展　望

トランザードデータアナリシス
Trunsored Data Analysis

廣瀬　英雄
Hideo HIROSE

概 要

trunsored model は truncated model と censored model との尤度による比較を直接行い，モデルの尤度比検定を行うことを目的として作られた，打ち切りデータを統一的に取り扱うモデルである．ここでは，まずこのモデルを用いたパラメータの推定法や検定法について述べ，次にこのモデルを発展させたいくつかの応用について概説する．最後に今後の展開について簡単に触れる．

1. はじめに

ある母集団からデータがサンプルされた場合，その母集団の背後にはある確率モデルが存在すると考え，データはそこからランダムにサンプルされたと考える．そのとき，サンプルされたデータから母集団の背後モデルを求め，将来への予測を立てるのが統計の役割である．つまり，ある確率モデルから順方向に（演繹的に）出てきたデータを用いて，逆方向に（帰納的に）モデルの推定や同定を行うのが統計と言える．この背後の確率モデルを仮定する際には，均質な（homogeneous な）データが得られているというモデルを考えないと議論が進みにくい．従って，通常は，ある確率分布モデル（例えば正規分布やワイブル分布）を仮定してそのパラメータを求めたり，時には分布そのものの選定（モデル選定）を行っている．

しかし，実際のデータを扱う場合，母集団が複数の特性をもつ確率モデルに支配されていると思われることがある．例えば，かなり早い段階で故障が起こる製品とゆっくり故障が起こる製品とが混在している場合である．あるいはもっと極端に，製品のいくつかはかなり頑丈でいつまでたっても壊れず（これを durable と言う），またいくつかは容易に壊れる製品（これを fragile と言う）からなっている母集団の存在が疑われることがある．ここでは，このような状況を考えて，サンプル総数 n が分かっているとき，ある時刻までにそれらのいくつか（r としよう）が故障したという条件下で，最終的に故障する総個数や故障への経過をどのように求めたらよいのか，その基本的な考え方と，いくつかの応用について概説する．図 1 に fragile と durable とが混在したときの概念図を示す．

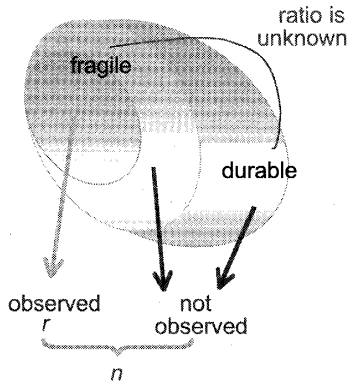


図 1　durable and fragile products are mixed

2. trunsored model

生存解析の典型的な教科書の1つである，Klein and Moeschberger[1]では，不完全データとして，censored data と truncated data の2つがあり，背後の確率モデルが均質であれば，サンプルデータの総数が分かっているとき censored data として，不明なとき truncated data として取り扱えば，背後にある確率分布のパラメータ(パラメトリックであれば）を求めることができると紹介されている．サンプルデータの総数は分かっているけれども，いつまでたっても壊れそうな気配のない製品が含まれていそうだという場合，もちろん censored data では具合が悪いので，故障しやすい製品の総数が分かっていないとして，truncated data として取り扱えば推定は可能になる．ただ，このときサンプル総数 n の情報は使っていない．

簡単のため，今，右側打ち切りだけを考え（censoring time を t_c とする），観測された故障数を r とすると，censored model, truncated model の場合，尤度関数はそれぞれ，

$$L_c = (1-F(t_c;\theta))^{n-r}\prod_{i=1}^{r} f(t_i;\theta), \qquad (1)$$

$$L_t = \prod_{i=1}^{r}\{f(t_i;\theta)/F(t_c;\theta)\}, \qquad (2)$$

で表される．先に述べたような場合，(2)を使って分布のパラメータ $\hat{\theta}$ を（例えば最尤推定法を用いて）求めれば，故障総数 m は，

$$m = r/F(t_c;\hat{\theta}), \qquad (3)$$

から自然に求めることができる．

この推定値 m がサンプル総数 n に近いとき，故障しそうにない製品は全く無いかもしれない．つまり，censored model で推定した方が適切かもしれない．このようなときには，censored model で取り扱った方がよいのか，truncated model で取り扱った方がよいのかを検定する必要が起こる．しかし，(1)，(2)の尤度は直接比較することができないので，尤度比検定を用いることができない．

そこで，尤度を(1)にそろえられるようなモデルとして，（全く壊れない製品と壊れやすい製品が混在している）混合分布モデルを考えることにする．全体に占める壊れやすい製品の割合を s として，

$$L_{tr} = (1-sF(t_c;\theta))^{n-r}\prod_{i=1}^{r}\{sf(t_i;\theta)\}, \qquad (4)$$

のような（自然に拡張された）混合モデル[2]を考えると，(4)を最大にするような $\hat{\theta}$ は，s に制限をつけなければ，(2)で求めた値と同じになり，(1), (4)の間の尤度を比較することが可能になる．つまり，(4)を使えば，censored model か truncated model かを尤度によって比較することが可能になり，(4)は，s=1 で censored model を表すことから，censored model, truncated model の両方を表すことができるようなモデルとなっている．そこで，このモデルを trunsored model[2]とした．実は，このモデルは，いろいろな場面で随分以前から使われていて，例えば工学系では，別名（limited failure population model）もついている[3]．工学系だけでなく医学薬学系にも応用されている．例えば，治療を受けた癌患者の生存モデルを表すとき，5年生存すれば治癒と見なされ，これを製品に置き換えると，治癒は壊れない製品と同等とみなすことができる[4]．

実は(4)で，s を割合と言いながら p という通常（確率を表す）文字を用いなかったのには理由がある．p を全体に占める確率と考えて p に $0 \le p \le 1$ の制限をつける考え方は自然であり，この制限のもとで Maller & Zhou[5] は censored model と truncated model との尤度比検定を行っている．ただ，この制限がついているため見かけ上計算が少し面倒になっている．しかし，この制限を外すと尤度比検定の計算に特別な工夫が要らなくなる．

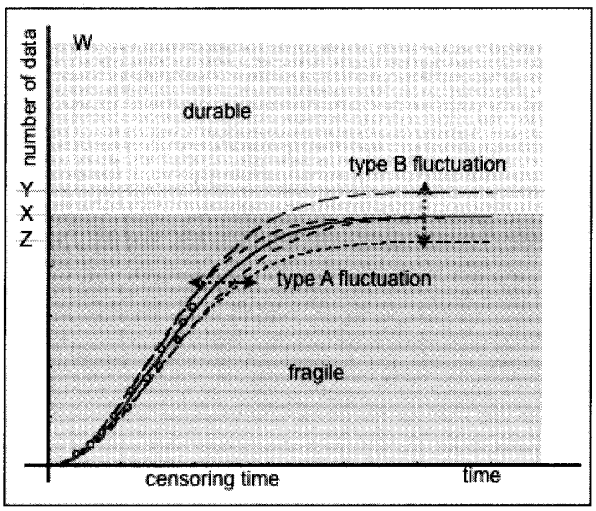


図 2　censored, truncated, and trunsored models

trunsored model が censored model や truncated model とどのような関係にあるのか，図で説明しよう．図 2 で，横軸は時間，縦軸は製品の（累積）総数を，また，○印は censoring time までに故障した製品を表す．censored data ならば総数 X（固定）のところが分かっているので観測データにより推定された確率分布にばらつきが起こるのは type A の形ということになる．一方，truncated data の場合，分布のばらつきは type B の方向にも起こる．この図の場合，fragile 部分は X のところまでを表しているが，総数 m は Y や Z のところにも収束するように推定される．つまり，type B の方向にも動きはある（X は上下に動く）が，結局は fragile の総数のところに収束する．あたりまえであるが，truncated data の場合 durable は全く考えていないということである．さて，trunsored data の場合，fragile 部分は type B の方向にも動くことは同じであるが，総数（W = fragile + durable）はあらかじめ分かっており（固定されており），durable と fragile の境界線は Y や Z のところにも上下する．これは truncated model と同じである．しかし，もし durable の割合が少ない場合，Y の位置が総数 W を超えるようなことも起こる．この場合，(4) では $s>1$ と推定される．これは不合理であって，Y の位置は W で押さえられていなければならないという立場が混合分布の考え方である．従ってこのとき，尤度の最大値は W の境界上でとられることになる．従ってこの値は，truncated model の最大尤度よりも小さい．一方，trunsored model ではこの制限を外しているため，s を自由に動かすことによって，尤度の最大値は truncated model との結果と一致する．結局，trunsored model は尤度を censored model に合わせながら truncated model（と censored model とを）を取り扱うモデルになっているということになる．そうすると，検定を行う際の仮説，

$$H_0 : s = s_0, \tag{5}$$

で，s_0 をパラメータ定義域の内点に設定できるので，尤度比検定の計算に特別な工夫は要らない．

2.1　electronic board failures[2)]

図 3 はある電子部品の出荷後のクレーム数を，6/30/1998 から 2/26/1999 の期間，ロットごとに示したものである．総出荷数は 2997，期間内のクレーム数は 133 であった．この図から，クレームがいつまで続くか，またクレームの総数はどのくらいになるかを見積もりたい．trunsored model を使ってこれらを推定すると，クレームの総数は 147 であった．尤度比を用いて 95%信頼区間を求めると[124, 173]，bootstrap からは[126, 175]が得られた．この例の場合，s=0.0491 なので，fragile と durable との境界が出荷総数を超えるような状況からはかなり離れているので，censored model を使うか truncated model を使うかの判断はほぼ自明である．

その後，3/1999 から 6/2000 の間にクレームは更に 42 個追加され，クレームの総数は 175 となった．この報告は最初の計算結果を出した後で受けたものである．

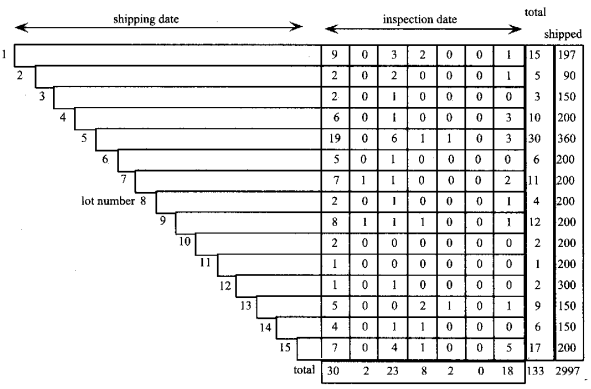


図 3　electronic board failures

3. mixed trunsored model

上に述べた trunsored model をいくつか組み合わせると，扱いにくいさまざまな実際問題に対応可能な場合がある．ここでは 2 つの場合について述べる．

3.1　type I mixed trunsored model

SARS の場合を考えよう．SARS による死亡率（case fatality ratio，略して CFR と言う）を，感染が蔓延するよりも早い段階で予測することは重要である．ここでの死亡率は，最終的な（収束値の）死者数を感染者数で割った値として求められる．さて，WHO は，死者，感染者，および治癒者の数を日ごとに報告してきたので[7]，これらを一緒に用いて死亡率を推定すると，死者数，感染者数だけから推定するよりも，更に推定精度が上がると考えられる．今，trunsored model を 3 つ用いて，死亡率を推定することを考えてみよう．つまり，

$$L_{tr}^{j}(\theta_j, s_j) = (1 - s_j F_j(t_c;\theta_j))^{n_j - r_j} \prod_{i=1}^{r_j} \{s_j f_j(t_i;\theta_j)\} \tag{6}$$

を感染者，死者，治癒者それぞれについてあてはめ，s_j の間にある制約

$$\varsigma(s_1, \cdots, s_J) = 0, \tag{7}$$

を与える．ここでは，ある地方の総人口に対する感染者の比，死亡者の比，治癒者の比を s_1，s_2，s_3 とすると，

$$s_1 = s_2 + s_3 \tag{8}$$

が成立する．このような(7)の制約がついた場合の mixed trunsored model[7]を type I のモデルと言うことにする．図 4 にその概念図を示す．

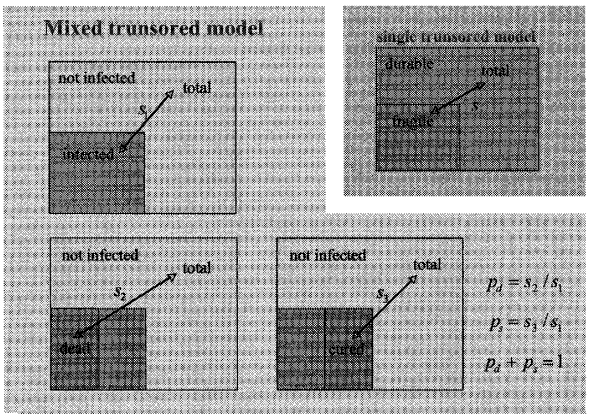


図 4　type I mixed trunsored model

このモデルの下で，SARS の CFR は Hong Kong で 17.16%と推定された．ただし，censoring time は 5/25/2003 の段階である．censoring time をいろいろ変えて CFR を求めてみたのが図 5 である．横軸は censoring time である．図から，終息段階における CFR の値を早い段階でも推定できていることが分かる．図には，死者と感染者のデータから truncated model を用いて推定した場合，および治癒者と感染者のデータを用いた場合も併記している．

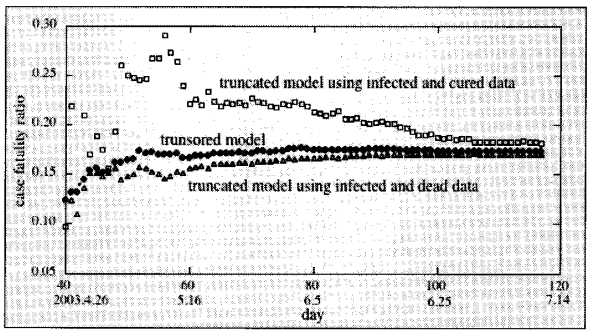


図 5　CFR estimated by trunsored model and truncated models[7]

また，trunsored model の場合 bootstrap を用いれば 95%信頼区間は[13.73, 19.04]%と計算された．更に SARS の終息段階における CFR の標準偏差の近似値は 0.58-1.07%程度となり，それほど小さい値とはならない．しかし，もし，truncated model を死者，および感染者の growth models にあてはめて，CFR を計算する場合，truncated model での終息値での標準偏差が 0 に近くなることを考えると，CFR の標準偏差も 0 に近い値になる．Canada, Taiwan, Singapore, Viet Nam など他の地域での CFR のばらつきに実際の値から考えて，CFR の標準偏差が 0 に近いことを理解するには，その地域特有の状況という解釈をするしかない．truncated model を使った結果と trunsored model を使った結果のどちらが受け入れられやすいだろうか．

3.2 type II mixed trunsored model

複数の trunsored model 間に制約条件があることは先と同様であるが，今度はモデルの中の支配パラメータなどに制約条件がある場合である．このような制約がついた場合の mixed trunsored model[8]を type II のモデルと言うことにする．例えば，故障に関するモードが複数存在し，重故障（1 つでもアイテムが故障するとシステム全体が止まる）と軽故障（複数のアイテムにこの故障が観察されてもすぐにはシステムの故障にはつながらない．また故障時刻は観測されず，検診などによってどの期間にこの状態になったかという情報のみが観測される）とが組み合わさって起こる現象の場合で，重故障と軽故障との間に類似の確率分布を仮定できる場合である．言い方を変えると，軽故障の情報だけでは推定値は得られないので，重故障の情報から支援して軽故障の状態を知ろうとする考え方になる．重故障と軽故障との間に類似の関係が見られるという物理的な条件，あるいは過去の経験に基づく条件などがあってはじめて可能な推測問題になる．この概念図を図 6 に示す．図で failure は重故障，malconditoned は軽故障を表す．

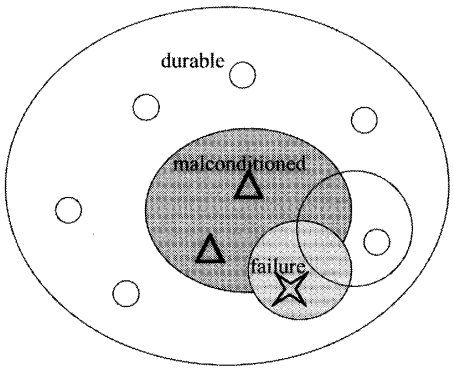


図 6　type II mixed trunsored model

例えば，数 10 万の単位の個数で設置されている電機設備に，致命的な故障が起こる場合と，すぐには故障には至らないが何らかの故障の引き金になる可能性が観察された場合で，しかもこれらはまれな現象と分かっているときなど，このモデルによって軽故障の広がりをおおよそつかむことができる．

4. 今後の展開

工場出荷時には均質に管理された製品も，フィールドで稼働するときには，出荷前には予想しなかった環境下に置かれていたり，想定していなかったストレスがかかったりするため，フィールドでは工場でさまざまな試験を行ったときよりも一般には速く劣化することが考えられる．図 7 は，まだ使用中の地下に埋設された電力ケーブルを掘り起こして，ストレスを上昇法によってかけたときの絶縁破壊データを表している．図は，劣化が単純な時間の関係（回帰）では表されず，いくつかの（単純な）劣化モードが混在している可能性を示唆している．4 つの回帰直線は，複数のモードを仮定したときに，最適なモード数を AIC によって求めた結果を示している[9]．

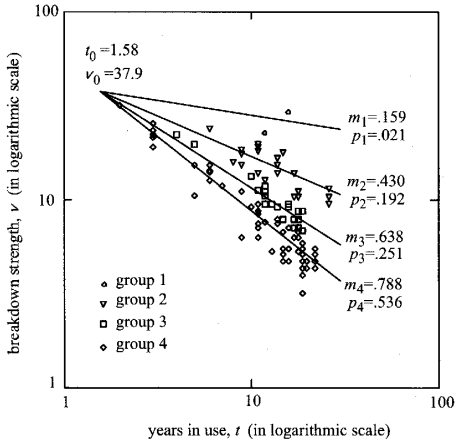


図 7　mixture Weibull power law model[9)]

ただし，これは，工場出荷時の加速試験による予測を補う目的で行った試みではあったが，相当なコストをかけてこのような試験を行っても，(個別の)ケーブルを交換すべきかどうかの判断にはまだ明快な回答を与えている訳ではない．

そこで，CBM (condition based maintenance) の考え方が生まれてくる．機器の状態を監視し，状態に変化が見られたら修理もしくは取り替えを行う方法である (図 8)．簡単そうに書いているが，状態の変化とその帰結との対応関係がつかめていないと監視は役に立たないので実は大変難しい．対応関係をつかむための基礎実験や調査を行い，それをもとにして対応関係のルールも作る必要がある[10)]．

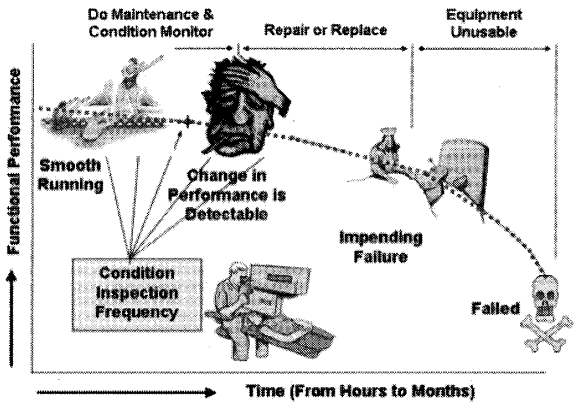


図 8　condition based maintenance

図 9 は変電設備で研究が進められているその 1 例を示したものである．対応関係を構成するのに，明示的なルールを提示できる決定木 (decision tree) の利用も行われている．図で EIA は絶縁設備を表す．

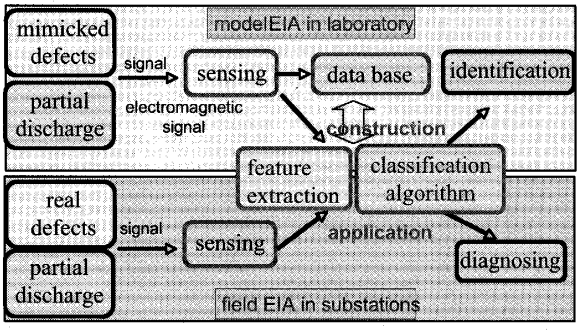


図 9　classification procedure in EIA diagnosing

図 9 では機器から出る信号をもとに状態を表す特徴量を求めて機器の状態を同定しているが，これを更に進めて，機器の状態の時刻変化をその機器に付随するさまざまな環境要因と併せ記録し，(たとえば決定木を用いて) 劣化を特徴付ける環境要因によって分類を行い，その環境要因下での劣化状態の変化を取り出すことも考えられる．例えば logistic regression tree[11)]はその 1 例である．図 10 にその概念を示す．劣化状態の経時変化を表すデータに censoring が起これば trunsored model で対処する必要性も考えられる．

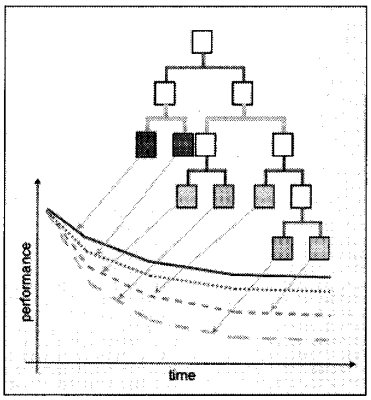


図 10　logistic regression tree

さて，個別の機器やシステムの状態の把握は当該のフィールドにとっては重要であるが，全ての機器を管理する経営側からは総合的な見方が必要になってくる．つまり，個々の機器やシステムの故障の起こる確率(probability)と起こった場合の帰結(consequence あるいは severity)との両方を同時に考えて，システムのどの部分から，あるいはどの機器から手当をしていけば経営的にリスクが最小になるかという観点からも考えあわせて総合的に判断しなければならない．図 11 に示すように，

$$\text{Risk} = \text{Probabiliy} \times \text{Consequence} \qquad (8)$$

の関係 [12]を把握しておくことも重要になる．図で，両軸は対数目盛りとして考えると(8)を理解しやすい．

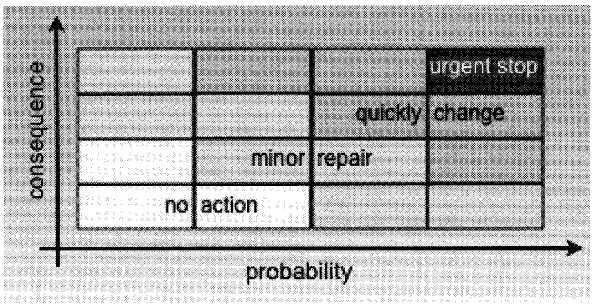


図 11　probability-consequence risk diagram

5. おわりに

ここではまず，trunsored model を導入した経緯と特徴，応用例を紹介し，次に mixed trunsored model を使うことによって更に複雑な問題への対処への可能性が高まったことを述べた．更に，精度の高いセンサーや状態の判断を高度に行うためのコンピュータやアルゴリズムの開発が格段に向上していく中で，preventive maintenance から monitoring maintenance あるいは condition based maintenance へと，より複雑ではあるが，きめの細かい信頼性解析への関心が高まっていることについて触れ，trunsored model が単独に取り扱われるだけでなく，今後はさまざまな外部（環境）要因と併せて考えていく方向性を示唆した．そして，最終的にはリスクについてどう立ち向かうかの経営判断も重要な要素であることについて触れた．

参考文献

1) J.P. Klein, and M.L. Moeschberger, *Survival Analysis: Techniques for Censored and Truncated Data*, Springer, (2000)

2) H. Hirose, The trunsored model and its applications to lifetime analysis: unified censored and truncated models, *IEEE Transactions on Reliability*, Vol.54, No.1, pp. 11-21 (2005.3)

3) W.Q. Meeker, Limited failure population life tests: application to integrated circuit reliablity *Technometrics*, 29, 1, pp. 51 - 65 (1987.2)

4) J. Boag, Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Statistics in Medicine*, 23, 3525-3543, (1949)

5) R. Maller, and X. Zhou, *Survival Analysis with Long-term Survivors*, Wiley, (1996)

6) WHO, http://www.who.int/csr/sars/country/en/ (2003)

7) H. Hirose, The mixed trunsored model with applications to SARS, *Mathematics and Computers in Simulation*, to appear.

8) H. Hirose, T. Matsumoto, Mixed trunsored model with application to the lifetime estimation of malcondition items, *RAMS2006,* 06RAMS_11A3, USA, (2006)

9) H. Hirose, Lifetime assessment by intermittent inspection under the mixture Weibull power law model with application to XLPE cables, *Lifetime Data Analysis*, Vol.3, No.2, pp.179-189 (1997)

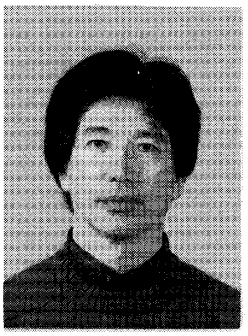
10) H. Hirose, et. al., A method to classify the signals from artificially prepared defects in GIS

using the decision tree method, *ISEIM2005 (2005 International Symposium on Electrical Insulating Materials)*, B6-8, pp.885-888, (2005)

11) L. Wei-Yin, Logistic regression tree analysis, *in Handbook of Engineering Statistics*, pp. 537-549, Springer (2006)

12) R. B. Jones, Risk Management Principles and Techniques, *RAMS2006,* 06RAMS_14C, USA, (2006)

（ひろせ　ひでお／九州工業大学）



廣瀬　英雄

1977年九州大学理学部数学科卒業．1988年工学博士（名古屋大学）．1977年（株）高岳製作所入社．1994年広島市立大学情報科学部情報数理学科教授．1998年九州工業大学情報工学部教授．電気学会，応用統計学会，日本計算機統計学会，日本統計学会，臨床研究・生物統計研究会，日本応用数理学会，情報処理学会，電子情報通信学会，日本数学会，日本OR学会，統計科学研究会，日本信頼性学会，コンピュータ利用教育協議会，日本工学教育協会，IEEE (Computer Society, Power Engineering, Reliability, Control, Dielectric and Insulation societies) ASA (American Statistical Association), IMS (Institute of Mathematical Statistics), MPS (Mathematical Programming Society), SIAM (Society for Industrial and Applied Mathematics, ACM (Association for Computing Machinery), AMS (American Mathematical Society).