情報学基礎　30－3
（1993．7．26）

データベース信頼性に関する考察

中村敏夫　　伊土誠一　　石垣昭一郎

ＮＴＴ情報通信網研究所

企業活動の結果として，大量の情報がデータベース（ＤＢ）化されてきた．この大量の情報が，企業の重要な資産になってきている．しかし，ＤＢのデータ品質の重要性が認識されつつあるにもかかわらず，データ品質についての検討は少ない．

ソフトウェアバグは一般的に単調減少であるのに比べ，データ不良は運用中のＤＢ更新のために必ずしも単調減少とは限らない．本稿では，このような特徴を持つデータの信頼性に関して，データ信頼性の推定法，及びデータ不良混入の要因分析結果を報告する．

A study of data reliability in databases

Toshio Nakamura　Seiichi Ido　Shoichiro Ishigaki

NTT Network Information System Laboratories

1-2356 Take Yokosuka-Shi Kanagawa 238-03 Japan

According to company's activity , a large amount of information have been stored in databases. And ,a large amount of information have been important property in company. Although data quality is recognized important , research for data quality is little.

Although software bug is generally on the monotonous decrease , data error is not always on the monotonous decrease , according to inserting and updating records in databases.

This paper describes a method of estimating data reliability and analyzing the factors deteriorating data reliability.

1．はじめに

企業活動の結果として，大量の情報がデータベース（ＤＢ）化されてきた．この大量の情報が，企業の重要な資産になってきている．しかし，ＤＢのデータ品質（data quality）* の重要性が認識されつつあるにもかかわらず，データ品質についての検討は少ない[1]．

ＤＢ中にデータ不良（data error）が存在することにより，企業活動に対し，以下のような悪影響を及ぼす．

(1) 企業内各種業務のミス／手戻りが発生する[2]．
(2) (1)に基づくお客様信頼の低下或いは苦情の発生を招く．
(3) (1)，(2)により，企業活動がコスト高になる．

ＤＢ中に潜在するデータ不良はシステムの不良の一種と見なされるが，ソフトウェアバグは一般的に単調減少であるのに比べ，データ不良は運用中のＤＢ更新のために必ずしも単調減少とは限らない．　本稿では，このような特徴を持つデータの信頼性に関して，データ信頼性の推定法及びデータ不良混入の要因分析について検討したので，これまでの検討結果を述べる．

2．データ信頼性とは

2．1　データ信頼性の定義

データ信頼性（data reliability）とは，ＤＢ中に存在するデータ不良，具体的には，データ項目値の誤りの度合を示すものであり，データ信頼性の評価尺度としては，データ不良率であり、以下の式で表す．

$$\text{データ不良率} = \frac{\text{データ不良件数}}{\text{全データ項目数}} = \frac{\text{データ不良件数}}{\sum_{\text{レコードタイプ数}} \text{レコード数} * \text{データ項目数}} \quad (2.1)$$

2．2　データ品質評価のための基本参照モデル[6]

まず，ＤＢは，現実世界の情報（ここでは，これを実体と呼ぶ）を写像したものである．

実体とは，例えば，設備に関するＤＢ（以下，設備系ＤＢと称する）では設備そのもの（交換機，線路，電話機，等），顧客に関するＤＢ（以下，顧客系ＤＢと称する）ではお客様注文（オーダ）などである．

ＤＢのデータ品質評価を行う上での基本参照モデルを図１に示す．

(1) ＤＢの成り立ち

ＤＢの成り立ちとしては，

① 実体を直接写像したもの，
② 写像した別ＤＢ（ＣＤ／ＦＤのデータも含む）から流通したもの，

の２種類に大別できる．

なお，図１において，ＤＢへの更新という観点からは，運用中における業務用プログラムからの更新（追加／変更）もある（例えば，ＮＴＴ社内のトラヒック予測システムＴＲＡＹＤ[5]のトラヒック予測データ，等）．

(2) 写像手段

写像の手段としては，以下の２つがある．

① 自動入力（例えば，光線路設備に対し，光ファイバコネクタに組み込んだ設備情報を記憶するＩＤモジュールと，設備情報を読み込み・転送するための通信装置（ＬＣＴＲ）などで，光コネクタの接続状況を取得し，また，ＬＣＴＲ接続のセンサにより，温度・浸水・酸素量等の環境情報を取得して，リアルタイムにＤＢを更新するシステム[2]など）

② 人手を介する入力（通常は，人手入力）

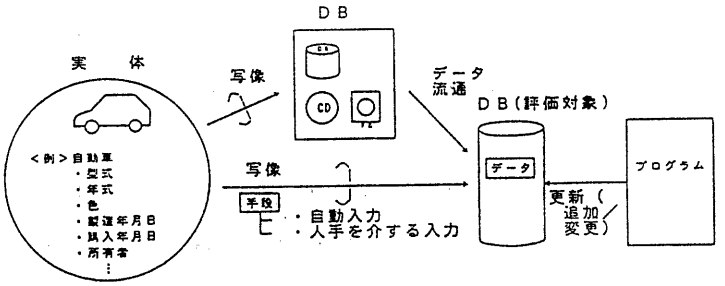

図１　データ品質評価基本参照モデル

Fig.1　Basic reference model for data quality in databases.

2．3　データ信頼性の特徴[6]

データ信頼性に関しては，ソフトウェア，或いは，ハードウェアの場合とは異なり，以下の特徴がある．

① ＤＢは通常，運用中に頻繁に更新（レコードの追加／データ項目値の変更／レコードの削除）され，それに伴い，データの不良も増加する傾向にあり（後述の図３のカーブ１），無視できない．

② ＤＢの個々のデータは，データ毎に色々なプロセスを経て生成／更新される．その各プロセスでは人やプログラムなどが係わるため，各局面でデータ不良が起きる可能性がある．

＊　本稿では，「データ品質」とは，狭義の意味での「データ信頼性」を指す．

3．データ信頼性向上フロー

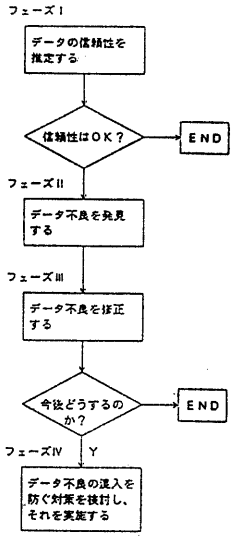

図2　データ信頼性向上フロー

Fig.2　Data reliability improvement flow.

我々が採用するデータ信頼性向上フローを，図2に示す．

［フェーズI］　データ信頼性の推定

データの信頼性を推定し，推定した信頼性が満足できない場合は，信頼性を向上させるために，フェーズII～IVを実施することになる．

ところが，DBのデータ信頼性については，DB中の全レコードについてその状態を把握しなければならないことから大変であり，ほとんど把握されていないのが現状であり，データ信頼性の把握のためのモデルが重要となる．データ信頼性の推定法については，4章で述べる．

［フェーズII］　データ不良の発見

本フェーズは，他フェーズに比べ，一番工数がかかっているところであり，特に発見後の正解値を得る作業に工数がかかる．データ不良の検出方法については5章で述べる．

［フェーズIII］　データ不良の修正

フェーズIIで求めた正解値をもとに，該データ不良を修正する．

［フェーズIV］　データ不良混入防止対策の検討・実施

新たなデータ不良混入を防止するための対策について検討し，それを実施する．具体的には6章で述べる．

4．　データ信頼性の推定法

4．1　データ信頼度成長曲線の考え方

ソフトウェアにおいては，従来からソフトウェアエラー発見過程やソフトウェア故障発生現象をモデル化したソフトウェアの信頼度成長モデル（software reliability growth model）が研究され，ソフトウェアの品質管理に利用されてきた（[3]，[4]など）．本稿では，ソフトウェアのモデルをベースにして，ソフトウェアとデータとの違いに着目したデータ信頼度成長モデル（data reliability growth model）を考えた．

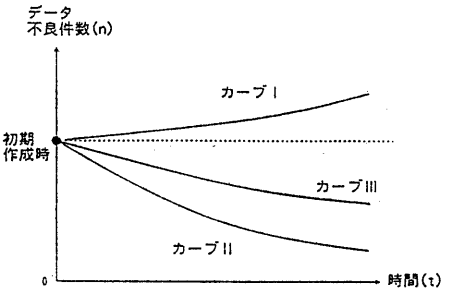

図3　データ信頼度成長曲線

Fig.3　Data reliability growth curve.

カーブI：運用中に何等の品質改善を行わない場合

カーブII：運用中に何等かの品質改善活動や運用中に発見される不良を修正する場合

カーブIII：カーブI　＋　カーブII

図3は，DBにおけるデータの信頼度成長曲線（reliability growth curve）の概念図である．

ソフトウェアの信頼度成長曲線では，発見した累積ソフトウェアバグ数で表すのに比べ，本稿でのデータ信頼度成長曲線では，各時点におけるデータ不良件数で表したのが，両者での大きな違いである．ここで，各々のカーブは以下の意味を持つ．

(1) カーブI

DBの運用中に何等の品質改善を行わない場合，データの品質はカーブIの曲線をたどる．レコード追加及びデータ項目値の修正に伴い，新たなデータ不良が発生することを意味している．

(2) カーブII

運用中に何等かの品質改善活動や運用中に発見されるデータ不良を修正すると，カーブIIの曲線を推移する．

(3) カーブIII

実際のDBシステムでは，(1)，(2)の両方のことが発生し，データの信頼度成長曲線は，カーブIIIのようになる．

ここでは，仮に，カーブIを"新規分のデータ不良（件数）"と呼び，カーブIIを"旧分あるいは既存分のデータ不良（件数）"と呼ぶことにする．

4．2　モデル式

ここで，データ信頼度成長モデルの式を，以下で表す．

$$① \quad y = F(x) = y_1 + y_2 \qquad (6.2.1)$$

但し，y　：（トータルな）データ不良件数
x　：経時的な変化（年あるいは月）
$F(x)$：xを変数（説明要因）とする関数
y_1　：新規分のデータ不良件数
y_2　：旧分のデータ不良件数

同様にカーブⅠ，カーブⅡの式を，以下で表す．

② $y_1 = F_1(x)$ (6.2.2)

③ $y_2 = F_2(x)$ (6.2.3)

4．3　対象ＤＢと測定の条件

(1) 評価対象ＤＢはある地域における顧客系ＤＢ，

①ＤＢ-A
②ＤＢ-B

の２システムであり，データ件数は各々約４０万件である．ＤＢ－Aは実体を直接写像したものであり，また，ＤＢ－Bは，その一部のデータ項目に関しては，ＤＢ－Aからデータ流通したものである．なお，データ不良の修正に関しては，流通ルートによる場合と，各々単独に修正される場合とがあった．

(2) データ信頼度の測定に当たっては，各ＤＢ間のデータ項目間の突き合わせ（突合２）を実施したとしても，最終的には，実体との突き合わせが必要となるなど，多大な人手と時間を要するものであるから，対象ＤＢに関しては現状１回／年実施していた．なお，測定ポイントが毎年同一月ではないため，経時的な変化の単位としては，月単位で表している．

(3) データ不良のうち，品質測定（／品質改善活動実施）ポイント間で，運用中に発見され，修正されているものについてのデータ不良件数については不明であり，それについては無視することにする．

4．4　モデル式のあてはめ結果

測定ポイントは，平成２年～４年の３ポイント（品質改善を実施）であるが，それを最小２乗法で，直線式及び８つの曲線式（分数／ルート／自然対数／べき乗／指数／修正指数／ロジステック／ゴンペルツ曲線）を当てはめて，あてはめの良い関数式を選択する．

(1) ＤＢ－Aのデータ信頼度成長モデル式の選択

結果を表１に示す．あてはめの良い関数式としては，修正指数曲線とロジスティック曲線があったが，分散比（Ｆ）の一番大きい修正指数曲線とする．

(2) ＤＢ－Bのデータ信頼度成長モデル式の算出

結果を表２に示す（但し，この場合は，変数がx^2になっている）．あてはめの良い関数式としては，修正指数曲線，ロジスティック曲線，ゴンペルツ曲線があったが，分散比（Ｆ）の一番大きい修正指数曲線とする．

求めた信頼度成長曲線（カーブⅢ）を，図４，図５に示す．また，カーブⅠ，Ⅱ，Ⅲの式をまとめて表３に示す．

表１　ＤＢ－A の信頼度成長曲線（カーブⅢ）の選択

関数式		決定係数 R^2	分散比 F	F検定による有意差判定
直線	$y=ax+b$	0．9454	17．30	
分数	$y=a(1/x)+b$	0．8114	4．30	
ルート	$y=a\sqrt{x}+b$	0．9200	11．49	
自然対数	$y=a\log x+b$	0．8887	7．98	
べき乗	$y=ax^b$	0．5230	1．09	
指数	$y=ab^x$	0．6877	2．20	
修正指数	$y=K-ab^x$	0．9993	1503．49	*
ロジスティック	$y=\frac{K}{1+ae^{-bx}}$	0．9858	69．20	
ゴンペルツ	$y=K\cdot a^{b^x}$	0．9957	233．61	*

＊＊　：　有意水準　1％　で有意といえる
＊　：　有意水準　5％　で有意といえる
無印　：　有意水準　5％　で有意といえない

表２　ＤＢ－B の信頼度成長曲線（カーブⅢ）の選択

関数式		決定係数 R^2	分散比 F	F検定による有意差判定
直線	$y=ax+b$	0．9651	27．623	
分数	$y=a(1/x)+b$	0．6846	2．170	
ルート	$y=a\sqrt{x}+b$	0．9241	12．181	
自然対数	$y=a\log x+b$	0．8601	6．149	
べき乗	$y=ax^b$	0．6315	1．713	
指数	$y=ab^x$	0．8675	6．546	
修正指数	$y=K-ab^x$	0．9989	924．322	*
ロジスティック	$y=\frac{K}{1+ae^{-bx}}$	0．9948	190．683	*
ゴンペルツ	$y=K\cdot a^{b^x}$	0．9976	412．918	*

＊＊　：　有意水準　1％　で有意といえる
＊　：　有意水準　5％　で有意といえる
無印　：　有意水準　5％　で有意といえない

但し、$x=\chi^2$とする

4．5　考察

考察を以下にまとめる．

(1) 係数に関する考察

① 係数Kは，理論的にはＤＢ初期作成時のデータ不良件数（初期値）と考えられ，それは，該当ＤＢが最初にどのようにして作成されたものかに依存するものと考えられるが，初期作成時のデータ不良件数は実測されていないため，推定でしかない．

また，ＤＢ－BのKの方がＤＢ－Aのそれより小さい．これは，ＤＢ－Bが料金に関係するＤＢであり，重要なものであるので，信頼性が元々高いものと判断される．

② 係数ａ及びｂに関しては，

・ＤＢのデータ品質改善活動における稼働のかけかたのレベル（全データに対して実施する／部分的に行う，また，修正を全て行う／重要なものだけ修正する，など）
・日常的なチェック体制のレベル（データ投入までのチェック体制，オーダ投入から料金計算などの業務を一貫して実施するなどの体制，など）
・担当者の意欲／意識などのレベル

などによる改善効果として決まってくるものと推察される．

なお，各係数の決め方あるいは係数値の大小関係の詳細な分析については，今後データ数を増やして検証していく必要がある．

(2) 新たなデータ不良に関して

評価対象ＤＢに関しては，ＤＢにデータ入力する時点までのチェック体制として，繰り返しチェックが行われていたが，それでもなおかつ新たなデータ不良の混入が避けられない状況を示している．

(3) 検出したデータ不良の修正について

品質改善活動などにより，検出したデータ不良が完全に修正されたならば、残りは新規分のデータ不良のみとなるはずであるが，実際には，全て修正されてはいなかったということである．

この理由としては，

① 修正すべきデータが，ＤＢ間にまたがった重複データの場合，全てのＤＢ間を同期して修正する必要があること，
② 正解値の把握に当たっては，最終的には実体との確認が必要となること，

など，慎重な配慮が望まれるため，プライオリティを付けて修正が実施されたためであった．具体的な事例としては，重要なデータ不良，例えばお金にかかわるようなデータ項目は必ず修正されているが，重要でないデータ不良，特に，お金に直結しないデータ項目であったり，同じ原因のデータ不良（投入漏れ等）がかなりの件数あるが，その内容を認識しているということで，そのままとしている場合などがあった．

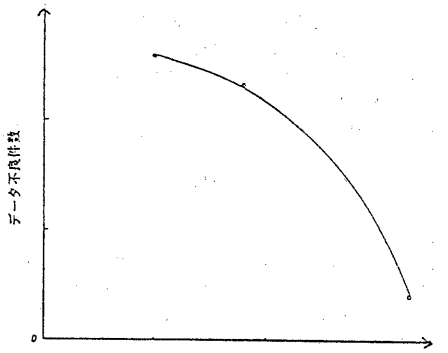

図4　ＤＢ－Ａのデータ信頼度成長曲線
Fig. 4　Data reliability growth curve of ＤＢ－Ａ.

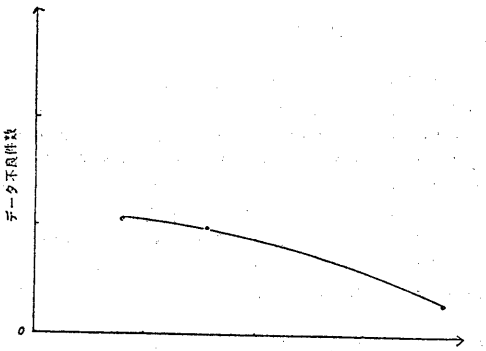

図5　ＤＢ－Ｂのデータ信頼度成長曲線
Fig. 5　Data reliability growth curve of ＤＢ－Ｂ.

表3　モデル式のあてはめ結果

	データベース DB－A	データベース DB－B
信頼度成長曲線 カーブIII	xを変数とする修正指数曲線 $y=K-a*b^{x}$ 但し、$b>1$ （5％で有意）	x^{2}を変数とする修正指数曲線 $y=K-a*b^{x^{2}}$ 但し、$b>1$ （5％で有意）
カーブI	x^{2}を変数とする指数曲線 $y=a*b^{x^{2}}$ 但し、$b>1$ （1％で有意）	x^{3}を変数とする分数曲線 $y=a(1/x^{3})+b$ 但し、$a<0$ （5％で有意）
カーブII	xを変数とする修正指数曲線 $y=K-a*b^{x}$ 但し、$b>1$ （1％で有意）	xを変数とするゴンペルツ曲線 $y=K*a^{b^{x}}$ 但し、$b>1$ （1％で有意）

5．　データ不良検出法

データ不良の検出方法としては，

① ＤＢと実体との突き合わせ（突合１）
② ＤＢ間での突き合わせ（突合２）
③ １ＤＢ内レコード間，或いはレコード内のデータ項目間の突き合わせ（突合３）
④ レコード内のデータ項目の値の範囲チェック（突合４）
⑤ データ項目の値の異常値を検出（突合５）

に分類でき，その方法及び特徴を，図６及び表４に示す．

ＤＢ間での突き合わせ（突合２）による方法は，データベース品質管理システムＤＱＳ[8]などで採用されているものである．また，突合２～５による方法では，最終的には実体との確認を行う必要がある．

ここで，通常使用されている方法としては、突合１と突合２が主と想定される．（なお，今回のモデル式算定に用いた測定結果も突合１と突合２によるものである．）

6．データ不良混入の機会と防止法

6．1　データ品質劣化の要因分析[6]

ＤＢのデータ品質を向上させるには，データが生成／更新される全てのプロセスを明らかにし，各々のデータ不良が発生したプロセスを特定し，データ不良の防止策を検討することが必要である．そのため，データ品質劣化要因分析用のモデル（model to analyze the factors deteriorating data quality）を作ることが重要となる．

実体の変化に対して，ＤＢは適正なタイムラグの許容範囲内で追従していく（デッドライン制御）必要があり，この追従のメカニズムとしては，

① 実体の変化を検出する検出メカニズム
② 実体の変化をＤＢに反映する反映メカニズム
③ 何等かのデータ誤りに伴う修正メカニズム
④ 業務用プログラムによる運用中での更新メカニズム

に分類でき，これらのメカニズム中の全ての処理（プロセス）において，データ不良が混入する可能性がある．

データ品質劣化要因分析モデル（図７）に基づき，顧客系ＤＢの２システムに関するデータ不良要因を分析した結果，データ不良は，全てヒューマン・エラー[7]であり，その内訳としては，

① （お客様注文等の）聞き取り誤り/記入誤り/確認誤り/連絡漏れ
② ＤＢへの投入時誤り／漏れ
③ データ不良検出後の修正漏れ／誤り

であった．

6．2　データ入力系のインテリジェント化について

２．２節で述べたように，設備系ＤＢのように，設備を扱うＤＢでは，設備自体に，実体変化を検出するハードウェア機構を付加することにより，ＤＢへの自動入力（セル

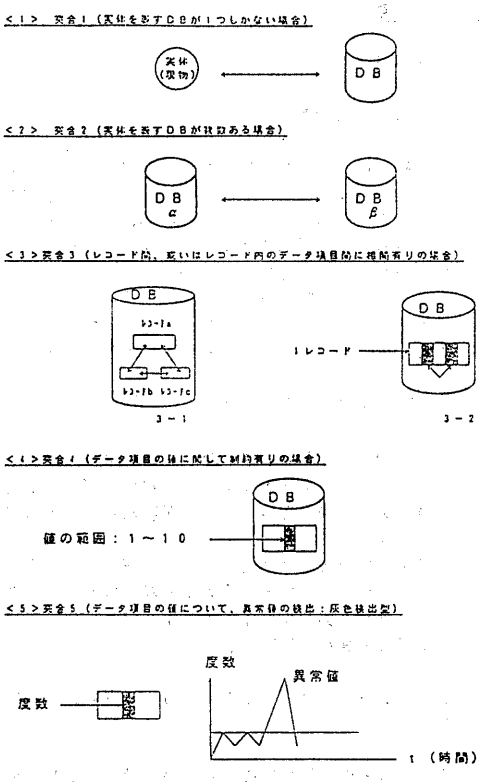

図６　データ不良検出方法
Fig. 6　Data error detecting method.

表４　データ不良検出方法

方法	説明	工数	適用方法
突合１	ＤＢと実体（現物）との突き合わせ	大 ・実体の把握に膨大な工数がかかるから	・実体を表すＤＢが１つしかない場合
突合２	ＤＢ間での突き合わせ	中 ・表現方法が異なるから ・センタが別にあるため，レコードの取り出し／移動が伴うから	・実体を表すＤＢが複数ある場合で，かつＤＢの成り立ちが違う場合
突合３	１ＤＢ内レコード間，或いは，レコード内のデータ項目間の突き合わせ	小	・レコード間に相関がある場合，或いは，レコード内のデータ項目間に相関がある場合 <例１>データ項目間の組合せチェック ①項目Ａ　～　項目Ｂ　～　項目Ｃ
突合４	（１ＤＢ内）レコード内のデータ項目の値の範囲チェック	小	・レコード内のデータ項目の値に対して制約がある場合 <例１> ①商品ｺｰﾄﾞ，住所ｺｰﾄﾞ，等
突合５	・データ項目の値について異常値を検出（例，分布の３σ以外を検出した場合） ・突合３の場合で，取りうる値の組合せが極めて珍しい等の場合） （反色検出型）	小	・レコード内のデータ項目の値に対して，時系列データの分析がある場合 ・レコード間に相関がある場合，或いは，レコード内のデータ項目間に相関がある場合

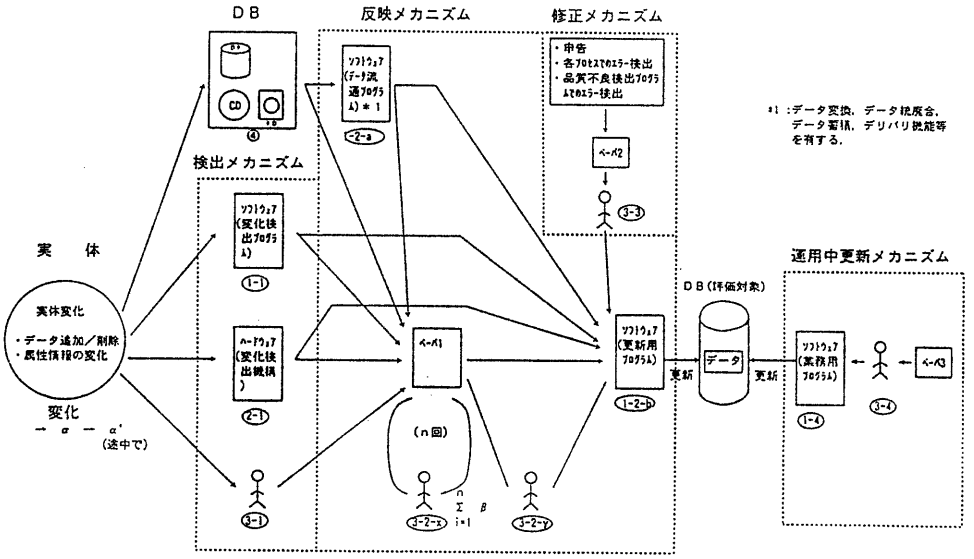

図7　データ品質劣化要因分析モデル
Fig. 7　Model to analyze the factors deteriorating data quality.

フ・インベントリ化）が可能な場合がある[2]が，お客様を対象とする顧客系ＤＢでは，対お客様であり，ヒューマン・エラーは避けられない．

そこで，新たなデータ不良混入を防止するための対策として，顧客系ＤＢにおけるデータ入力系のインテリジェント化について考察する．

(1) ソフトウェアの場合と同様に，データについても，後になればなる程、正解値を得るための手間がかかり，修正困難となっているから，新たなデータ不良混入を防ぐための対策としては，お客様注文の受付時に、リアルタイムに正しさをチェックする機構が必要である．

(2) 上記(1)の方法で，窓口でのお客様注文の受付時，厳密にチェックする方法も考えられるが，それを実施すると，チェックのためにターンアラウンドタイム（ＴＡＴ）が長くなり，業務処理が回らなくなる可能性がある．適度なＴＡＴを保証するために，どこまでのレベルまでをチェックすべきかということが重要となる．

新たなデータ不良混入を防ぐための，「入力系のインテリジェント化」については，いまだ解決されたわけではない．

7．おわりに

本稿では，データ信頼性の推定法，データ不良混入の原因分析などについて報告した．以下に，今後の課題をまとめる．

(1) データ信頼度成長モデル
本稿で示した経時的な変化に基づくデータ信頼度成長モデルにより，データ信頼性の推移状況の把握が可能となったと考えるが，予測式として，次年度以降の傾向予測に使えるかどうかの検証は，今後の課題である．また，少ない工数で推定するために，サンプリング結果などによる方法で精度良く推定する方法についても検討していく必要がある．

(2) データ不良混入防止策
本稿で示したデータ品質劣化要因分析モデルにより，データ不良の混入プロセスを特定することは可能となったが，新たなデータ不良混入の防止策については，今後の課題である．

参考文献

[1]Huh Y U et al. : Data quality, Information and software technology, Vol. 32, No. 8, pp. 559-565(1990).
[2]山下他：リアルタイム光線路設備管理システム，ＮＴＴＲ＆Ｄ，Vol. 41, No. 10, pp. 1233-1242(1992).
[3]山田：ソフトウェアの品質評価に関する考え方と動向 ソフトウェア信頼度成長モデルに基づく定量的品質評価法，情報処理，Vol. 32, No. 11, pp. 1189-1202 (1991).
[4]山田，高橋：ソフトウェアマネジメントモデル入門 － ソフトウェア品質の可視化と評価法，共立出版(1993).
[5]原田：明日のネットワークを築くトラヒック予測システム（ＴＲＡＹＤ），ＮＴＴ技術ジャーナル，Vol. 4, No. 8, pp. 86-87(1992).
[6]中村他：データベース品質劣化要因分析モデルの提案，情処学会第45回全国大会，6P-4(1992).
[7]菅野：ヒューマン・エラーのメカニズム，日科技連(1980).
[8]田中他：データベース品質向上を支援する新システム － データベース品質管理システム（ＤＱＳ）の開発 －，ＮＴＴ技術ジャーナル，Vol. 4, No. 11, PP. 67-68(1992).