II — 9　　2，3の水文事象の異常値解析

信州大学工学部　正員　荒木　正夫
信州大学工学部　正員　寒川　典昭
信州大学工学部　学生員○田中　信治

1　概要

我々が水文統計量を解析していくとき，資料の中で少数のかけはなれた観測値に出くわす場合がある．本稿はこのような観測値の異常性を客観的に評価するために北川[1]のモデルを導入し，それによって年最大日降水データの異常値解析，水文系列の分布特性の検討をはかるものである．

2　異常値解析ベイズモデル

n個のデータの中にk個の異常値が含まれているとき，正常値x_iは平均μ，分散σ^2，異常値x_jは平均μ_j（$J=1,2,\cdots,k$），分散は正常値と同じσ^2を持つ正規分布に従うものとする．このとき，k個の異常値に対する$\mu_1 \leq \mu_2 \leq \cdots \leq \mu_k$の割り当てにより，k！通りのモデルが考えられることになる．モデルのパラメータμ，σ^2の最尤推定値$\hat{\mu}$，$\hat{\sigma}^2$は次式で与えられる．

$$\hat{\mu}=\frac{1}{n-k}\sum_{l=1}^{n-k}x_i(l) \quad \text{-----(1)} \qquad \hat{\sigma}^2=\frac{1}{n}\left\{\sum_{l=1}^{n-k}(x_i(l)-\hat{\mu})^2+\sum_{l=1}^{k}(x_j(l)-\hat{\mu}_l)^2\right\} \quad \text{-----(2)}$$

ここで，$\hat{\mu}_l$は各異常値に割り当てられた平均値とする．次に，最大対数尤度は次式で表わされる．

$$L(x|J,\hat{\mu},\hat{\mu}_j,\hat{\sigma}^2)=-\frac{n}{2}\log 2\pi\hat{\sigma}^2-\frac{n}{2} \quad \text{-----(3)}$$

ここに，$J=\{j(1),\cdots\cdots,j(k)\}$は，n個のデータのうちk個のデータが異常値である集合を表わす．さらに，最大対数尤度と平均対数尤度との差を求めることにより，平均対数尤度の不偏推定量が求められ，結局モデルの尤度$P(x|J)$は次式で定義される．

$$P(x|J)=\exp\left\{L(x|J,\hat{\mu},\hat{\mu}_j,\hat{\sigma}^2)-\frac{n(k+2)}{n-k-3}\right\} \quad \text{-----(4)}$$

異常値の個数の事前分布を一定値$P(k)=\alpha$とし，n個のデータからk個の異常値が現われる組み合わせの数${}_nC_k$と，k個の異常値へのk個の平均値の割り当ての仕方k！を考慮すると，特定のモデルの事前確率は，

$$\pi(J)=\frac{\alpha}{{}_nC_k\,k!}=\alpha\frac{(n-k)!}{n!} \quad \text{-----(5)}$$

で与えられる．従って，各モデルの事後確率は，

$$\pi(J|x)=P(x|J)\,\pi(J) \quad \text{-----(6)}$$

となり，k！通りの場合を加え合わせると$x_{j(1)},\cdots,x_{j(k)}$が異常値である事後確率は次式となる．

$$\pi(x_{j(1)},\cdots,x_{j(k)}|x)=\sum^{k!}\pi(J|x) \quad \text{-----(7)}$$

最後に，個々のデータの異常性は，そのデータを異常値として含むモデルの事後確率の合計，すなわち周辺事後確率Πとして表わされる．

3．データ数と周辺事後確率

平均70.51，標準偏差30.73の正規母集団から，10個きざみで10～100個のデータを発生させた場合の異常値解析を行ない，それぞれ上側5位までのデータについての結果を表1に示した．これより，データ数が大きくなるに従って，異常性を表わす周辺事後確率が小さくなっていくことがうかがわれる．

このことは，小標本の水文資料からは異常値と判断されたデータでも，データ数が整備されると正常値とみなされる可能性があることを示している．

4．水文事象への適用例

4－1．年最大日降水データの異常値解析

菅平地点の年最大日降水データ55年分に対して，図1はデータをトーマスプロットしたものであり，表2は代表的な異常値が持つ周辺事後確率を示したものである．特に昭和56年のデータ215mm/dayは94.5%という非常に大きな周辺事後確率を持っており，トーマスプロットの結果からも他のデータから推定される母集団とはかけはなれたものであることが分かる．このデータについて対数正規法，岩井法で確率年を計算すると，それぞれ1000年以上，320年という結果が得られたが，異常値解析の結果はこのような確率年の議論に疑問を投げかけるものである．

表1　データ数による周辺事後確率の変化

n	順位	1	2	3	4	5
10	x	117	114	94	82	81
	II	.04	.04	.01	.01	.01
20	x	168	116	116	99	98
	II	.58	.03	.03	.01	.01
30	x	125	115	114	91	90
	II	.07	.03	.03	.01	.01
40	x	150	136	136	120	105
	II	.24	.10	.10	.03	.01
50	x	141	139	115	112	110
	II	.23	.21	.03	.02	.02

n	順位	1	2	3	4	5
60	x	151	135	121	117	113
	II	.37	.12	.03	.02	.02
70	x	131	127	120	117	115
	II	.03	.02	.01	.01	.01
80	x	148	133	124	122	121
	II	.12	.03	.02	.01	.01
90	x	161	144	129	129	117
	II	.26	.07	.02	.02	.01
100	x	164	135	125	121	121
	II	.28	.03	.01	.01	.01

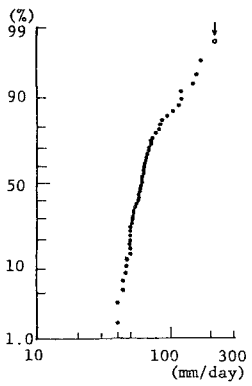

図1　年最大日降水量のトーマスプロット

4－2．水文系列の分布特性の検討

高棹・池淵[2]は水文系列の分布特性を検討したが，日単位以上の場合に満足のいく結果が得られなかった．彼らも指摘しているように，その原因の一つに異常値の存在が考えられる．従って，ここでは木曽福島雨量観測所のデータを取り上げ，前出のモデルでいくつかのデータの異常性を計算して，平均系列と標準偏差系列の安定性を標準偏差で評価した．その結果が表3であり，ここでAは異常値を除かない場合，Bは周辺事後確率5%以上の異常値を除いた場合である．年単位以外では，Aに較べBは平均より標準偏差の方が安定しており，このことは高棹らの議論に妥当性を与えるものである．ただし，変動係数で評価した場合は一概にそのようなことは言えなかった．

表2　主な異常値とその周辺事後確率

x	215	170	158	148	120
II	.945	.812	.713	.552	.043

表3　$\bar{x}$, σ の標準偏差比較

時間単位		標準偏差 A	標準偏差 B
6月15日	$\bar{x}$	.484	.190
	σ	.164	.107
9月15日	$\bar{x}$	.418	.649
	σ	.229	.188
6月中旬	$\bar{x}$	.153	.215
	σ	.105	.098
9月中旬	$\bar{x}$	.403	.380
	σ	.191	.114
6月	$\bar{x}$	.102	.109
	σ	.080	.051
9月	$\bar{x}$	.121	.058
	σ	.161	.054
3,4,5月	$\bar{x}$	.036	.047
	σ	.036	.027
6,7,8月	$\bar{x}$	.048	.018
	σ	.066	.013
9,10,11月	$\bar{x}$	.047	.070
	σ	.083	.032
12,1,2月	$\bar{x}$	.122	.181
	σ	.078	.035
年	$\bar{x}$	.121	.074
	σ	.280	.167

$\bar{x}$：平均系列，σ：標準偏差系列

5．あとがき

このモデルは，水文資料の異常値解析において個々のデータに客観的な異常性の評価を与える点で，有用かつ実用的な方法と言えよう．今後，異常値の分散が正常値と異なった値を持つ場合の理論式を導出するとともに，正常値と異常値の属する分布型として指数分布，ガンベル分布を検討したいと考えている．

参考文献　1）北川源四郎：異常値解析ベイズモデル，数理科学 NO.213，1981.3

2）高棹琢馬・池淵周一：エントロピー的にみた降雨・流出変換特性とそのモデル化，京都大学防災研究所年報第23号 B－2，1980.4