離散分布の合成による渇水時流況のモデル化の研究
Stochastic modeling by the combination of two discrete distributions of flow-series in dry season

名古屋工業大学博士後期課程学生　鈴　木　正　人　Masato SUZUKI
名古屋工業大学工学部　長　尾　正　志　Masashi NAGAO

Flow series is considered by combination of the following two components. The one for the domain of smaller quantity of flow is a simultaneous correlated binomial distribution, the other for the larger quantity of flow is an independent binomial or Poisson distrubution. The whole flow model is composed of the linear combination of the distribution functions for those two models. The decision of boundary for two components and the selection of optimal model are judged by AIC(Akaike's Infomation Criteria). Through the analysis on the simulated data, the estimated results for the model are well coincident with the original parameters. In addition, by applying the model for the observed flow data at the drainage of Makio Dam, the decision of boundary and the characteristics of divided flow are considered as the reasonable results.

Keywords:stochastic modeling, flow series, dry season, AIC

１．研究の概要

渇水時流況を、量的・時間的に離散化して経験分布を求めると、離散化単位の取り方にもよるが、ほぼ全体の数％にあたる個数の量の大きなデータが存在し、分布形が不連続になったり、全体の積率を左右したりしてモデル化に大きな影響を及ぼすことがよくある。そこで、本研究は流入量の大きい範囲と小さい範囲は異なった統計的性質を持つと仮定し、量の小さい範囲は自己相関性を考慮した相関同時二項分布で、量の多い範囲は無相関の分布でそれぞれモデル化を行う。ついで、両モデルの混合によって全体のモデルを表現する。全体的なモデルの適否はＡＩＣによって評価し最適なモデルを採択する。すなわち、渇水時流況の離散モデル化とデータの分割を同時に行う手法を示す。また、実データへの適用計算を行い、渇水時流況に対する好ましい統計モデルの選定を検討する。

２．対象とするデータ[1)]

量的・時間的に離散化された流入量系列ｑｔ（ｔ＝１，２，・・・，Ｎ）を対象とする。ここでいう量的・時間的な離散化とは、ｑｔが０，１，・・という無次元の整数値を持つことを意味する。また、ｑｔ+1はｑｔに引続き生起した流入量を意味する。ここで、

ｎ(i)≡（ｑｔ＝ｉ）の発生個数，（ｔ＝１，２，・・，Ｎ）（ｉ＝０，１，・・，ｃ）

ｎｉｊ≡（ｑｔ＝ｉ｜ｑｔ+1＝ｊ）の発生個数，（ｔ＝１，２，・・，Ｎ－１）（ｉ，ｊ＝０，１，・・，ｃ）

ｃ：発生流入量の上限

と定義する。なお、上式に関しては、次式が成り立つ。

$$\sum_{i=0}^{c} n(i) = N \quad \cdots\cdot(1) \qquad \sum_{i,j=0}^{c} nij = N-1 \quad \cdots\cdot(2)$$

ここでｎ(i)は経験分布の周辺分布に、ｎｉｊは同時分布に相当した表現である。

３．モデル化に用いる理論分布

モデル化には、周辺分布として、二項分布とポアソン分布を、相関同時分布として二項分布を用いる。これらの理論分布を以下に示す。まず、二項分布周辺分布は

$$Pb(i)=\binom{r}{i}(1-a)^{r-i}a^{i}\quad(i=0,1,..,r)\qquad\cdots\cdot(3)$$

$$Pb(i)=0\qquad(i=r+1,\cdot\cdot)$$

また、条件付き分布は、次式となる。

$$Pb(j\mid i)=\sum_{s=0}^{\min(i,j)}\binom{i}{s}\binom{-i+r}{j-s}\{a(1-\rho)+\rho\}^{s}\times\{1-a(1-\rho)\}^{s+r-i-j}a^{j-s}(1-a)^{i-s}(1-\rho)^{j+i-2s}\qquad\cdots\cdot(4)$$

したがって同時分布Pbijは、

$$Pbij=Pb(i)\times Pb(j\mid i)\qquad\cdots\cdot(5)$$

と表される。平均、分散、相関係数はそれぞれ、

$$E(q)=ra,\ V(q)=ra(1-a),\ Corr(qt,qt+1)=\rho\qquad\cdots\cdot(6)$$

であり、パラメータrは正整数で生起変量の上限を意味する。

つぎに、ポアソン分布周辺分布を次式で示す。

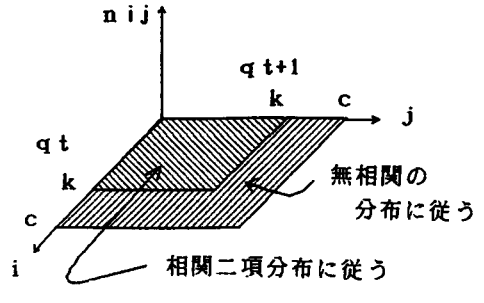
$$Pp(i)=\lambda^{i}\cdot\exp(-\lambda)/i!\qquad\cdots\cdot(7)$$

4．渇水時流況の離散化モデル[1)]

（1）モデルの概念

流量時系列を自己相関性の違いにより、低水、高水流量集団にそれぞれ分割する手法[2)]が提案されている。本研究では、同時分布のモデルを取り扱うことで、自己相関性の有無により渇水時流況を二つの集団に分割し、自己相関性を考慮したモデル化および母集団の分割を同時に行う。

分割の境界をi，j＝k（k＝0，1，・・，c）として量の少ない範囲は、一次の自己相関性を考慮した相関同時二項分布に、またそれ以外の範囲は相関のない二項、ポアソンのどちらかの分布に従うと仮定する。相関同時二項分布を用いたのは、従来、著者らが利水用貯水池問題を取り扱う際に、渇水期における貯水池系への入力として、この分布を用いた解析が有用性をもつ[3)]ことによる。まず、二つの領域で別個に求めたモデルを合成して全体のモデルを構成する（図－1参照）。モデルを合成する際、量の多い範囲と少ない範囲とが時間的に連続しておこるかどうかで2つの合成の方法を取る。一つは、量の多い範囲と少ない範囲とは異質の集団であり、これらが時間的に連続して生起することはないというモデルである。したがってこのモデルでは、量の多い範囲と少ない範囲との同時確率はほとんど0に等しい。もう一つは量の多い範囲と少ない範囲とが時間的に連続して生起するというモデルで、両者の同時確率は合成された周辺分布の積で与える。



図－1　モデルの概念図

合成されたモデルをMODELm(k)と表記する。ここで、kは分割の境界、mはモデル番号で表－1に示すような分布の組合せを意味する。たとえば、MODEL1(0)，MODEL3(0)は全てのデータが無相関のポアソン分布に従うとするモデルであり、モデル番号に関係せずMODELm(c)は全データが相関二項分布に従うモデルである。

表－1　用いたモデルと分布の組合せ

	量の少ない範囲	量の多い範囲	両範囲の連続生起
MODEL1	相関二項分布	周辺二項分布	しない
MODEL2	相関二項分布	ポアソン分布	しない
MODEL3	相関二項分布	周辺二項分布	する
MODEL4	相関二項分布	ポアソン分布	する

（2）　パラメータの推定

先述の仮定により、量の少ない範囲のモデル、つまり同時、周辺二項分布のパラメータ推定には、データとして、

経験分布n(i)（i＝0，1，・・，k）およびnij（i，j＝0，1，・・，k）を用いる。また、量の多い範囲、つまり周辺二項、ポアソン分布のパラメータ推定にはデータとして経験分布n(i)（i＝k＋1，・・，c）を用いる。なお各々範囲外（たとえば、量の多い範囲では、i＝0，1，・・，k）のデータはn(i)＝0とする。

パラメータの推定方法は、積率法、最尤法などがあり一概にどの方法がよいかは判断しかねるが、著者らが数値実験的に検討した結果[4]二項分布のパラメータ推定において最尤法が積率法に比べて有効であったので、ここでは最尤法を用いる。なお、二項分布は量の多い範囲と少ない範囲の双方でパラメータ推定を行うので、それらを区別するため、量の少ない範囲のパラメータに添字１（r1，a1，ρ1）を量の多い範囲のパラメータに添字２（r2，a2）を付す。

（３）　モデルの混合

ここでは、相関同時二項分布と独立な同時分布としての二項またはポアソンの両分布を合成する。まず量の多い範囲と少ない範囲が時間的に連続して生起しないとするモデルでは、たとえば、モデル番号２の場合、両範囲の同時分布を合成して、次式で表される。

$$pqij=\alpha\times Pbij+(1-\alpha)\times Pp(i)\times Pp(j) \quad (i,j=0,1,\cdots,c) \cdots\cdots(8)$$

上式において、pqijは合成された同時分布を、またαは全データの内相関二項分布に従うデータ数の割合を意味しており、本研究では次式で近似しておく。

$$\alpha=NL/(N-1), \quad NL=\sum_{i,j=0}^{k} nij \cdots\cdots (9)$$

いま、ポアソン分布のパラメータ推定にはデータとして、経験分布n(i)（i＝k＋1，・・，c）を用いるので、推定された分布においてPp(j)≒0（j＝0，1，・・，k）となり式（８）において、量の多い範囲と少ない範囲との同時確率pqij（iまたはj＝k＋1，・・，c）はほとんど0に等しくなる。

量の多い範囲と少ない範囲が、時間的に連続して生起するとするモデルは、たとえばモデル番号４の場合、まず周辺分布の合成を行い、ついで相関二項分布と周辺分布の積を合成して次式のように表される。

$$pq(i)=\alpha\times Pb(i)+(1-\alpha)\times Pp(i) \quad (i=0,1,\cdots,c)$$

$$pqij=\alpha\times Pbij+(1-\alpha)\times pq(i)\times pq(j) \quad (i,j=0,1,\cdots,c) \cdots\cdots(10)$$

上式で、pq(i)は合成された周辺分布を意味する。またαは式(9)の値を用いる。

（４）　ＡＩＣの計算[5]

MODELm(k)のＡＩＣをＡＩＣm(k)と表記し、これを計算する。ＡＩＣm(k)は次式で定義される。

$$AICm(k)=-2\times LLm(k)+2\times 自由パラメータ数,$$

$$LLm(k)=\sum_{i,j=0}^{c} nij\times ln(pqij) \quad \cdots\cdots(11)$$

自由パラメータ数は、相関二項分布で３個、周辺二項分布で２個、ポアソン分布で１個、ついで合成比率が１個であるから、一般的にMODEL1(k)，MODEL3(k)は６個、MODEL2(k)，MODEL4(k)は５個である。特別な場合として、全データが相関二項分布に従うとするMODELm(c)では３個、全データが周辺二項分布分布に従うとするMODEL1(0)，MODEL3(0)では２個、全データがポアソン分布に従うとするMODEL2(0)，MODEL4(0)では１個となる。

各モデル番号、各分割のうちでＡＩＣが最小のモデルを最適なモデルとして採択する。一般に複雑な混合をすればするほど、自由パラメータ数の増加、すなわちＡＩＣの増大に連なり、そのようなモデルは採択されにくくなる。しかしそのようなモデルがあえて採択された場合には、流量時系列の母集団が統計的特性の異なった複数の集団から形成されていたといえるだろう。普通あまり自由パラメータ数の多いモデルはパラメータ推定が不安定になるなどの理由で分布としては適切でないが、本研究では、ＡＩＣだけで、モデルの適否を判断するため、自由パラメータ数が５個や６個といったモデルを採択することも起りえる。

５．模擬発生データによる検討

パラメータの再現性を検討するため、模擬データに対し、最適なモデルを策定する。量の少ない範囲の母集団として、ｒ＝５，ａ＝0.4，ρ＝0.6の相関二項分布を、量の多い範囲の母集団としてｒ2＝10，ａ2＝0.5,0.6,0.7,0.8,0.9の周辺二項分布を仮定した。それぞれの範囲の分布を、混合の比率α＝0.90,0.95,0.99の３種で合成し、全体のモデルを構成した。合成は、量の多い範囲と少ない範囲とは時間的に連続して生起しない、つまり先述のMODEL1に従う形で行った。合成された分布に従う、全データ数Ｎ＝100,500,1000個の頻度分布を模擬データとし、MODEL1およびMODEL2によってモデル化した。二項分布の積率は、式（６）で表現されるから、量の少ない範囲の母集団の平均は、2.0、量の多い範囲の母集団の平均は、それぞれ5.0,6.0,7.0,8.0,9.0となり、パラメータａ2が小さいほど両母集団の平均値に差がなくなり、データの分割が難しくなることが予想される。

表2　最適モデルの選定例（ｒ1＝５，a1＝0.4，ρ＝0.6，ｒ2＝10，a2＝0.8，α＝0.9，N=500）

No	α	r1	a1	Corr.	MODEL1			MODEL2	
					r2	a2	AIC1	λ	AIC2
0	0.	−	−	−	10	0.2565	4422.5034	2.5648	3995.3823
1	0.3077	1	0.7040	0.3140	10	0.3363	4309.3496	3.3626	3962.0828
2	0.6215	2	0.6590	0.3980	10	0.4492	4034.4795	4.4920	3750.4202
3	0.8300	3	0.5836	0.4828	24	0.2634	3405.2039	6.3214	3383.2705
4	0.8988	4	0.4880	0.5580	10	0.7900	3118.4221	7.9000	3164.8184
5	0.9069	5	0.3981	0.6100	10	0.8152	◎3051.6665	8.1522	○3148.0219
6	0.9130	6	0.3321	0.6390	10	0.8302	3060.2493	8.3023	3156.3796
7	0.9332	7	0.2904	0.6692	10	0.8697	3092.9851	8.6970	3173.5242
8	0.9636	25	0.0876	0.7520	10	0.9278	3217.4058	9.2778	3247.6682
9	0.9899	25	0.0978	0.7710	10	1.0000	3412.1382	10.0000	3335.8369
10	1.0000	25	0.1026	0.7760	−	−	3411.8601	−	3411.8601

表－３パラメータ推定精度の比較(N=100)

α a2	0.5	0.6	0.7	0.8	0.9
0.90	×	×	×	×	△
0.95	△	△	×	×	△
0.99	×	×	×	×	×

表－４パラメータ推定精度の比較(N=500)

α a2	0.5	0.6	0.7	0.8	0.9
0.90	×	△	△	○	○
0.95	×	△	△	○	○
0.99	△	△	△	△	○

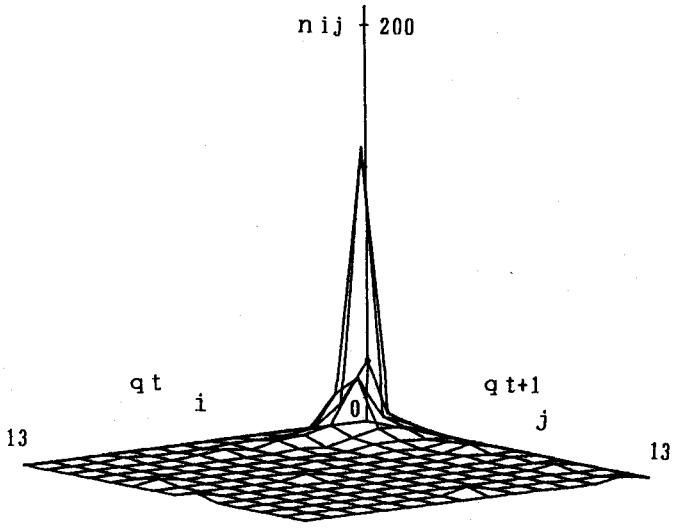
表－５パラメータ推定精度の比較(N=1,000)

α a2	0.5	0.6	0.7	0.8	0.9
0.90	×	×	○	○	○
0.95	×	△	△	○	○
0.99	×	△	△	△	○

最適モデルの策定例として、量の多い範囲の母集団として、ｒ＝１０，ａ＝0.8の周辺二項分布を合成の比率α＝0.9で合成した分布に従うデータ数500個の離散分布に対して、MODEL1，MODEL2でそれぞれモデル化した結果を表－２に示す。表の中でNoは、分割の境界、すなわち３．（１）で述べたｋに対応する。まず、全体でＡＩＣを比較すると、MODEL1のNo５がＡＩＣが最小なモデルとなっている。２番目にＡＩＣが最小なモデルであるMODEL1のNo６とのＡＩＣの差は、約９あるので、このモデルが全体の中で、最適なモデルといって良いであろう。合成される前の母集団のパラメータや、合成の比率もほぼ正確に推定されており、パラメータの再現性も非常に良好である。MODEL2の中でＡＩＣが最小なのは、No５であり分割の境界は正しく推定されている。これは、主体となる分布（この場合相関同時二項分布）のモデルが適切であれば、量の多い範囲のモデルがそれほど適切でなくても、分割の境界は正しく推定できることを意味していると思われる。

つぎに、データ個数Ｎや混合率α、さらに量の多い範囲の母集団の統計的性質がパラメータ推定に与える影響を考察する。Ｎ＝100，500，1000の場合で、αやパラメータａ2とパラメータ推定精度との関係を表３～５に示す。表中で×は両方の分布でのパラメータ推定の失敗を、△は量の少ない範囲のパラメータ推定の成功を、○は両方の分布でのパラメータ推定の成功を意味する。パラメータ推定の成否は、推定誤差が±５％以下であれば、成功したと判断した。まずＮ＝100の場合は、αやａ2に関係なくほとんどパラメータ推定に失敗している。本研究では、比較的自由パラメータ数の多いモデルを考えているため、100個というデータ個数は少ないと思われる。Ｎ＝500の場合は、αが一定の場合ａ2が大きくなるほどパラメータの推定精度はよ

くなっている。これは、混合率が等しいときは、両分布の統計的性質が異なるほど推定精度（特に量の多い範囲のパラメータ推定）がよくなることを意味する。また、ａ2が一定の場合、ａ2が小さい、すなわち両範囲の平均が近い場合は、αが大きくなるほど、推定精度がよくなる。逆にａ2が大きい、すなわち両範囲の平均が離れている場合には、αが大きくなるほど推定精度は悪くなる。両範囲の平均が近いと、αが大きくなることにより量の多い範囲のデータ個数が減少し、量の少ない範囲のパラメータ推定に雑音の減少として推定精度がよくなるような影響を与えるためと思われる。しかし、データの分離が行われ難いことにより、量の多い範囲のパラメータ推定は、ほとんど期待できない。両者の平均が離れていると、データの分割は行われやすく、量の少ない範囲のパラメータ推定はデータ個数が多いこともあり、ほぼ正確に行われる。またαが大きくなると、量の多い範囲のデータ個数の減少に連なり、パラメータの推定精度が悪くなる。Ｎ＝1,000の場合、全体的な傾向はＮ＝500の場合とほとんど同じで、その傾向がより顕著に現れているようにみえる。



図－２　牧尾ダム流域での渇水期流量の同時頻度分布

これら、模擬データによる検討により、本研究で示した方法でデータの分離、モデル化をする場合には、少なくとも数百個のデータと、ある程度、両母集団の統計的性質が異なっていることが必要だと思われる。しかし、たとえパラメータが正確に再現されなかったとしても、得られたモデルは、データ個数との兼ね合からみてデータをよく表現しているモデルと考えるべきであろう。

６．実データに対する適用計算

牧尾ダムの1969年～1986年冬期渇水期（１２月１日から翌年２月２８日まで）日流入量に対し、適用計算を行った。データの組は、各水期ごと（全１７期間分）および全渇水期を統合したもの、の１８組であり、各組を、単位期間１，３，５日、単位量＝単位期間×３ m^3／secにより離散化し、MODEL1～MODEL4でモデル化を行い、最適なモデルを選定した。

表－６　最適モデルの選定例（1969～1986年、牧尾ダムの冬期渇水期流入量）

No	α	r1	a1	Corr	MODEL1，3				MODEL2，4		
					r2	a2	AIC1	AIC3	λ	AIC2	AIC4
0	0.	－	－	－	13	0.1164	2076.768	2076.768	1.5131	1975.198	1975.198
1	0.6961	1	0.7958	0.7352	13	0.2399	1741.075	1741.043	3.1183	1658.080	1680.177
2	0.8824	2	0.5090	0.7830	13	0.3761	1756.075	1593.677	4.8889	1659.270	1552.288
3	0.9346	3	0.3810	0.7550	13	0.4923	1778.161	1581.625	6.4000	1673.342	1557.093
4	0.9510	5	0.2418	0.7620	50	0.1440	1698.496	1587.996	7.2000	1684.377	1587.183
5	0.9673	10	0.1270	0.7460	31	0.2677	1737.981	1637.154	8.3000	1715.178	1637.167
6	0.9771	50	0.0263	0.7322	21	0.4422	1769.815	1685.871	9.2857	1738.184	1686.718
7	0.9837	50	0.0270	0.7038	17	0.6000	1799.182	1716.085	10.2000	1756.012	1718.235
8	0.9902	9	0.1547	0.6342	13	0.8974	1941.289	1797.502	11.6667	1801.806	1804.432
9	0.9902	10	0.1391	0.6400	13	0.8974	1920.020	1788.121	11.6667	1797.990	1795.051
10	0.9935	11	0.1289	0.5880	13	0.9615	1965.420	1853.074	12.5000	1816.760	1862.964
11	0.993	12	0.1182	0.5912	13	0.9615	1958.409	1845.213	12.5000	1814.104	1855.104
12	0.9967	13	0.1129	0.5751	13	1.0000	1983.689	1894.810	13.0000	1829.807	1902.636
13	1.0000	14	0.1082	0.5380	－	－	1963.041	1963.041	－	1963.041	1963.041

選定例と

して、全渇水期の日流量を、単位期間5日で離散化したデータの同時頻度分布を図－2に、最適モデルの選定結果を表－6に示す。図－2に示した頻度分布をみると、量の大きなデータが量の少ないデータの集団から、飛び離れて存在していることが解る。さてこれら量の多いデータを量の少ないデータから分離したモデル化を行こなう。表－6のモデル選定結果をみると、全体では、MODEL4(2)がＡＩＣが最小で４つのモデルのうち、最適なモデルであることが解る。混合率αは0.88、自己相関係数ρ＝0.78であるから、全体の９割が自己相関性の強い流量時系列成分から構成されていることが推測される。２番目にＡＩＣが小さいのは、やはり、MODEL4(3)であるから、モデルとしては、MODEL4が適しているようだ。分布形は同じ組み合せで、量の多い範囲と少ない範囲とが時間的に連続して生起しないとするMODEL2のうちでＡＩＣが最小なのは、MODEL2(1)だが、MODEL4(2)とのＡＩＣの差は、100以上もある。また、MODEL1とMODEL3のＡＩＣを比較すると、全体的にみて、MODEL3のＡＩＣの方が小さくなっている。これらの結果から、このデータの場合は、量の多い範囲と少ない範囲が時間的に連続して生起するとしたモデルの方が妥当なように思える。

表－7　各単位期間における最適モデル数

単位期間	MODEL1	MODEL2	MODEL3	MODEL4	MODEL(c)
1日	2	2	5	7	2
3日	3	0	11	3	1
5日	2	0	10	5	1

ついで、各渇水期ごとに最適なモデルを求め、単位期間ごとに最適モデルの内訳を表－７に示す。表中でMODEL(c)というのはモデル番号に関係せず、全データが相関二項分布に従うとするモデルである。まず、全体的にみて、MODEL(c)は数ケースしかないので、今回のデータの場合は、量の多い範囲と少ない範囲とにデータを分割するモデルの方が適していたようだ。また、単位期間１日の場合は、かなり平均的にどのモデルも採用されているが、単位期間３日、５日ではMODEL3が最も多く採用されているので、全体的にみても、量の少ない範囲と多い範囲が連続して生起するというモデルの方が妥当なようである。

なお、このようなモデルの使用方法としては、たとえば、分割されたデータのうち、自己相関の強い流量時系列成分だけを貯水池系への渇水時入力として採用するのが実用的であろう。

７．まとめ

本研究では、渇水時流況データを流量時系列成分とそれ以外との成分に分離してモデル化する手法を示した。モデルとして、相関同時二項分布と、周辺二項分布、ポアソン分布の二つの分布を合成したものを用いた。しかし、基本的に離散分布であれば、どんな分布からモデルを構成してもよい。また、モデルを構成する分布の数も二つにかぎられるわけではない。それぞれのデータに適した分布を用いることで、データの分割や、統計的特性の単純化をより正確に行うことができると思われる。しかし、採用分布が異なった場合でも、本研究で示した手法で、ＡＩＣ基準によった最適モデルの選定が可能となろう。

８．参考文献

１）鈴木正人・長尾正志：ＡＩＣを用いた相関渇水時流況の離散モデル化の研究，第４４回年次学術講演会概要集Ⅱ部門，1989

２）室田明・神田徹：利水を対象とした流量時系列の解析について，水理講演会講演集，1969

３）鈴木正人・長尾正志：２段階推移モデルによる相関離散分布を受ける貯水池理論，土木学会論文集，第４１１号／Ⅱ－１２，pp.161-168,1989

４）鈴木正人・外山康彦・長尾正志：渇水時流況の表現のための離散化確率モデルの選定，土木学会中部支部講演概要集，1988

５）坂元慶行・石黒真木夫・北川源四郎：情報量統計学，共立出版社，1983