自動データ処理における不良データの除去

(Removal of Inappropriate Data in Automatic Data Processing)

石　川　甲子男*

1. まえがき

電子計算機によって Data(観測値，事務資料など)を処理する場合，Data 作成から Data 処理までの過程に人間が介在することは，Data 処理の自動性が失われることと，人為的誤差が入り得ることなどから極力避けなければならない．

このため通常，Data 作成機(観測装置，事務資料用パンチ・カード作成機など)と計算機とを直結させる方法，両者の間に Teletype などを結合して無線または有線によって Data を計算機に供給する方法等がとられる．もちろん，この場合 Card とか Tape とかを媒介に用いる方が便利な場合が多い．

こういう場合に特に問題となるのは，人為的誤差にくらべれば非常に少いが，次のような誤差が Data に加わる恐れがあることである．すなわち，

(1) Reading error

(2) Mechanical error

で，(1)は人間が観測装置等で目盛を誤読取りすると同様の誤りで，読取値の符号化が誤って行われたために生ずるもの．(2)は Data 作成機から計算機までに生ずる突発的かつ瞬間的誤動作のために生ずる誤りや，空電などが Data に影響したために生ずる誤りを仮りにこう名付けたものである．これらを除くため，いわゆる ADP(Automatic Data Processing)において，すでに Data の取捨を計算機内で自動的に行うよう考えられているが，経験的な状態を計算機に設定して行うものが多く，一般性を欠くうらみがある．本報告は Data の取捨を Information Theory の形をかりて，やや定量的に扱うことを試みたものである．

2. Filter 理論の応用

Data 作成機から得られる Data を一つの Information と考えれば，前節の各誤差は，不必要な Data であって，Noise とみなすことができる．したがって，これらの Noise を取除くためには，適当な Filter を如何に設計するかということに帰する．(ただ問題の性質上，Noise の

* 建設省国土地理院　昭和35年3月31日応用物理学連合講演会にて発表　昭和35年5月17日受理

加わった Information は Noise とともに捨てることとする.)

いまこの Filter として，次の3つが考えられる.

(1) Band Filter

(2) Statistical Filter

(3) Personal Filter

次に，これらの Filter の性質，用途等についてのべる.

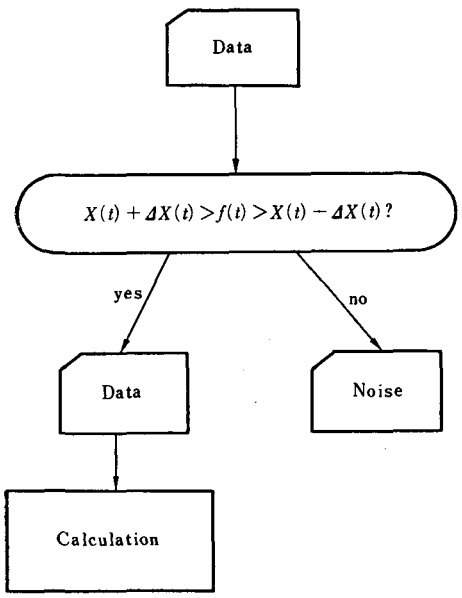
3. Band Filter

Band Filter は，Data すなわち Information の内容が予め一定範囲にあることが想定される場合，極端に予想値から外れた値をもつ Noise を除去するためのもので，もっとも primary な設計でよい. したがって，この種の Filter はあえて Filter というまでもなく，すでに各方面で用いられているが，一応順序として導入しておく.

いま Data ; $f(t)$ がある範囲 $X(t)\pm\Delta X(t)$ の中にあり，かつその外にないことが分っているとき(一般に Data のごく大略の予想値は得られるので)，この $X(t)\pm\Delta X(t)$ を形式的に Band filter の伝達関数 Y_B とおいて

$$Y_B=X(t)\pm\Delta X(t) \qquad (1)$$

と書くことができる. ここに $X(t)$ は $f(t)$ に予想される概略の平均値で，たとえばこの Filter に通すべき Data の値が，少くとも桁数のみは変わらないことが分っているときは $X(t)\fallingdotseq\frac{1}{2}(10+1)10^n=5.5\times10^n$ (n は Data に予想される桁数)となる. また $\Delta X(t)$ はこの Filter の Band width で Data に予想される上限の値と下限の値との差の 1/2 をとればよい. 上の例では $\Delta X(t)\fallingdotseq\frac{1}{2}(10-1)10^n=4.5\times10^n$ となる.



第1図 Flow Chart of Band Filter

この Filter では極端に大きな Noise を取除くのが目的なので，$\Delta X(t)$ は安全をみて幾分大きくとった方がよい. この Filter で濾波できなかった Noise は次の Statistical Filter で除くことができるからである.

また t は時間を表わし，もし Data が時間的に変動するものであれば $Y_B(t)$ も時間の関数にならなければならない.

この Filter は実際には，計算機内に Data が $X(t)\pm\Delta X(t)$ の範囲を出たときは Data として採用しないような計算の program を設定しておけばよい. これは各 Data が $X(t)+\Delta X(t)$ より

小さいか，$X(t)-\Delta X(t)$ より大きいかを判断させて，Yes ならばその Data を通過させ，No ならば濾波させることにより容易にできる．

この場合注意すべきことは，濾波した Data，すなわち Noise を，全く消してしまわないで一応 punch out しておくことである．これは次の(2)，(3)の Filter についても同様で，ある実験の段階において Noise として捨てた Data が，後程必要になることがあるからである．この間の関係を図示すると第1図のようになる．

4. Statistical Filter

Statistical Filter は，Data のばらつきの度合から推定して Band width を定め，極端に他の Data から離れている Data を除くもので，実際には Interval estimation の考えを拡張して，次のように設計する．ただし Noise のない Data は Ergodic 集合をつくり，定常性を満足しているものとする．この仮定は通常の観測から得られる Data に対してはすべて成立する．

いま与えられた Data $f(t)$ が区間$(\tau_1,\ \tau_2)$にある probability α(いわゆる信頼区間)は，

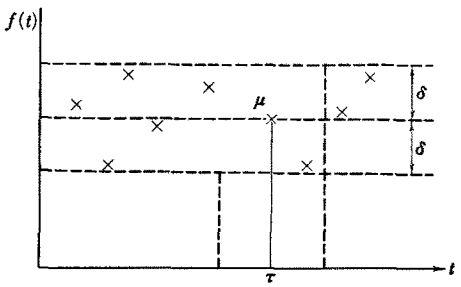
$$\alpha=P(\tau_1<f(t)<\tau_2) \tag{2}$$

で与えられる．もし i 番目の Data $f(t_i)$ に Noise が入って極端に $f(t_i)$ の値が大または小になったときは，この区間$(\tau_1,\ \tau_2)$に，この i 番目の Data $f(t_i)$ が入らないように α を定める．

たとえば，Data の属する母集団が，Normal Distribution $N(\mu,\ \sigma^2)$ に従う場合，$f(t)$ と $f(\tau)=\mu$ の差が，Standard Deviation に著しく外れたための Measure として選んだ任意の正数 δ を越えない probability α は

$$\begin{aligned}\alpha&=P[|f(t)-f(\tau)|<\delta]\\&=\frac{1}{\sqrt{2\pi}\cdot\sigma}\int_{\mu-\delta}^{\mu+\delta}e^{-\{(t)-\mu\}^2/2\sigma^2}\cdot df(t)\\&=\frac{1}{\sqrt{2\pi}}\int_{-\delta/\sigma}^{+\delta/\sigma}e^{-u^2/2}du\\&=\frac{2}{\sqrt{2\pi}}\int_0^{\delta/\sigma}e^{-u^2/2}du\end{aligned} \tag{3}$$

で与えられる(第2図参照)．すなわち



第2図

$$\begin{aligned}\alpha&=P[-\delta<f(t)-\mu<\delta]\\&=\frac{2}{\sqrt{2\pi}}\int_0^{\delta/\sigma}e^{-u^2/2}du\end{aligned} \tag{4}$$

ここで $\delta=\lambda\cdot\sigma$ とおけば

$$\begin{aligned}\alpha&=P[\mu-\lambda\cdot\sigma<f(t)<\mu+\lambda\cdot\sigma]\\&=\frac{2}{\sqrt{2\pi}}\int_0^{\lambda}e^{-u^2/2}du\end{aligned} \tag{5}$$

この式より，Data 作成機で得られた Data を Sample として算出した Standard Deviation σ

と，適当に選んだαによって定まるλとから信頼区間が定まる(この場合の信頼区間とは，通常の区間推定におけるものとは若干意味が異なる)．したがって，この区間を Statistical Filter の Band width にとればよい．αは正確には Data 作成機の性質，性能(あるいは Data そのものの性質)などにより予め適当な値を設定すべきであるが，2，3 の実験例(Table 1)から，Niose を取除いてなおかつ Data を極力保存するためには，$\alpha=0.80(80\%)$程度が良好であることが分る．すなわち例において，第1例では，Data が揃っていてこの Data はすべて有効であると考えれば，$\alpha=0.90(90\%)$で$\delta=1.32$となり表中の Data はすべて濾波されないので適当であるが，Data の性質上，表中の4番目の Data 64990.8 が Noise とみなされるような場合は$\alpha=0.80(80\%)$が適当となる．

第 1 表

Example 1	
	64992.6
	1.4
	1.5
	0.8
	1.7
	2.9
	2.3
	(5.1)
	3.1
	(644992.3)
Mean	64992.0

Table of $\alpha=\frac{2}{\sqrt{2\pi}}\int_0^{\lambda}e^{-u^2/2}du$

$\alpha\cdot100$	λ
99	2.5758
90	1.6449
80	1.2816
70	1.0364

	$\sigma=0.80$		$\sigma'=1.32$	
$a\cdot100$	$\lambda\cdot\sigma$	range	$\lambda\cdot\sigma$	range
99	2.06	−0.02～4.10	3.04	−1.11～5.69
90	1.32	+0.72～3.36	2.17	+0.12～4.46
80	1.03	1.01～3.07	1.69	0.60～3.98
70	0.83	1.21～2.87	1.37	0.92～3.66

また Noise を仮定するため，この Data の中1つだけ数値を変えて( )の中のようにすると，信頼区間は第1表右下のようになって，この場合は$\alpha=0.90$でも 0.80 でも( )中の Data は完全に濾波される．第2表は別の例で，この場合でも$\alpha=0.80$が適当なことが分る．

第 2 表

Example 2	
	1.6
	1.7
	1.4
	1.0
	4.0
Mean	1.9
$\sigma=$	1.3

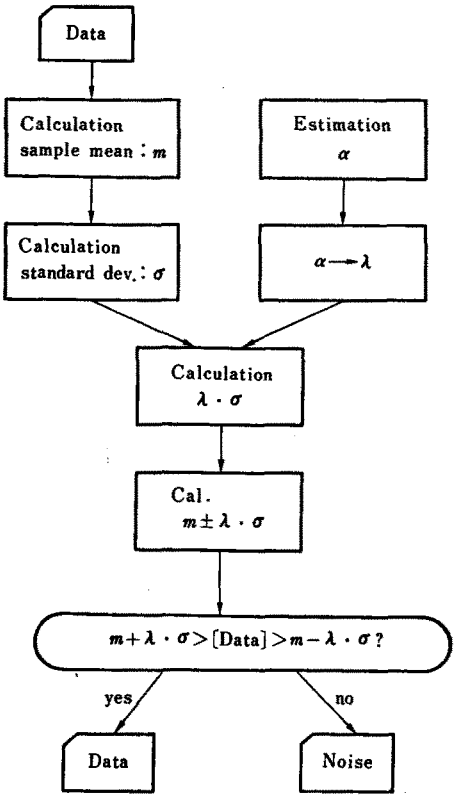
$\alpha\cdot100$	$\lambda\cdot\sigma$	range
99	3.35	−1.41～+5.29
90	2.14	−0.20～+4.08
80	1.67	+0.34～+3.61
70	1.35	+0.59～+3.29

第 3 表

Example 3

	1.7
	1.4
	5.0
Mean	2.7
$\sigma=$	2.0

$\alpha\cdot100$	$\lambda\cdot\sigma$	range
90	3.29	−1.29～+5.29
80	2.56	−0.56～+4.56
70	2.07	−0.07～+4.07



第3図 Flow Chart of Statistical Filter

また Data の数が非常に少なくて，ただ2つしかない場合では，Noise があっても何れが Noise か不明であることは，人間が判断する場合と同様であるが，3 つ以上あれば，その中の1つのみ，Noise が混入した場合には，相当性質の悪い Data でも濾波できることが実験例3(第3表)で確められる．

実際には，供給された Data をいったん計算機内に蓄え，Standard Deviation を計算してから，予め与えられた α より定まる λ によって δ を計算して Band width として先の Data を通過させる．すなわちいったん入れた Data を feed back する形になる．このようにして定めた Statistical Filter の伝達関数 Y_S は

$$Y_S(t)=X(t)\pm S(t,\ \alpha) \qquad (6)$$

とかける．ここに $X(t)$ はこの Data の平均値，$S(t,\ \alpha)$ は上述の Band width である．$Y_S(t)$ が時間の関数となることがあるのは，Band Filter の場合と同様である．Statistical Filter の Data の流れを図示すると第3図のようになる．

5. Personal Filter

前節までの2つの Filter は，取扱う Data の固有の性質にはあまり関係しないが，Personal Filter は取扱う Data の固有の性質を考慮したものである．したがってこの Filter は取扱う Data 個々について設計する必要がある．

一般に，相異なった取り方によって作成(または観測)された Data の組を $G_1(t)$，$G_2(t)$，$G_3(t)\cdots\cdots G_n(t)$ とする．これらが一般に

$$f[G_1(t),\ G_2(t),\ G_3(t)\cdots\cdots G_n(t)]\leqq\beta \qquad (7)$$

なる関係が分っておれば，各 Group に属する Datum を取り出して組合わせた結果，上の式を

満たすかどうかを調べれば，Noise が発見できる．すなわち i 番目の Datum $G(t_i)$ について

$$f[G_1(t_i),\ G_2(t_i),\cdots\cdots G_n(t_i)]\leqq\beta$$

が満足していて，

$$f[G_1(t_{i+1}),\ G_2(t_i),\cdots\cdots G_n(t_i)]\leqq\beta$$

が満足していなければ，明らかに $G_1(t_{i+1})$ の Datum が Noise となる．

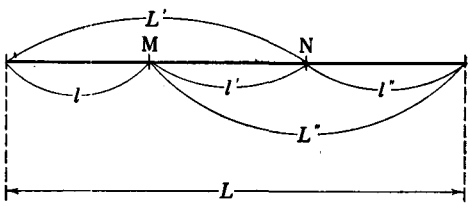
Personal Filter の伝達関数 $Y_P(t)$ は

$$Y_P(t)=X(t)\pm P(t,\ \beta) \qquad (8)$$

とかける．ただし，$P(t,\ \beta)$ は異なった Group に属する Data 間に(7)式の関係があるときに定まる Band width である．

Personal Filter の例

いま Personal Filter の最も簡単な例として，長い物体を測定する場合，第4図のように2点 M, N で分割して測定し Data として l の Group l_1, l_2, $\cdots\cdots l_k$, l' の Group l_1', l_2', $\cdots\cdots l_k'$, l'' の Group l_1'', l_2'', $\cdots\cdots l_k''$ を得たとする．このとき L', L'' の値があたえられていて，i 番目の Data l_i と l_i', l_i' と l_i'' の組合わせをとって，それぞれの和が $L'\pm\varepsilon'$ または $L''\pm\varepsilon''$ を越えていなければ，その Data は通過させるが，何れかを越えていればその Data の一方は濾波する．すなわち l_i+l_i' が $L'\pm\varepsilon'$ の範囲内にあって，$l_i'+l_i''$ が $L''\pm\varepsilon''$ を越えていれば，l_i'' は濾波させ逆に $l_i'+l_i''$ が $L''\pm\varepsilon''$ の範囲内にあって，l_i+l_i' が $L'\pm\varepsilon'$ を越えていれば，l_i は濾波させる．ただし両方がそれぞれ $L'\pm\varepsilon'$, $L''\pm\varepsilon''$ を越えているときは，どの Data が Noise かは不明となる．したがってこの場合は，これらの Data はすべて濾波する方が無難である．特にこのような場合が生じやすいときは，(1)の Band Filter の Band width をなるべく狭くしてこれを避けるようにする．



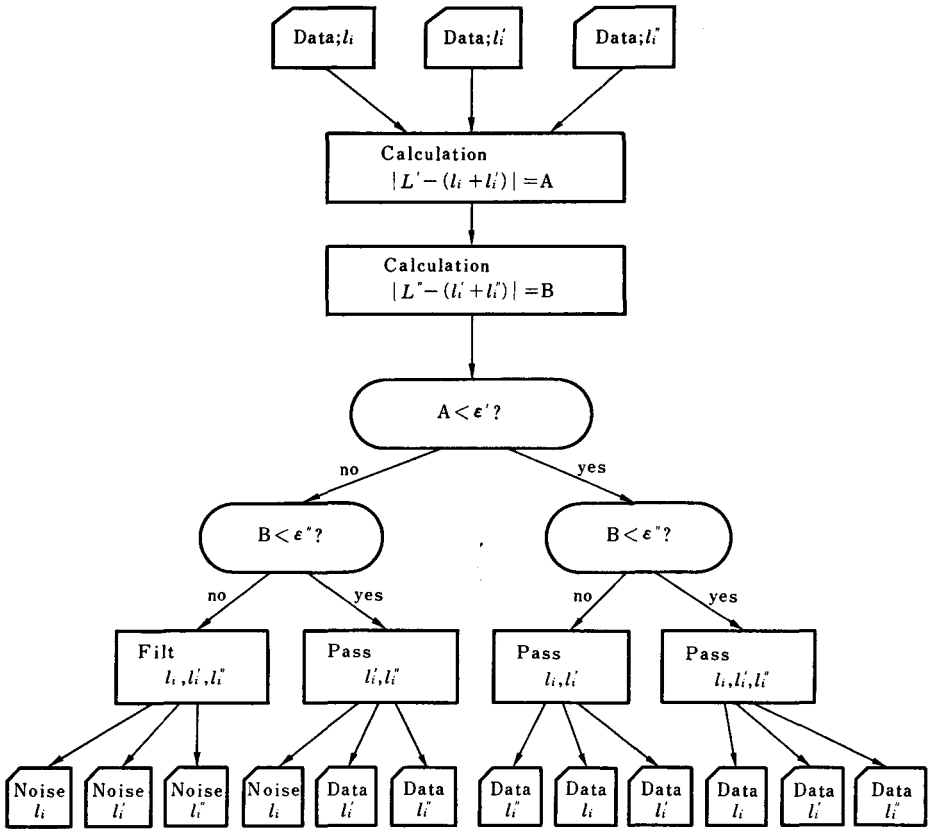
第4図 Example of Measurement of Length

この例における Personal Filter の Data の流れは第5図のようになる．

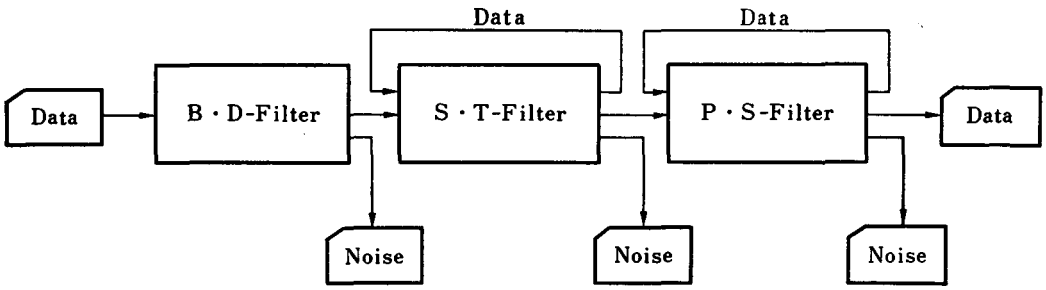
6. 結　　び

以上3つの Filter の状態を図示すると第6図のようになる．

これらの Filter は Data の性質によってただ1つのみでよいことがあり，2つ必要なときもあり，またすべてを使うこともある．そのときに応じて適当に組合わせて用いる．そして若干の時間的遅れの後，Noise を取除いた Data を取出し正常の計算に移れる．



第5図 Flow Chart of Personal Filter



第6図 Diagram of Combination of Filter

参 考 文 献

1) Goldman, S.: Information Theory, Prentice-Hall, New York, 1953.
2) Bell, D. A.: Information Theory and Its Engineering Applications, Sir Isac Pitman & Sons, London, 1955.