第13卷第1期
1991年 3月

重庆建筑工程学院学报
Journal of Chongqing Institute of Architecture and Engineering

Vol 13.No.1
March.1991

一种用于随机突变量的数据可靠性处理算法①

度 强　　段文泽

(重庆工业管理学院)　(重庆建筑工程学院)

摘 要 对于随机突变参量的数据可靠性处理，本文提出了一种新的基于自适应预报的数据处理器算法。该算法将被测参量的变化规律用一参数慢时变的时间序列模型描述，以被测参量变化趋势的自适应一步预报值及其95%置信限构成判据，对测量数据进行处理，剔除或抑制其中的不良数据。仿真实例表明本算法克服了现有大多数算法在处理随机突变数据时遇到的困难，效果明显。

关键词 数据处理，随机变量，时序分析，自适应预报

1 概 述

控制系统中最主要的信息之一就是测量数据，它们对系统的正常，可靠运行至关重要。通过各种方法确保这些数据的准确可靠，是设计控制系统时的一项基本工作，一般包括硬件和软件的措施。硬件措施主要是从提高系统的抗干扰能力着手，而软件措施则是通过一些可靠性处理算法，自动剔除或抑制不良数据。计算机的广泛应用，为采用软件方法进行实时，高效的可靠性处理带来极大便利，国内近年来也有一些这方面的文献介绍[1] [2] [3]。

一般控制系统的被测参量总可分为两大类：缓变量和随机突变量（随机性大，波动剧烈）。目前常见的数据处理算法，原理大多比较简单，其优点是编程，实现容易，实时性好，在处理缓变参量时能满足要求。但共同弱点是算法中判断不良数据的判据经验性和盲目性较大，因而，在处理随机突变量时难以把握适当的数量指标，使用效果不佳。为克服上述弱点，本文针对随机突变量的测量数据，提出了一种新的基于自适应预报的数据处理器算法。该算法的特点是用一参数慢时变的时间序列模型，描述被测参量的变化规律，以被测参量变化趋势的一步预报值，及其95%置信限构成判据，对测量数据进行判别和处理，剔除或抑制其中的不良数据；同时对时序模型的参数进行在线修正，以适应系统或环境特征的变化。由于该算法是建立在随机模型预报的基础上，自然已经考虑了被测参量随机性的影响，使得对其测量数据的处理不再是盲目或经验性的，因而，从原理上克服了现有大多数算法在处理随机突变数据时遇到的困难，效果较为明显。

2 算法简介

①本文1990年8月31日收到

2.1 时序模型的建立

设某随机突变量各时刻的状态值构成序列$\{y_t\}$，$t=\cdots,1,2,3,\cdots$．当对影响$\{y_t\}$某时刻数值的决定性因素了解甚少，且容易获得大量$\{y_t\}$的历史数据时，一般总可将$\{y_t\}$用时间序列模型 ARIMA(p, d, q)或季节性模型描述．由于这些模型在经零均值平稳化处理后都可表达为 ARMA(p, q)模型，且任何 ARMA(p, q)模型都可用阶数足够高的 AR(p)模型逼近到任意精度[3][6]，因此，本算法中将使用 AR(p)模型描述$\{y_t\}$的变化规律，这样可大大减少建模和计算工作量．对$\{y_t\}$建立时序模型的过程，主要体现为对其施行零均值平稳化处理和识别其 AR(p)模型阶 p 的过程．

为叙述方便，现设$\{y_t\}$服从周期为 S 的季节性时序模型，则建模时首先去除季节性因素影响，将$\{y_t\}$化为平稳序列．对$\{y_t\}$取差分运算，令

$$W_t^* = \triangle_S y_t = y_t - y_{t-S} \tag{1}$$

则通常序列$\{W_t^*\}$就已是平稳序列（这也可以通过考察$\{W_t^*\}$的自相关函数来判断）．$\{W_t^*\}$的样本均值

$$\hat{\mu} = \frac{1}{N-S}\sum_{t=S+1}^{N} W_t^* \tag{2}$$

式中 N 为样本容量．令

$$W_t = W_t^* - \hat{\mu} \qquad (t=s+1,\cdots,n) \tag{3}$$

则$\{W_t\}$就是零均值平稳序列．

零均值平稳序列$\{W_t\}$可用 AR(p)模型描述

$$W_t = \varphi_1 W_{t-1} + \varphi_2 W_{t-2} + \cdots + \varphi_p W_{t-p} + a_t \tag{4}$$

式中$\{a_t\}$为零均值白噪声，方差 $D[a_t]=\sigma_a^2$．p 称为模型的阶，可采用 AIC 准则、BIC 准则、F 检验等方法确定[5]．

2.2 自适应一步预报及其 95%置信限

本算法中判别不良数据的依据是被测量状态的一步预报值及其 95%置信限．预报方法采用了平稳线性最小方差预报，递推公式为

$$\begin{cases} \hat{W}_k(1) = \varphi_1 W_k + \varphi_2 W_{k-1} + \cdots + \varphi_p W_{k-p+1} \\ \hat{W}_k(2) = \varphi_1 \hat{W}_k(1) + \varphi_2 W_k + \cdots + \varphi_p W_{k-p+2} \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ \hat{W}_k(p) = \varphi_1 \hat{W}_k(p-1) + \varphi_2 \hat{W}_k(p-2) + \cdots + \varphi_{p-1}\hat{W}_k(1) + \varphi_p W_k \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ \hat{W}_k(l) = \varphi_1 \hat{W}_k(l-1) + \varphi_2 \hat{W}_k(l-2) + \cdots + \varphi_p \hat{W}_k(l-p) \end{cases} \qquad (l>p) \tag{5}$$

则有状态的一步预报公式

$$\hat{W}_k(1) = \varphi_1 W_k + \varphi_2 W_{k-1} + \cdots + \varphi_p W_{k-p+1} \tag{6}$$

可见$\hat{W}_{k(1)}$只依赖于$\{W_t\}$的 K 时刻以及它以前的一共 p 个时刻的值W_k，W_{k-1}，

…，W_{k-p+1}．式中φ_1，φ_2，…，φ_p 为 AR(p)模型的参数，其精确值无法知道，使用时只能用其估计值$\hat{\varphi}_1$，$\hat{\varphi}_2$，…，$\hat{\varphi}_p$代替．这时一步预报公式变为

$$\hat{W}_k(1)=\hat{\varphi}_1 W_k+\hat{\varphi}_2 W_{k-1}+\cdots+\hat{\varphi}_p W_{k-p+1} \tag{7}$$

本算法中参数估计采用了 Yule-Walker 矩估计方法，对于（4）所表达的 AR（P）模型，需估计的参数包括 $\varphi_1,\varphi_2,\cdots,\varphi_p,\sigma_a^2$ （σ_a^2为$\{a_t\}$的方差，在求取预报置信限时要用到）．其 Yule-Walker 矩估计所满足的方程组为：

$$\begin{bmatrix}\hat{\varphi}_1\\ \hat{\varphi}_2\\ \vdots\\ \hat{\varphi}_p\end{bmatrix}=\begin{bmatrix}1 & \hat{\rho}_1 & \hat{\rho}_2 & \cdots & \hat{\rho}_{p-1}\\ \hat{\rho}_1 & 1 & \hat{\rho}_1 & \cdots & \hat{\rho}_{p-2}\\ \vdots & \vdots & \vdots & \vdots & \vdots\\ \hat{\rho}_{p-1} & \hat{\rho}_{p-2} & \hat{\rho}_{p-3} & \cdots & 1\end{bmatrix}^{-1}\begin{bmatrix}\hat{\rho}_1\\ \hat{\rho}_2\\ \vdots\\ \hat{\rho}_p\end{bmatrix} \tag{8}$$

$$\hat{\sigma}_a^2=\hat{r}_0-\sum_{j=1}^{p}\hat{\varphi}_j\hat{r}_j \tag{9}$$

(8)，(9) 中$\hat{r}_j$,$\hat{\rho}_j$分别为$\{w_t\}$的样本自协方差函数和样本自相关函数，求取公式为

$$\hat{r}_j=\frac{1}{N-S}\sum_{t=s+1}^{N-j}W_tW_{t+j},\qquad (j>0) \tag{10}$$

$$\hat{\rho}_j=\hat{r}_j/\hat{r}_0$$

对于大多数控制系统，随着系统和环境特征的变化，其被测参量变化的规律性也会稍有改变．这反映在模型（4）中，主要体现为其参数$\varphi_1,\varphi_2,\varphi_3,\cdots,\varphi_p,\hat{\sigma}_a^2$的慢时变性．为提高预报的准确性，应使参数估计适应于这种变化．在本算法的自适应一步预报中，每次获得新的被测参量状态值，都将对估计参数进行更新．

置信限是预报可信度的数量指标，反映用预报值代替真值的准确程度．若置信区间分布得越狭窄，则用预报值代替真值的准确性就越高，亦即预报的准确性越高．对于（4）所示 AR（P）序列的平稳线性最小方差预报，用$\hat{W}_k(l)$替代 W_{k+l}的 95%置信上、下限为

$$\hat{W}_k(l)\pm 1.96\sigma_a\sqrt{1+G_1^2+\cdots+G_{l-1}^2} \tag{12}$$

式中，$G_1\sim G_{l-1}$为 Green 函数[5, 6]，由（12）式可见置信区间的分布仅与预报步长 l 有关．由于本算法正是要通过预报值去判别测量数据的真实性，因此要求尽可能提高预报的准确性，由（12）显见，可以利用一步预报及其 95%置信限．这时用$\hat{W}_k(1)$替代 W_{k+1}的 95%置信上、下限为

$$\hat{W}_k(1)\pm 1.96\sigma_a \tag{13}$$

值得注意的是，由（7）得到$\hat{W}_k(1)$是序列$\{W_t\}$中的状态一步预报值，而系统的测量数据是序列$\{y_t\}$中的状态观测值．因此，要利用预报值对测量数据的真实性进行判断，还必须将$\hat{W}_k(1)$转换为序列$\{y_t\}$中的状态一步预报值$\hat{y}_k(1)$．由(1)，(2)，(3)知

$$W_t=y_t-y_{t-s}-\hat{\mu} \tag{14}$$

从而

$$y_{k+1}=y_{k+1}-y_{k+1-s}+y_{k+1-s}-\hat{\mu}+\hat{\mu}$$
$$=y_{k+1-s}+\hat{\mu}+W_{k+1} \tag{15}$$

因此

$$\hat{y}_k(1)=y_{k+1-s}+\hat{\mu}+\hat{W}_k(1) \tag{16}$$

(16)式就是$\hat{W}_k(1)$与$\hat{y}_k(1)$之间的转换公式.

2.3 基于自适应预报的数据处理器算法

仍然考虑随机突变量各时刻状态值构成的序列$\{y_i\}$，设y_{k+1}为序列在$k+1$时刻的状态真值，y^*_{k+1}为其观测值，$\hat{y}_{k+1}$为经本算法处理后状态真值的估值（即工程上认可的状态值）．本算法首先通过$\{y_k, y_{k-1}, \cdots, y_1\}$对$y_{k+1}$实施自适应一步预报，得到$\hat{y}_k(1)$．然后就可以对$y^*_{k+1}$进行判断，看其是属于正常采样值还是不良数据，并分别按不同情况加以处理，最后确定出$\hat{y}_{k+1}$．

本算法中采用工程上惯常使用的不良数据处理准则，即设置一低门限L_{min}和一高门限L_{max}，当y^*_{k+1}在L_{min}之内时，认为y^*_{k+1}就是该时刻被测参量的状态值y_{k+1}；当y^*_{k+1}在L_{max}之外时，认为y^*_{k+1}是不良数据，将其剔除；当y^*_{k+1}介于L_{min}和L_{max}之间时，保留y^*_{k+1}一部分，即对其进行抑制．

以上处理原则的关键，在于L_{min}和L_{max}（即判据）的设置．本算法在此利用了$\hat{y}_k(1)$的95%置信区间，$[\hat{y}_k(1)-1.96\sigma_a,\ \hat{y}_k(1)+1.96\sigma_a]$，

令

$$L_{min}=1.96\hat{\sigma}_a \tag{17}$$

因为根据置信区间的定义：

$$P[|y_{k+1}-\hat{y}_k(1)|\leqslant 1.96\sigma_a]=95\% \tag{18}$$

故可认为当$|y^*_{k+1}-\hat{y}_k(1)|<L_{min}$时，$y^*_{k+1}$就是正常的被测参量状态值$y_{k+1}$．再令

$$L_{max}=4\hat{\sigma}_a \tag{19}$$

即当$|y^*_{k+1}-\hat{y}_k(1)|>L_{max}$时，认为$y^*_{k+1}$是不良数据，须予以剔除．这时$\hat{y}_{k+1}$取预报值$\hat{y}_k(1)$．

当$L_{min}<|y^*_{k+1}-\hat{y}_k(1)|<\dfrac{L_{min}+L_{max}}{2}$时，$y^*_{k+1}$作为状态真值的概率逐渐减小，应逐步加大$\hat{y}_k(1)$的加权成份．而当$\dfrac{L_{min}+L_{max}}{2}<|y^*_{k+1}-\hat{y}_k(1)|<L_{max}$时，应倒过来以考虑$\hat{y}_k(1)$为主、并逐步减小$y^*_{k+1}$的加权成份．于是，将权系数设计为

$$R=R_{max}-\frac{4R_{max}}{(L_{max}-L_{min})^2}\left[|y^*_{k+1}-\hat{y}_k(1)|-\frac{L_{min}+L_{max}}{2}\right]^2$$
$$=R_{max}\left(1-\frac{1}{1.0404\hat{\sigma}_a^2}\right)\left[|y^*_{k+1}-\hat{y}_k(1)|-2.98\hat{\sigma}_a\right]^2 \tag{20}$$

可见 R 为一变权系数，自变量是 $\left|y^{*}_{k+1}-\hat{y}_k(1)\right|$，其变化规律遵循一开口向下的抛物线，见图 1。

图中权系数的最大值为 R_{max}，其值应根据被测参量的特点以及预报误差的大致范围而定。例如可使 $\left|y^{*}_{k+1}-\hat{y}_k(1)\right|$ 中被保留下的部份（即 $R\left|y^{*}_{k+1}-\hat{y}_k(1)\right|$）不超过自适应一步预报的均方根误差。

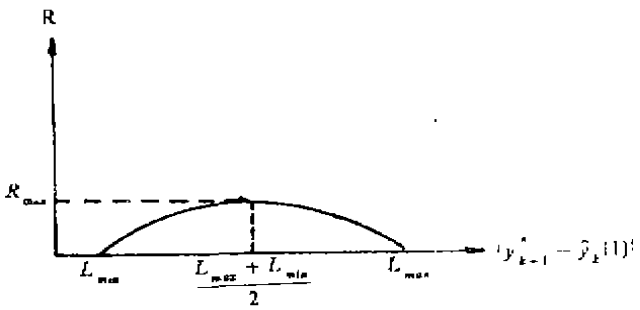

图 1　权系数 R 的变化规律

考虑到上述各点，可以将对观测值 y^{*}_{k+1} 的处理原则总结为以下公式

$$\hat{y}_{k+1}=\begin{cases} y^{*}_{k+1}, & \text{当}\left|y^{*}_{k+1}-\hat{y}_k(1)\right|\leqslant L_{min}\text{时}, \\ \hat{y}_k(1), & \text{当}\left|y^{*}_{k+1}-\hat{y}_k(1)\right|\geqslant L_{max}\text{时}, \\ y_{k+1}+Sgn\left(\hat{y}_k(1)-y^{*}_{k+1}\right)\cdot R\left|y^{*}_{k+1}-\hat{y}_k(1)\right|, & \\ \qquad \text{当}L_{min}<\left|y^{*}_{k+1}-\hat{y}_k(1)\right|\leqslant\dfrac{L_{min}+L_{max}}{2}\text{时}, & \\ \hat{y}_k(1)+Sgn\left(y^{*}_{k+1}-\hat{y}_k(1)\right)\cdot R\left|y^{*}_{k+1}-\hat{y}_k(1)\right|, & \\ \qquad \text{当}\dfrac{L_{min}+L_{max}}{2}<\left|y^{*}_{k+1}-\hat{y}_k(1)\right|<L_{max}\text{时} & \end{cases} \tag{22}$$

式中 R 按(20)式求取。

这样就得到了被测参量在 $k+1$ 时刻经过可靠性处理的状态值 $\hat{y}_{k+1}$，而 $\hat{y}_{k+1}$ 又将作为历史数据被用于 $k+2$ 时刻的状态预报，如此继续下去。为了体现预报和数据处理的自适应性，每次获得新的状态估值 $\hat{y}_{k+1}$，便对预报模型的估计参数进行更新。本算法中更新过程采用限定记忆方法，即保持样本长度不变，每增加一个新样本值就去掉最初的一个老样本值。

综上所述，可将本算法流程归纳如下

1) 赋 $N=k$，N 为样本长度，k 为起始时刻。

2) 利用(1)、(2)、(3)式将 $\{y_N, y_{N-1}, \cdots, y_1\}$ 零均值平稳化，得 $\{W_N, W_{N-1}, \cdots, W_{s+1}\}$。

3) 利用(8)、(9)、(10)、(11)式作模型(4)的参数估计得 $\hat{\varphi}_1, \hat{\varphi}_2, \cdots, \hat{\varphi}_p, \hat{\sigma}_a^2$。

4) 利用(7)式作一步预报得 $\hat{W}_N(1)$，通过(16)式得 $\hat{y}_N(1)$。

5) 由(17)、(19)式置 L_{min} 和 L_{max}，读入 y^{*}_{k+1}，通过式(22)对 y^{*}_{k+1} 进行处理得 $\hat{y}_{k+1}$。

6) 赋 $y_i=y_{i+1}$，$(i=1, 2, \cdots, N-1)$及 $y_N=\hat{y}_{k+1}$，$k=k+1$ 返回 2)。

需要说明的是，为叙述方便，以上在本算法的推导过程中，均假定被测参量状态变化规律所构成的序列$\{y_t\}$服从季节性时序模型．一般而论，$\{y_t\}$当然也可服从其它的时序模型，如 ARIMA (p, d, q)．这时算法流程中步骤 2)、4) 的零均值平稳化过程和 $\hat{W}_N(1)$ 与$\hat{y}_N(1)$ 的转换过程公式略有不同[5]，只须对程序稍作修改即可．

3 本算法的适用场合

本算法是一种在线实时数据处理算法，既适用于被测参量随机波动剧烈且对数据处理要求较高的工业控制系统，也适用于一般实验设备的数据实时处理．本算法以计算机作为计算工具，由于算法较为复杂，要求计算机的功能较强，因此特别适用于具有较强功能控制机的计算机集中控制系统，以及上位机功能较强(如 IBM PC—XT)的集散控制系统．

4 仿真结果及评价

本算法以某自来水厂递阶控制系统为对象进行了仿真研究．在水厂递阶控制系统中，送水泵站出口流量是一个随机波动剧烈（受用户用水量支配）且对控制算法的实现至关重要的参量，数据处理要求较高．又由于该系统本身需要在其上位机中对送水泵站出口流量未来若干时刻的状态作出在线实时预报，因此可以方便地将本算法加入到系统的控制软件中．

4.1 水量预报仿真研究

取某水厂提供的送水泵站实测出口流量十五天的数据（采样周期为 1 小时，共 360 个数据，$k=1, 2, \cdots, 360$）作为真实的状态值．取样本长度 $N=336$，预报起始时刻 $k=336$，预报 24 点．用自适应预报法与多层递阶预报法进行实验对比．表 1 给出了 1 步

表 1

时刻 k	k+1 时刻真值	自适应预报值	多层递阶预报值	时刻 k	k+1 时刻真值	自适应预报值	多层递阶预报值
336	122	131	126	348	154	157	158
337	150	120	118	349	140	141	141
338	125	143	150	350	131	148	157
339	189	178	178	351	125	144	139
340	103	111	111	352	160	149	150
341	104	131	118	353	122	179	176
342	116	142	151	354	165	153	162
343	160	148	152	355	160	130	132
344	135	128	130	356	139	129	128
345	117	121	130	357	129	140	130
346	123	152	153	358	128	108	106
347	117	125	132	359	156	180	168

预报值及相应的真值．表 2 给出了预报结果的统计指标．

表 2

指　标	均方根误差	平均相对误差	计算时间(秒)
自适应 1 步预报	20.67	12.70%	7
多层递阶 1 步预报	20.53	12.34%	14

比较两种预报方法，准确性相差无几，但自适应预报的计算时间却只有多层递阶预报的一半，因而更具实用性．平均相对误差虽然偏大，是因为水厂管理水平低，流量计量不准，造成数据序列平稳性很差之故．文[7] 采用同一多层递阶预报法，只因样本数据平稳性较好，误差就在 5%以内．随着管理水平，测量准确度的提高，预报准确度也必跟着提高．随着数据可靠性算法的采用，效果也将更好．

4.2 数据可靠性处理仿真

将前述 15 天的出口流量数据作为真实状态值，对其中 24 个状态值y_{k+1}($k+1=337$，338，…，360)，根据现场经验规定误差的最大范围，再利用计算机中的 RND 函数产生 6 个随机干扰迭加于其中，然后将此混有 6 个不良数据的序列值作为出口流量的观测值y^{*}_{k+1}($k+1=337$，338，…，360)．这批观测值经本算法处理后得到$\hat{y}_{k+1}$，再与y_{k+1}进行比较．此时本算法中样本长度 $N=336$，起始时刻 $k=336$，R_{max} 取 2／5，数据用 AIC 准则建模为 AR(24)．

图 2 为混有不良数据的流量观测值y^{*}_{k+1}与其状态真值y_{k+1}的比较，图 3 为经本算法处理后的状态真值的估值$\hat{y}_{k+1}$与y_{k+1}比较．y^{*}_{k+1}与y_{k+1}的平均相对误差为 14.75%，$\hat{y}_{k+1}$与y_{k+1}的平均相对误差为 2.53%．算法用编译 BASIC 编程，在 GW—0520DH 微机上实现，处理一个观测数据需机时 7 秒．

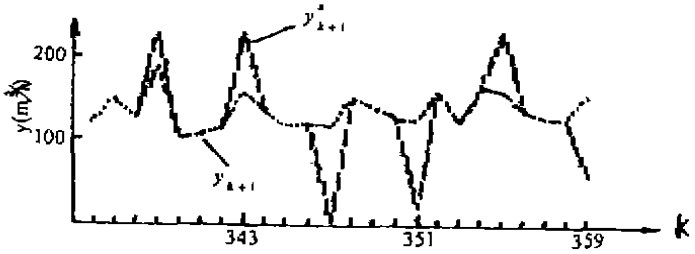

图 2　混有不良数据的观测值与状态真值比较

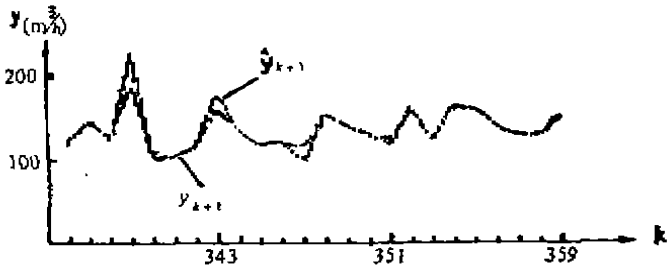

图 3　经本算法处理后的状态估值与真值比较

从以上仿真结果可见，本算法完全可以从随机突变参量的观测数据中剔除或抑制不良数据，使误差指标下降到令人满意的水平．从实时性方面看，对一般采样率不是太高的对象均可适用．此外，以上水厂提供的原始数据平稳性还比较差，随着递阶控制的实施，通过自适应算法的调节，将使出口流量数

据序列逐渐变得更加平稳．这将使预报的准确性更加提高，并导致估计值$\hat{\sigma}_a^2$随之减小，因而使本算法中判据的门限分布得更狭窄，进一步加强本算法的数据处理能力．而如果对本问题采用现有一般的数据处理算法，由于数据本身随机波动剧烈，不易根据经验设置适当的判据，故很难获得满意的处理效果．

参 考 文 献

1 段生荣．连续变化量监测过程中排除干扰数据的一种方法．计算机应用研究．1986.3

2 李济芳．电网微机监控系统抗强电磁干扰的措施．微型机与应用．1988.6

3 王正光等．数据采集与处理．国防工业出版社，1985

4 贾沛璋、宋征桃．最优估计及其应用，科学出版社，1984

5 安鸿志等，时间序列的分析与应用，科学出版社，1983

6 杨位钦，顾岚．时间序列分析与动态数据建模，北京工业学院出版社，1986

7 段文泽，刘士荣．城市供水系统负荷量的分时分地预报，发展战略与系统工程，学术期刊出版社，1987

（编辑：刘家凯）

A RELIABILITY PROCESSING ALGORITHM SUITABLE FOR STOCHASTIC AND SUDDENLY VARYING DATA

Du Qiang

(Chongqing Institute of Industrial Management)

Duan Wenze

(Chongqing Institute of Archit. and Engin.)

ABSTRACT In regard to stochastic and suddenly varying data, a new kind of data processor algorithm based on adaptive prediction is presented. The algorithm describes the changing pattern of measured variables by means of time series models with slowly time-varying parameters,forms the criterion by use of the adaptive and its 95% belief limit,and processes the measured data so as to reject or restrain outliers mixed in them. The simulation example indicates that the algorithm surmounts difficulties which most current algorithms encounter when processing stochastic and suddenly varying data and that it is of obvious effects.

KEY WORDS data-processing, stochastic variable, time seriesanalysis, adaptive prediction