文章编号: 1006- 9348(2009) 02- 0134- 05

一种基于新型数据预测方法的 MICA仿真研究

白丽娜[1,2], 高 翔[2], 苑明哲[1], 曹景兴[3]
(1. 中国科学院沈阳自动化研究所, 辽宁 沈阳 110016;
2. 沈阳化工学院信息工程学院, 辽宁 沈阳 110142;
3. 沈阳新松机器人自动化股份有限公司, 辽宁 沈阳 110168)

摘要: 在多元统计过程监控中, 为解决因未知过程数据统计分布而产生误报漏报的现象, 提出一种结合多向独立元分析法(MICA)和广义相关系数(GCC)数据预测的综合方法, 进行在线监控过程的仿真。MICA分析方法能有效分解各变量的关联关系, 且不需考虑建模数据是否符合正态分布, 用此方法计算的独立元变量能更好地描述过程的变化规律。为提高预报未来过程故障的能力, 提出用广义相关系数法进行数据预测: 确定与运行轨迹相似的监控模型库中的轨迹, 并使其相应部分承接于运行轨迹之后。现场采集聚氯乙烯聚合过程的数据进行仿真, 仿真结果显示: 对于在线监控和在线故障诊断方面, 这种新型预测方法优于其它传统处理预测问题的方法。

关键词: 广义相关系数法; 间歇过程; 多向独立元分析法; 故障诊断
中图分类号: TP 277 文献标识码: A

Simulation of MICA Based on a New Type of Data Prediction

BAI Li- na[1,2], GAO Xiang[2], YUAN Ming- zhe[1], CAO Jing- xing[3]
(1. Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang Liaoning 110016, China;
2. The Information and Engineering School, Shenyang Institute of Chemical Technology, Shenyang Liaoning 110142, China;
3. SiaSun Robot& Automation CO., LTD, Shenyang Liaoning 110168, China)

ABSTRACT: In Multivariate Statistical process monitor, an approach of combining Multiway Independent Component Analysis with Generalized Correlation Coefficients (GCC) is presented to deal with the unknown statistic distribution of the data for overcoming the phenomenon of improper diagnosis and to carry out the simulation of online monitoring process. MICA could separate the correlation of variables without considering whether the model data follow the normal distribution or not, and the independent component variables are able to describe the variation of the process well. Furthermore, to improve the ability of predicting future process fault, GCC method is also presented for predicting unknown future data to ascertain the proper corresponding multi- trajectories in history model library and copy the part from the current time point as supplement for the running trajectories being tested. The data from polymerization process of polyvinyl chloride are sampled for simulation of online process monitoring, and the results show that the new prediction method is more effective than other traditional ones for online process monitoring and fault diagnosis.

KEYWORDS: Generalized correlation coefficients(GCC); Batch process; Multiway independent component analysis (MICA); Fault diagnosis

1 引言

间歇过程具有变量繁多, 变量之间相互关联, 过程变化剧烈且存在有限操作周期等特点。多元统计分析是间歇过程进行监控和故障诊断的一种有效方法。Nomikos & MacGregor于 1995年曾经提出用多向主元分析法(MPCA)解决间歇过程的离线和在线监控的问题[1], 这种方法随后得到了很大发展[2-4]。然而, 许多实际生产过程的数据并不满足MPCA所设想的正态分布, 此时运用 MPCA方法是必产生误报和漏报。MICA方法由于将原始变量变换为独立元变量,

基金项目: 国家高技术研究发展计划(863计划)(2006AA04Z185), 辽宁省教育厅项目(2005320)
收稿日期: 2007- 10- 10 修回日期: 2008- 01- 12

不需考虑数据是否符合正态分布,减少了误差的产生[5]。

间歇过程中,开发有效的在线监控方法,就能预知过程的异常并及时解决以保证产品质量。而在线监控方法的实施不可避免的问题是对现有的过程数据进行数据预测。Nomikos & MacGregor提出三种在线预测数据的方案[1]。第一种方法是轨迹经过归一化处理后(其均值为 0),将轨迹的未知部分填充为全 0。第二种方法是用最近的当前时刻值,将轨迹的后续部分补成最近值的直线族。这两种方法忽视了过程的变动特征。第三种方法是按轨迹的长度决定模型中需要投影的部分的大小,这样做的缺点是忽视了整体模型的作用,容易造成漏报。

本文提出一种新型预测数据的方法:广义相关系数法是表征多元轨迹矩阵之间相似关系的一种运算。在监控模型库中,计算正在运行中的多元轨迹与监控模型库中各条轨迹相应部分的相似特征,取广义相关系数最高的某批次数据的监控点的后半部接于正在运行中的多元轨迹,并用于 MICA 仿真。与三种传统方法相比,广义相关系数法更接近于实际过程,仿真的控制指标和故障诊断较为准确。

在间歇过程在线监控的仿真中,以监控时间点为基准,用广义相关系数法对监控时间点后的未知数据进行预测,得到完整的数据批次;将此完整批次投影到已建立好的 MICA 建立的模型中,计算出 SPE统计量,判断其是否超出控制限,如果超限,则说明此批次存在异常;最后,根据变量对 SPE的贡献大小,作出直观的贡献图,初步诊断哪个变量发生故障。

2 独立元分析和多向独立元分析

2.1 独立元分析(ICA)

设用来建模的数据矩阵 x 经过独立元分解有

$$x = As + E \tag{1}$$

其中,$x \in R^{d\times n}$为数据矩阵,n为采样时间,d为观测变量个数,$A \in R^{d\times m}$为混合矩阵,$s \in R^{m\times n}$为独立元矩阵,m为独立元个数,且 $d \geqslant m$,$E \in R^{d\times n}$为残差矩阵。独立元矩阵 s的估计方法是先估计出信源分离矩阵 W,然后再按式(2)估计出独立元矩阵 $\hat{s}$。

$$\hat{s} = Wx \tag{2}$$

ICA 的第一步是漂白化,将采样时间为 k 的矩阵 x 的协方差为:$R_x = E(x(k)x^T(k))$,按以下公式分解:

$$R_x = U\Lambda U^T \tag{3}$$

漂白变换后有:

$$z(k) = Qx(k) = QAs(k) = Bs(k) \tag{4}$$

其中,Q 可以由 $Q = \Lambda^{-1/2}U^T$ 求得,B 为正交矩阵。问题由寻求一个任意的满秩矩阵 A 化简为寻找一个正交矩阵 B。可以按照式(5)估计出 $s(k)$。

$$\hat{s}(k) = B^T z(k) = B^T Q x(k) \tag{5}$$

从式(2)和(5)可以得到 W 和 B 的关系:

$$W = B^T Q \tag{6}$$

可采取负熵的定义估计 B,使得第 i 个独立元为 $\hat{s}_i = (b_i)^T z$,并满足非高斯分布。由于熵在随机变量中是不确定性的平均度量,负熵是典型的非高斯分布的度量,熵在随机变量中是不确定性的平均度量,为了获得一个对于高斯分布为零的非高斯分布的度量,负熵 J 定义如下:

$$J(y) = H(y_{gauss}) - H(y) \tag{7}$$

其中,y为随机变量,$f(y)$为密度函数。y_{gauss}是与 y 方差相同的高斯随机变量。而用式(7)要求对密度函数的概率进行估计。为有效地估计负熵,Hyvarinen 提出一种简单近似的方法:

$$J(y) \approx [E\{G(y)\} - E\{G(v)\}]^2 \tag{8}$$

其中,y假设其均值为零,方差为单位方差,v 是具有零均值单位方差的高斯变量,G 是任意非二次方的函数。通过广泛的选择 G 可以获得很好的负熵近似值。Hyvarinen证明函数 $G(u) = \frac{1}{a_1}\log\cosh(a_1 u)$,$(1 \leqslant a_1 \leqslant 2)$对一般目的有着较好的效果。基于负熵的近似式,Hyvarinen 介绍了如何迭代地计算矩阵 B[5]。

计算过 B 之后,分别从式(5)、(6)可以得出 $\hat{s}(k)$和分离矩阵 W。

2.2 ICA 排序及独立元个数的选择

计算出独立元矩阵 s后,影响监控效果的重要一步是对独立元矩阵进行排序即对信源分离矩阵 W 的行向量 w_i 进行排序,根据欧几里得范数 (L_2) $\arg\max \|w_i\|_2$ 的原理,按其范数的大小进行排序。独立元矩阵 s 的顺序与信源分离矩阵 W 一致。ICA 独立元个数 m 的选择用于监控中的独立元个数必须适中,太多将会扩大噪声,太少的独立元则不能显示出过程的主要特性,近而导致监控性能差。独立元个数的选择是保留排序后 W 的前 m 行,并基于这样的假设:拥有最大欧式范数 W 的行向量对独立元矩阵的变化最显著。可根据 W 的每行的范数的累积百分比决定[7],叫做累积范数贡献率。

2.3 多向独立元分析(MICA)

间歇过程的数据以三维数组 $\underline{X}(I\times J\times K)$的形式堆放在一起建模:第一维为批次 I,第二维为变量 J,第三维为时间 K。按 Nomikos& Macgregor的展开方式:在每一时刻,垂直于时间轴,将三维数据切成多个批次 × 变量的数据片,并按时间顺序从左至右依次展开,形成一个超宽的矩阵 X $(I\times JK)$。对此矩阵运用 ICA 方法,且依照 ICA 的算法将矩阵转置为 X $(JK\times I)$。将数据矩阵 X $(JK\times I)$归一化后,分解为独立元向量 s、负荷矩阵 A 和残差矩阵 $\underline{E}$的乘积:

$$\underline{X} = \sum_{r=1}^{m} s_r \otimes A_r + \underline{E} \text{ or } X = \sum_{r=1}^{m} s_r a_r^T + E = X^* + E \tag{9}$$

其中,$\otimes$表示 Kronecker乘积($\underline{X} = s\otimes A$ 是 $\underline{X}(i,j,k) = s(i)A(j,k)$),$m$ 表示保留的独立元个数。

建模完成后,为进行监控需要计算两种统计量:对于某个批次 i的 I^2 统计量涉及过程变量的系统部分,定义如下:

$$I^2(i) = \hat{s}(i)^T \hat{s}(i) \tag{10}$$

其中, $\hat{s}(i)$为 $\hat{s}$的第 i个列向量。

对于某一批次 i的 SPE统计量表示过程变量残差部分, 定义如下:

$$SPE(i) = e(i)^T e(i) = (x(i) - \hat{x}(i))^T (x(i) - \hat{x}(i)) \quad (11)$$

其中, $\hat{x}$可以用下列公式计算:

$$\hat{x}(i) = A\hat{s}(i) \quad (12)$$

在 MICA的监控中, 独立元并不服从象正态分布那样一个特殊的分布, 所以 I^2和 SPE统计量的控制限可以通过核密度估计 (KDE)来求取[5]。

3 基于广义相关系数的在线监控

3.1 一种广义相关系数法

多变量的矩阵方差和协方差的计算结果都是矩阵, 无法如同传统的两随机变量的相关系数的计算直接以协方差 (一个有理数)除以两个方差的乘积的平方根 (有理数)的方式表达。矩阵特征根之和可以粗略描述矩阵的总体特征, 因此提出用它们的方差或协方差矩阵的迹来求广义相关系数。

设一个待测多元轨迹 $a(n \times d)$。n为采样时间点, d为变量的个数。另一来自模型库 Ω中的历史批次的多元轨迹可表示为 $b(n \times d)$。它们之间的广义相关系数为:

$$\rho(a, b) = \frac{tr(R_{ab})}{\sqrt{tr(R_a)\, tr(R_b)}} \quad (13)$$

其中 R是计算方差和协方差的函数。$R_a = E(a(k)a^T(k))$, R_a的求法同理; $R_{ab} = E(a(k)b^T(k))$。

3.2 模型参考轨迹弥补预测的在线监控

在线运行时的轨迹并没有将未来时间的部分展现出来, 不能直接应用 MICA法, 不可避免的问题是进行数据预测, 当预测出完整的过程批次数据后再使用 MICA方法进行监控。广义相关系数法提供了一种当前轨迹和监控模型轨迹的比较方式。它基于这样的假设: 如果通过计算, 运行轨迹的现有部分和对应模型库中的某条轨迹监控点前的广义相关系数最高, 就可以认为它们的相应部分最相似, 那么其后的部分相似的程度也最高[8]。将从监控模型库中选出的广义相关系数最高的轨迹相应部分补充到现有的运行轨迹后, 完成数据的预测。监控模型库 Ω中的历史批次应尽量包含各种正常和异常的批次, 以便可以正确地选择出相似轨迹。

4 仿真实例

仿真采用聚氯乙烯 (PVC)实际生产过程的数据。过程采样频率为每 5秒种采集一次, 共 3200个采样点, 持续约 4.5小时。依次选取温度设定值、釜内温度、夹套入口温度、进水温度、挡板出口温度、夹套出口温度、釜内压力、挡板流量、夹套流量和搅拌功率等 10个观测变量用于多元统计模型计算。系统采用良好的 50个批次按照 2.3节所述建立 MICA模型。取完整的 10个测试批次 (含正常批次和故障批次), 投影到 MICA模型上, 以进行离线监控。其它在线监控方法的结果与离线监控方法的进行比较, 与离线监控结果越近的说明此在线监控方法越优。监控方法使用 SPE统计量和 I^2统计量, 如 2.3节所述。

4.1 各种在线方法与离线方法的统计量的比较

在线监控中需设置监控时间点, 由于 PVC在早期和中期的动态过程变化最为剧烈, 对于数据预测方法来说会存在很大误差, 所以在早期的第 500点和中期的 1800点设置监控点可以比较出各个在线预测数据方法性能的优良。分别用平均轨迹法, 最近值法和部分模型的方法和本文提出的广义相关系数的预测方法对 10个 PVC过程进行数据预测, 方法参考 3.1、3.2节所述。并用 MICA方法实施在线监控, 将其结果和离线方法进行比较。监控结果越接近离线监控结果的说明此数据预测方法越优越。选择 MICA监控方法的 6个独立元, 其累积范数贡献率为: 86.23%。其监控结果如图 1所示。

图 1中的 (a)、(b)为监控时间点设置在 500点的监控结果, 此时正处于过程反应变化比较剧烈的早期, 出现的数据量较少各种预测方法不容易正确预测未来情况。从图 1 (a)、(b)可以看出, 无论是 SPE统计量还是 I^2统计量基于广义相关系数预测数据的方法均能正确预测出未来数据的状态, 其监控结果与离线监控结果最为接近。而其它方法均产生不同程度的误报或漏报。随着反应时间的增加, 过程反应趋近于平缓, 但在反应的中期仍有所波动。图 1中的 (c)、(d)显示了在监控时间点为 1800点的监控仿真情况, 从图中可以观察到随着仿真时间的增加, 数据量的增大, 基于广义相关系数法的数据预测方法更接近离线监控结果。而其它方法仍有漏报现象, 如: 在线均值法对第 9批次的漏报 (如图 1(d)), 在线部分模型法对第 2批次的漏报 (如图 1(c)、(d))。

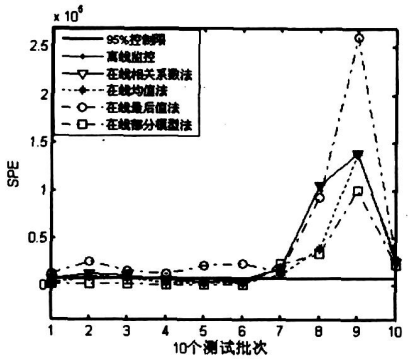
4.2 各种在线方法和离线方法故障诊断结果的比较

过程监控只是发现过程存在异常与否, 而人们希望找到问题的所在。贡献图 (contribution plot)是一种能根据变量对误差贡献的大小, 从图形中直观表达出各变量占 SPE或是 I^2的比例大小, 从而协助人们确定问题变量的初步方法, 在确定了一个或几个变量出现问题后, 可以综合其它的智能方法, 如因果图, 故障树等逆向反推求出故障原因。

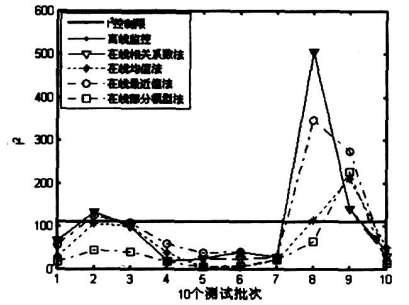
图 2的 (a)和 (b)分别给出了第 2批次的早期 500点时的 SPE和 I^2的贡献图。从图中可以看到, 两种统计方法即 SPE和 I^2统计量的广义相关系数与离线方法的贡献图最相似, 而其它方法未能正确找出故障变量, 有的甚至看不出哪一个变量贡献最大 (如图 2(a)的在线均值法贡献图)。与离线诊断结果相同, 基于广义相关系数的数据预测方法的诊断结果是 SPE检查出第 9变量夹套流量, 而 I^2检查出是第 4变量进水温度对故障的贡献较大。综合两种因素, 可以进一步寻找故障原因。

4.3 在线过程监控的实时性

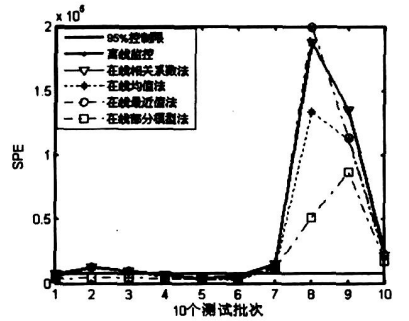
从过程采样到统计量的计算结果, 需要消耗一定的时



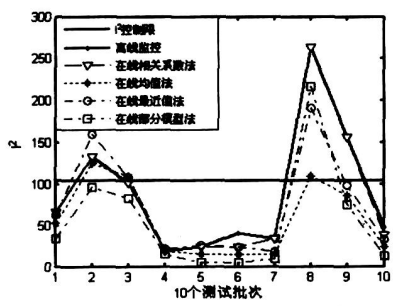
(a)500点，SPE统计量



(b)500点，T^2统计量



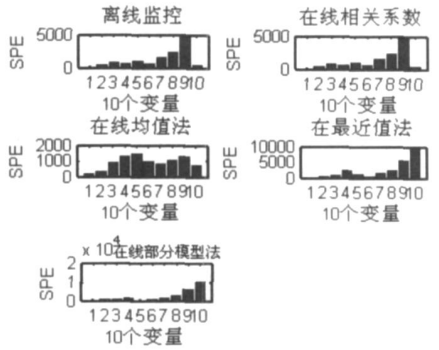
(c)1800点，SPE统计量



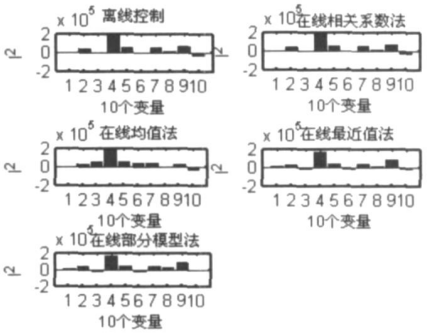
(d) 1800点，T^2统计量

图 **1** 各种数据预测方法监控结果

间，这部分的时间对实时性的影响需要考虑。仿真所用的语言是 **MATLAB**7.0，在 Pentium4 处理器，3.06GHz主频，内存



(a)SPE的贡献图



(b) T^2的贡献图

图 **2** 各种监控方法在过程早期两种统计量的贡献图

为 512M 的环境下运行。检测一个批次需耗时 33.4s，而采样频率为 5s/次。这意味着出现结果需延时 7拍左右，可以满足实时监控的要求，随着计算机的硬件设施的改善，时延将大大缩小。

5 结语

运用 MICA方法对 PVC间歇过程进行监控，可以通过建模后，新批次多元投影的方法计算 SPE和 T^2，有效地发现过程的异常；佐以贡献图的手段，得以迅速地查找出故障感染的变量，为最终判别故障的原因提供依据；而广义相关系数法能够在在线过程中，通过计算广义相关系数来判断轨迹的相似性以供替代。有效地预计了过程的未来，降低了计算的误差。但在实际运用时要注意模型库的选取应尽量庞大，以涵盖可能的变化，这样才能预测得准确。

参考文献：

[1] Paul Nomikos, John F MacGregor. Multivariate SPC Charts for Monitoring Batch Process[J]. Technometrics, 1995, 37(1): 41-59.

[2] Paul Nomikos, John F MacGregor. Muti-way partial least squares in monitoring batch processes[J]. Chem. Intell. Lab. Sys. 1995, 30: 97-108.

[3] N B Gallagher, B M Wise. Application of multi-way principal

component analysis to nuclear waste storage tank monitoring[J]. Comput Chem. Eng 1996, 20: 739 – 744

[4] T Kourti Multivariate dynamic data modeling for analysis and statistical process control of batch processes[J]. start– ups and grade transitions, J Chemom. 2003, 17: 93– 109.

[5] Chang Kyoo Yoo, Jong– Min Lee. On– line monitoring of batch processes using multiway independent compinent anslysis[J]. Chemometrics and Intelligent Laboratory Systems 2004, 71: 151 – 163

[6] 何宁. 基于 ICA – PCA 方法的流程工业过程监控与故障诊断研究[D]. 浙江大学博士学位论文, 2004.

[7] Jong Min Lee, ChangKoo Yoo, In Beum Lee. Statistical process monotoring with independent component analysis[J]. Journal of Process Control, 2004, 14: 427– 485

[8] 高翔, 白丽娜. 基于广义相关系数的多元轨迹预测及数据恢复方法[C]. 2007中国控制与决策学术年会论文集, 120 – 122

[作者简介]

白丽娜(1982–), 女(满族), 辽宁省沈阳市人, 在读硕士研究生, 从事多元统计过程控制, 故障诊断和模式识别等方面的研究。

高 翔(1967–), 男(汉族), 吉林市人, 副教授, 博士, 从事多元统计过程控制, 故障诊断和模式识别等方面的研究。

苑明哲(1971–), 男(汉族), 辽宁省抚顺市人, 副研究员, 中国计算机用户协会仿真应用分会理事, 主要研究方向为分布式控制系统和工业过程先进控制与优化。

曹景兴(1978–), 男(汉族), 辽宁省沈阳市人, 助理工程师, 主要研究方向为工业自动化。

(上接第 133页)

[2] J Ma, D Zhou. Fuzzy set approach to the assessment of student– centered learning [J]. IEEE Transactions on Education, 2000, 43 (2): 237– 241

[3] Y H Li, Y Q Hu. A Model of Multilevel Fuzzy Comprehensive Evaluation for Investment Risk of High and New Technology Project [C]. In 2006 International Conference on Machine Learning and Cybernetics, Dalian, 2006. 1942 – 1947.

[4] S A Farghal, M S Kandil, A Elmitwally. Quantifying Electric Power Quality via Fuzzy Modeling and Analytic Hierarchy Processingp [C]. IEE Proceedings– Generation, Transmission and Distribution, 2002, 149(1): 44– 49.

[5] 陈衍泰, 陈国宏, 李美娟. 综合评价方法分类及研究进展[J]. 管理科学学报, 2004, 7(2): 69– 79.

[6] D Ramot, R Milo, M Friedman. Complex fuzzy sets[J]. IEEE Transactions on Fuzzy Systems, 2002, 10(2): 171 – 186

[7] L X Wang. Analysis and design of hierarchical fuzzy systems [J]. IEEE Transactions on Fuzzy Systems, 1999, 7(5): 617 – 624

[8] 程乾生. 属性识别理论模型及其应用[J]. 北京大学学报, 1997, 33(1): 12– 20.

[9] T L Satty. The Analytic Hierarchy Process[M]. New York: McGraw– Hill Inc, 1988

[10] 秦寿康. 综合评价原理与应用[M]. 北京: 电子工业出版社, 2003.

[11] M H J Bollen, D D Sabin, R S Thallam. Voltage– sag Indices – Recent Developments in IEEE P1564 Task Force[C]. in Proc CIGRE/IEEE PES International Symposium on Quality and Security of Electric Power Delivery Systems, 2003. 34 – 41

[12] 陈伟, 郝晓弘, 林洁. 基于属性识别和 AHP的电能质量综合评价体系和方法[J]. 电气技术, 2006, 7(5): 26– 30

[作者简介]

陈 伟(1976–), 男(汉族), 甘肃甘谷人, 硕士, 讲师, 研究方向为电能质量分析与控制技术。

郝晓弘(1960–), 男(汉族), 甘肃泾川人, 硕士, 教授, 博士生导师, 研究方向为网络控制, 电能质量控制等。

林 洁(1977–), 女(汉族), 甘肃兰州人, 硕士生, 讲师, 研究方向为电能质量控制技术。

附表 **1** 某供电系统 **PCC**点实测数据统计表

样本序号	I_{11}	I_{12}	I_{13}	I_{14}	I_{15}	I_{16}	I_{17}	I_{18}	I_{19}	I_{110}	I_2	I_3
1	5	2	1.67	0.89	0	4	3	0	0	1.15	0.116	2.47
2	11	2	1.40	0.78	1	1	1	0	0	1.55	0.06	2.39
3	13.7	2	1.11	0.78	0	1	0	0	0	1.20	0.65	2.40
4	7.3	1.1	1.56	0.85	0	2	2	0	0	1.15	0.092	3.50
5	6	2	1.53	0.75	1	3	0	0	0	0.90	0.07	3.19
6	3	1	0.96	0.71	0	1	0	0	0	0.94	0.06	1.61
7	5	2	1.17	0.73	1	1	0	0	0	1.05	0.06	2.59
8	4	2	2.08	0.96	5	8	0	0	0	1.87	0.12	4.28