	10.716		10.16
B	葡萄样品	13	19	4	16	27	22	
	综合评价指标	9.395	9.753	8.45	9.348	9.135	9.529	
		17	24	5	20	26		平均分
		9.901	9.706	9.071	9.817	9.139		9.38
C	葡萄样品	25	8	14	11	10		平均分
	综合评价指标	8.571	9.003	9.204	8.662	9.204		8.93
D	葡萄样品	12	18	6	7	15	1	平均分
	综合评价指标	6.984	7.623	8.985	8.897	7.309	7.79	7.93

酿酒葡萄（红）的等级划分

竞赛题2：葡萄酒评价

3. 分析酿酒葡萄与葡萄酒的理化指标之间的联系。

- 葡萄的理化指标xi，酒的理化指标yj.
- 典型相关分析：对每组变量xi和yj分别构造线性组合，将两组变量之间的相关性转化为两个变量之间的相关性进行研究。
- 典型相关系数的显著性检验，判断两个典型变量之间相关性是否显著。

竞赛题2：葡萄酒评价

4．分析酿酒葡萄和葡萄酒的理化指标对葡萄酒质量的影响，并论证能否用葡萄和葡萄酒的理化指标来评价葡萄酒的质量？

- 逐步回归分析：将葡萄酒的质量评分z作为因变量，葡萄的理化指标xi和酒的理化指标yj都作为自变量. 用此回归方程来评价葡萄酒的质量。由于指标比较多，可以采用逐步回归的方法对指标进行筛选。可以发现，只有少数理化指标对评价酒的质量是显著有价值的。

祝各位在2018年竞赛中取得好成绩

谢谢！