まず検討されるべき対象となるイベントの性質であるが，利用できるデータの性質を勘案して最適な分析手法を用いることが重要となってくる。山口の「イベントヒストリー分析」『統計』(2002-2003, [1]-[15]) の連載では，離散時間ロジットモデル[4]・[5]，等比ハザードモデルの拡張 1：時間区分定率モデル（piecewise constant rate model）[6]，等比ハザードモデルの拡張 2：Cox モデル[7]・[8]，加速時間モデル（accelerated failure-time model）[12]，ネステッドロジットモデル（nested logit model）[14]・[15]が詳細に紹介されている。本稿では，汎用性が高い離散時間ロジットモデルと等比ハザードモデル，その拡張版である Cox モデルについてまとめる。

それぞれの分析手法の特徴をまとめる前に，イベントヒストリー分析に共通するいくつかの理論的概念を以下に示す。

- **ハザード率**：ある時点のイベントのハザード率は，
  ➔ (a) ある個人について，イベントがその時点より前に起こらなかった場合の，
  (b) その時点でそのケースにイベントが生起する確率を示す

- **基底時間**： リスク開始時間からの継続時間
  ex.) 離婚 → 基底時間＝初婚期間

  ※ リスク開始からイベントの生起までの経過時間自体の影響を考慮できる

- **基底ハザード関数**：独立変数の影響が全くないとき（経過時間の影響しかないとき）のハザード率の関数

  → ハザード確率のオッズが，時間とともにどのように変化するかを特徴づける
  → 年齢が基底時間の場合 ex.) 初婚・初出産・死亡など
  イベントの生起率の年齢によって変化するパターンを表す
  → ある事象からの継続時間が基底時間の場合 ex.) 離婚・離職など
  ある事象から何年(月)後にイベントが生起するかというパターンを表す

- **比例ハザード性**：推定結果で算出されるハザード率の比は，どの時点でも一定であるという仮定のもとで算出させる性質のこと （山口 2002[3])

4－1．離散時間ロジットモデル（山口 2002[4]・[5]）

このモデルはデータの時間の単位が「年」で集計される離散時間モデルとして適用できるモデルである。データ形式はパーソン・ピリオド・データ形式を用い，推定法は最尤（推定）法（maximum likelihood estimation）である。最尤法は，モデルの推定に対し最も発生確率が高まるような母数（パラメタ：parameter）の組合せを選択する（＝「尤度」を最大にする）方法である。尤度とは，イベントが生起した場合，そのイベント生起時でのイベント生起確率を示し，イベントが起こらず右センサーされたときは，センサーされたときまでの生存確率（survival probability：その時点より前にはイベントが起こらないという確率）を掛け合わせたものである（山口 2002[3]）。利点は，(1) ロジスティック回帰分析を利用できる（パーソン・ピリオド・データを用いることで，ロジスティック回帰分析のモジュールによって離散時間ロジットモデルと同様の結果が得られるという意味），(2) 時間依存の独立変数をモデルに組み込みやすい，(3) ハザード率の時間分布を仮定しなくてよいといったものがある。欠点は，(1) パーソン・ピリオド・データの作成が難解である，(2) データのレコード数が多くなるといったものがある。参考として，連続時間を仮定できる場合にこのモデルを適用すると，Cox モデルと同様の結果が得られるという性質がある。

4－2．等比ハザードモデル（山口 2002[6]）

このモデルはデータの時間の単位が「月」で集計される連続時間モデルとして適用できるモデルである。等比ハザードモデルにはいくつかの種類があり，基底ハザード関数を使用しない Cox モデルと基底ハザード関数を使用するパラメトリックモデルがある。Cox モデルは次項でまとめるが，パラメトリックモデルは指数分布・ワイブル（Weibull）分布・ゴンパーツ（Gomperz）分布など理論分布が仮定され，最尤法によって推定されるモデルである。しかし「筆者（注：山口）の経験ではパラメトリックなモデルがデータに適合するのは稀」であるため，「特に分布の仮定を置かず階段変数やスプライン関数などで近似する」セミパラメトリックなモデルが「望ましい」という（pp.65）。セミパラメトリックには，一般的に与えられた時間区分内でハザード率は一定で時間区分間でのみ変わると仮定する時間区分別定率モデル（piecewise constant rate model）が多用される。利点は，「離散時間ロジットモデルに比べ，回帰係数のイベント生起率を表すハザード率の比という，よりなじみのある数量で解釈ができる」（pp.70）。

4－3．Cox モデル（山口 2002[7]・[8]）

このモデルはデータの時間の単位が「月」で集計される連続時間モデルとして適用できるモデルである。他のモデルとの差異は，母数（パラメター）の推定に対し部分最尤（推

定）法（partial likelihood estimation）を用いる点にある。部分最尤法は，規定ハザード関数の影響を統制するような尤度の算出をもたらす方法である。「イベントがいつ起こりやすいか，起こりにくいか」というハザード率の時間的変化を示す基底ハザード関数の影響を弱くし，「ある状態では他の状態よりも何倍くらいイベントのリスクが高いか」といった「相対的リスク」（relative risk）を表すことができる。相対的リスクはロジスティック回帰分析における（対数）オッズ比に似た概念であるが，(1) センサーされた観測値の情報を偏りなく取り入れられる，(2) 時間とともに変化する独立変数を用いることができるというイベントヒストリー分析の長所を失うことなく，適用できる点に画期性がある。とはいえ，利用上の制約はある。第 1 に，標本数が少ないと最尤法に比べ推定値の精度と正規性の精度が著しく低下する。第 2 に，Cox モデルはイベントの生起が同時であるとみなされる組合せの割合が多くなると問題が生じることから，第 1 の制約とは逆に，標本数が大きすぎると推定が困難になるというものである。イベントの発生単位を小さく分割（年→月単位）することや分析手法選択に関し最尤法を利用するものに変更することなどでこの問題を回避する必要がある。第 3 に，基底ハザード関数の推定を削除した分析であることから，独立変数と時間変化に交互作用があるとき，その影響の解釈にあいまいさを残すことになる。利点は，(1) ハザード率の時間分布を仮定しなくてもよい，(2) 若干数であれば時間とともに変化する独立変数をモデルに組み込むことができる，(3) 連続時間モデルを仮定しているため，情報のロスが少ないといった点があげられる。

5．結果の記述

統計ソフトなどでイベントヒストリー分析によって得られる分析結果は，ロジスティック回帰分析と同様に，回帰係数 B と指数変換された回帰係数 EXP(B)という形で出力される場合が一般的である。回帰係数 B に関しては，「他の変数が一定のときの，独立変数 Xi が 1 単位増加したときの Y の増加量」を表すが，指数変換された回帰係数 EXP(B)に関しては，ロジスティック回帰分析では，確率の（対数）オッズ比であったのに対して，離散時間ロジットモデルではハザード確率のオッズ比を表し，等比ハザードモデルではハザードの比として示される。そして，それらの値はどの時点においても一定であるという比例ハザード性の仮定のもとで算出される（交互作用がある場合は，その限りではない）。

次に独立変数の係数に関して，独立変数の変化量に対するハザード率の増加率の推定方法については以下のとおりである。

- 1 変量の変化によるハザード率の増加率の推定　$\Delta r = (EXP(B_i)^n - 1)*100$
  r：予測されるハザード率の上昇，B_i：変数 i の係数，n：変数 i の増加単位
- 2 変量の変化によるハザード率の増加率の推定　$\Delta r = (EXP(B_i)^n * EXP(B_j) - 1)*100$
  B_j：変数 j の係数，k：変数 j の増加単位

最後に，イベントヒステリーデータを用いた「統計表表現とその 2 次分析」(山口 2002[9]) についてまとめる。イベントヒストリー分析によって提示できる記述統計量はサバイバル確率とハザード確率である。「基底時間をカテゴリー化した変数を含む幾つかのカテゴリー変数をクロスさせた組合せの各セルに対応するサバイバル確率やハザード確率を特定のモデルを用いずに推定する」(山口 2002[9]) である。サバイバル確率はハザード確率と異なり，クロス統計は基底時間カテゴリー以外の時間に依存する変数とクロスさせることはできないという制約がある。

サバイバル確率についてのノンパラメトリックな推定方法には，カプラン・マイヤー法 (Kaplan-Meier Method) があり，「センサー時とイベント時との同時値の取り扱いについての仮定以外は，特別な仮定を何も置かない推定法」(山口 2002[9]) である。ハザード率については，サバイバル関数とは異なり単調減少関数ではないので，技術的な困難がある。ただし，累積ハザード関数を用いたノンパラメトリックな推定については，ピーターセン法やネルソン法などがある。

6．「21 世紀成年者縦断調査」を用いたイベントヒストリー分析の一例

ここでは，イベントヒストリー分析を「21 世紀成年者縦断調査」を用いて分析する。分析テーマは「第 2 子出生タイミングに対する社会経済的地位の影響」とする。使用するデータは「21 世紀成年者縦断調査」の第 1-2 回連続標本 (ar02.dat, 15rireki-data-sample) である。

分析対象者は第 1 子を有する女性に限定し (4,536 ケース)，第 2 子を持ったかどうかの妊娠履歴 (さらに，第 1 子出生年月が結婚年月よりも早いケースを除いた 4,376 ケース) に，出生コーホート，第 1 子出生時年齢，学歴，初職 (学卒後はじめて就いた仕事) の推定値を得るシンプルなモデルを作成した。

分析手法は，Cox モデルを用いる。作成したデータ形式は，期間データである。リスクの開始時期は第 1 子出生時点を用い，リスクの終了はイベントの発生，すなわち第 2 子出生時点とイベントが起きずにリスク期間を終えた右センサー時点とした。右センサーは分析対象者の内，第 2 子出生時年齢が最年長である 43 歳時とする。調査時点において 43 歳未満で，イベントが発生しないケースについては，調査時年齢が右センサーとなる。

分析モデルは 2 つである。モデル 1 は第 1 子出生時年齢を実数で投入し，モデル 2 では第 1 子出生時年齢を 5 歳階級カテゴリーで投入し，「25-29 歳」をリファレンスカテゴリーとした相対確率を推定している。出生コーホートは，「190-74 年」生まれをリファレンスカテゴリーとし，その他を 5 年階級別カテゴリーで投入している。学歴は，「大学」と「大学院」を選択したケースを高等教育とした高等教育ダミーを作成している。また，在学中が 10 ケースあったことから，在学中ダミーを作成し統制変数として用いている。初職は，9 カテゴリーあった職業分類を「会社役員・自営業主」，「自営業」，「正規の職員」，「パート・

アルバイト」,「派遣・契約・嘱託職員」にまとめ，現在職業及び職業履歴で非該当であったケースと在学中のケースを「無職・学生」に値の再割り当てを行っている。分析モデル

表 4　Cox モデルによる第 2 子出生タイミングの推定結果

説明変数	モデル1		モデル2		ロジスティック回帰	
	B	EXP(B)	B	EXP(B)	B	EXP(B)
【出生コーホート】						
1958-64年	0.680 **	1.973	0.830 **	2.294	2.359 **	10.581
1965-69年	0.443 **	1.558	0.447 **	1.564	1.293 **	3.645
1970-74年†						
1975-79年	-0.783 **	0.457	-0.747 **	0.474	-1.538 **	0.215
1980-84年	-2.077 **	0.125	-1.805 **	0.164	-3.119 **	0.044
【第1子出生時年齢】						
・実数	-0.137 **	0.872				
・16-19歳			0.701 **	2.015	1.979 **	7.239
20-24歳			0.614 **	1.849	1.429 **	4.176
25-29歳†						
30-34歳			-1.194 **	0.303	-1.945 **	0.143
35-39歳			-1.209 *	0.298	-2.705 **	0.067
【学歴】						
高等教育ダミー	0.026	1.026	-0.024	0.976	-0.014	0.986
在学ダミー	-0.010	0.990	-0.009	0.991	0.038	1.039
【初職】						
会社役員・自営業主	0.170	1.186	0.136	1.145	0.222	1.248
自営業	-0.104	0.901	-0.031	0.969	0.019	1.019
正規の職員†						
パート・アルバイト	-0.113 +	0.893	-0.017	0.983	-0.086	0.917
派遣・契約・嘱託職員	0.079	1.082	0.136	1.145	0.184	1.202
無職・学生	0.209 *	1.233	0.299 **	1.349	0.520 **	1.681
定数					-0.047	0.954
イベント	2451		2451		2451	
センサー	1413		1413		1413	
全体	3864		3864		3864	
-2Log Likelihood	37816.58		37831.33		4258.317	
χ2乗値	700.7		633.4		872.8	
自由度	12		12		15	

** p < .001, * p < .05, + p < .10

におけるケース数は，全 3,864 ケースの内，イベントが発生したのは 2,451 ケース，イベントが発生せず右センサーされたのは 1,413 ケースである。また，モデル 2 と同じ変数構成のモデルでロジスティック回帰分析結果を行った。推定結果は表 4 の通りである。

推定結果について，出生コーホートはモデル 1・2 ともに「1970-74 年」生まれコーホートに比べて，以前のコーホートでは正の効果，そして相対確率は「1970-74 年」の発生確率を 1 とした場合，「1958-64 年」はおよそ 2 倍，「1965-69 年」はおよそ 1.5 倍の相対確率を示し，以後のコーホートでは負の効果，「1975-79 年」はおよそ 1/2，「1980-84 年」は 1/10 の相対確率を示していることより，「1970-74 年」コーホートに比べ最近のコーホートほど

第 2 子出生の発生確率が低いことを示している。第 1 子出生時年齢も同様の結果がみられる。ただし,「1975-79 年」・「1980-84 年」は一般的に,産み始めの時期であり第 1 子出生が多く,加齢効果の影響も考慮に入れる必要がある。

学歴については,統計学的に有意な結果はみられなかった。これは予備推定において,「高校」をリファレンスカテゴリーにおいた相対確率においても同様の結果であった。初職については,「正規の職員」に比べ,モデル 1 では「パート・アルバイト」である場合,負の効果を示し,「無職・学生」である場合,正の効果を示している。モデル 2 では「無職・学生」の場合,正の効果がみられた。

モデル 2 とロジスティック回帰分析との比較においては,出生コーホートと第 1 子出生時年齢でロジスティック回帰分析の推定値が高めに推定されている。これは,ロジスティック回帰分析ではタイミング効果を考慮することができないため,加齢効果が推定値に大きく影響を与えているものと考えられる。学歴と初職については,ほぼ同水準であるといえる。

最後に,特定のモデルを用いない生存関数(サバイバル確率)の推定を行う。ここでは,一般的なカプラン・マイヤー法を用いた。図 3 から図 6 に生存関数の分布を示した。分布から得られるカテゴリーごとの差は Cox モデルによる推定結果に符合する。

7. おわりに

本稿では,パネル調査データのミクロ分析として有力な分析手法であるイベントヒストリー分析を山口一男「イベントヒストリー分析」『統計』(2002-2003, [1]-[15]) を通じてレビューしてきた。解説の手引きとして,21 世紀成年者調査や 21 世紀出生児調査を用いた分析を提示したが,両調査はまだ開始して間もないことから年次のデータの蓄積が多くなく,パネル調査の長所である各年の属性の変化を活かしきるにいたらなかった。とはいえ,今後継続的にデータの蓄積が可能になるにつれ,分析対象となるイベントのケース数が増加することによって,分析可能範囲は拡がることは必至である。

参考文献

Allison, Paul D. 1984."Event History Analysis-Regression for Longitudinal Event Data, Sage Publications, Inc.

大橋泰雄・浜田知久馬, 1995.『生存時間解析 SAS による生物統計』東京大学出版会.

山口一男, 2002-2003.「イベントヒストリー分析(1)〜(14)」『統計』52(9)〜53(11).

図３　カプラン・マイヤー法による出生コーホートの生存関数（サバイバル関数）分布

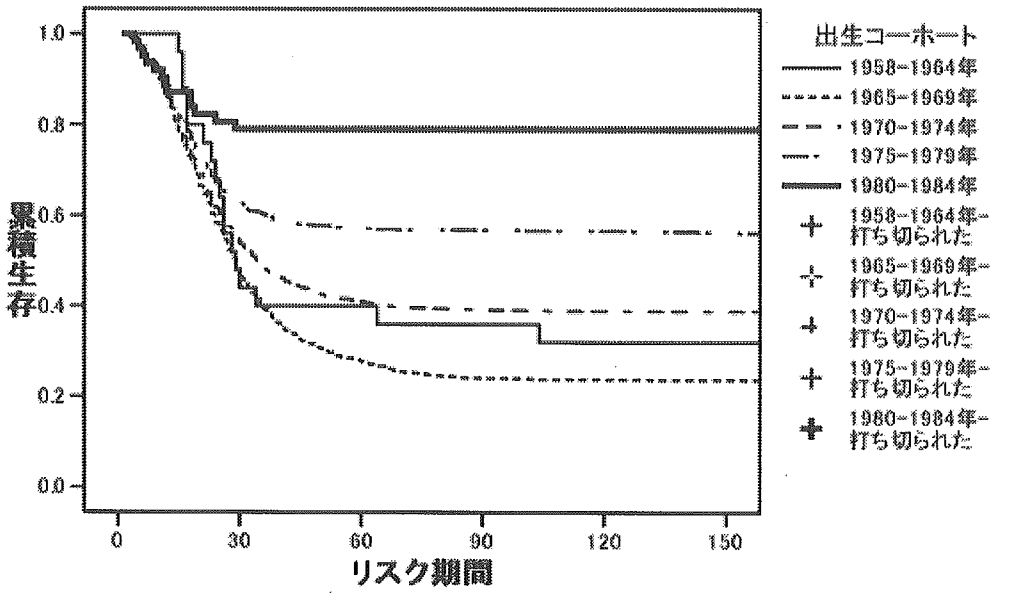


図４　カプラン・マイヤー法による第１子出生時年齢の生存関数（サバイバル関数）分布

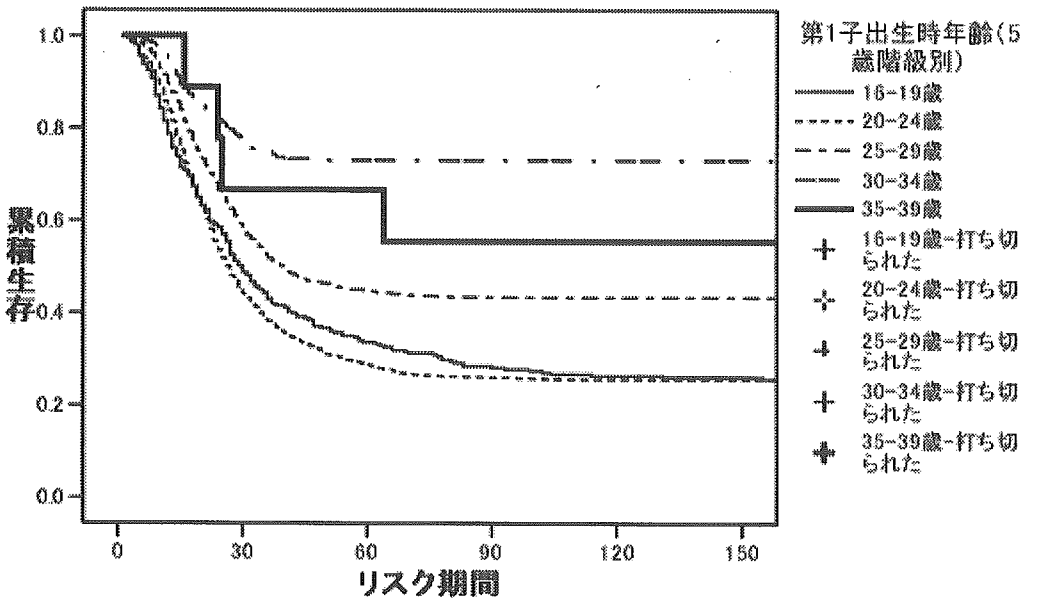


図５　カプラン・マイヤー法による学歴の生存関数（サバイバル関数）分布

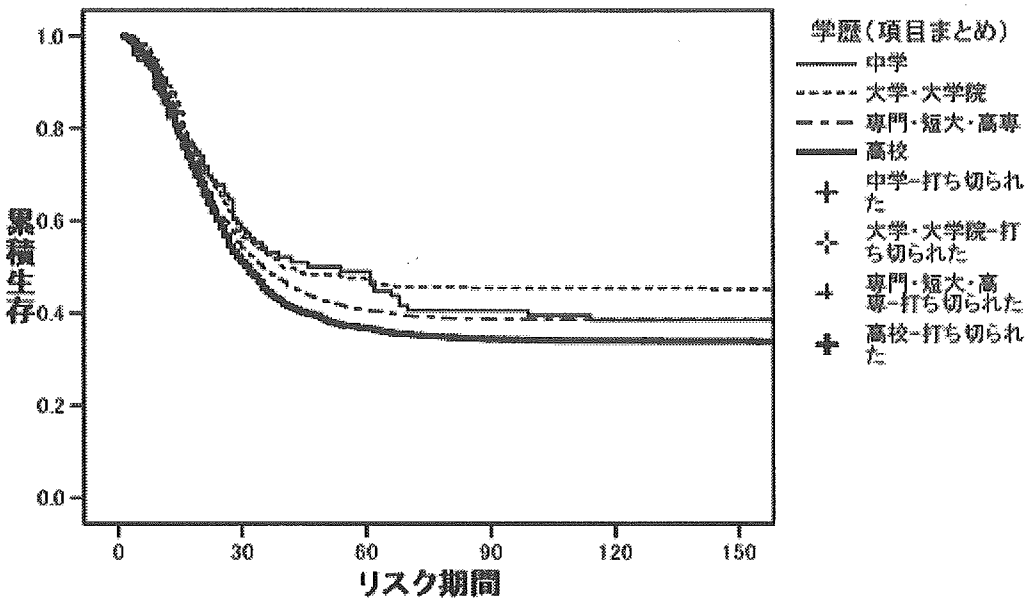
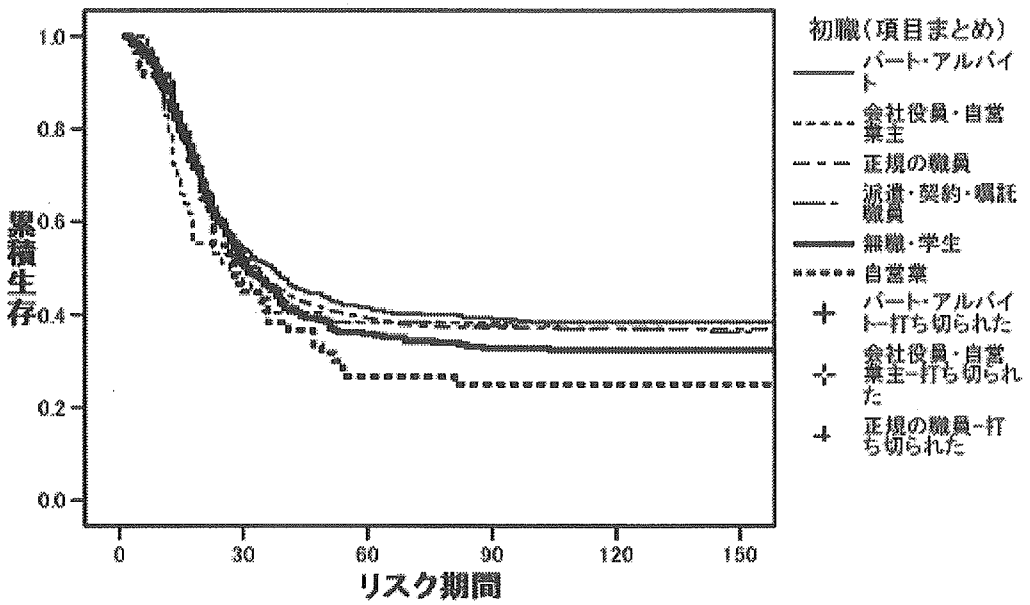


図６　カプラン・マイヤー法による初職の生存関数（サバイバル関数）分布



4　縦断調査マイクロシミュレーション分析の基礎システムの開発

金子 隆一

1．はじめに

パネル調査(縦断調査)の分析法の一つとして、マイクロシミュレーション分析がある。マイクロシミュレーションとは、各種属性を持った個人の集団をコンピュータ上に構成して、おのおのの行動や状態変化を発生させることにより、集団の変化を再現するシミュレーション手法である。対象集団の将来予測、行政制度・施策の効果の予見をはじめ、行動メカニズムの解明や統計手法の精度評価など、幅広く応用される。一方、パネル調査は、抽出された標本内の同一対象(個人、世帯)を追跡しながら継続的に調査し、対象者の変化とその要因を記録して行くものであり、その枠組みやデータ構造はマイクロシミュレーションにきわめて近く、その分析手法として親和性が高いといえよう。実際、諸外国においては、社会政策、税制等の制度・施策の評価や検討のためにパネル調査に基づいたマイクロシミュレーション分析が行われている。

本研究では、諸外国の事例について検討を行った後、21 世紀縦断調査を基にしたマイクロシミュレーション分析を行うための基礎フェーズに対する支援システムの開発を行った。すなわち、パネル調査データの管理情報を基に、シミュレーション分析に必要となる標本モデルをシミュレーション言語（現行では C++）と連携しながら生成するシステムを作成した。システムは、本事業で構築を行ったデータマネジメントシステムの一環として開発されており、統合的に扱うことができる。本システムによれば、縦断調査データに即したさまざまなタイプのマイクロシミュレーション分析を比較的簡単に展開することができる。

2．パネル調査とマイクロシミュレーション

マイクロシミュレーションとは、各種属性を持った個人の集団をコンピュータ上に構成して、おのおのの行動や状態変化を発生させることにより、集団の変化を再現するシミュレーション手法である。とくに縦断型マイクロシミュレーション longitudinal micro-simulation と呼ばれるものは、個人の経時的変化を模擬するもので、パネル調査データとの親和性が高く、対象集団の変化の将来予測、行政制度・施策の効果の予見をはじめ、行動メカニズムの解明や統計手法の精度評価など、既存の統計分析に止まらない多くの応用と可能性を持っている。パネル調査で捉えられた標本をシミュレーションモデルとして再現すれば、さまざまな仮想的条件や仮定の下での標本の変化を観察することが可能であり、それらを実際の変化と比較すれば、仮定の現実的な妥当性を評価することができる。

実際、諸外国においては、社会政策、税制等の制度・施策の評価や検討のためにパネル調査に基づいたマイクロシミュレーション分析が盛んに行われている。カナダでは早くか

ら統計局においていくつかのモデルが開発され、長年にわたって政策シミュレーションに用いられている。そのうち SPSD/M と呼ばれるものは、さまざまな横断調査や行政情報を組み合わせて構築された標本データベースを基にしたシミュレーションモデルであり、主として税制や所得分析に用いられている。また、縦断型のモデルとしては、LifePaths と呼ばれるモデルがある。これは国民を代表する標本について、ライフコース全体をシミュレートする能力があり、個人や世帯を対象とした政策の評価や世代間公平性などの分析に用いられている。POHEM は、健康・疾病に関する縦断型のシミュレーションモデルである。さらに、汎用的なシミュレーションを構築するシステムとして、Modgen という言語が開発されている。これらはすべて統計局のインターネットサイト上に説明書と共に公開されている。アメリカ政府によって実施されているシミュレーション分析とともにこれらを評価した論文集が見られる(Lewis and Michel (eds.) 1990)。アメリカ政府からはマイクロシミュレーションの実施に関する説明資料が公刊されている(Citro and Hanushek (eds.) 1991)。また、この他にも欧米各国（イギリス、ドイツ、オランダ、オーストリア、フィンランド、スウェーデン、デンマーク、ノルウェー、カナダなど）の社会政策、税制等をテーマにしたマイクロシミュレーションの実施に関する個別論文を含んだ論文集が見られる(Harding 1996)。

縦断型マイクロシミュレーションは、21 世紀縦断調査についても、その主要なテーマである結婚・出生・子育てなどの発生メカニズムと決定要因の解明や、制度・施策効果の評価を行う有力な手法となるほか、脱落をはじめとするパネル調査特有の統計分析上の困難に対して、さまざまな条件下におけるそれら統計手法の妥当性や精度を検証する有効な手段を与えると考えられる。

本研究では、21 世紀縦断調査データを活用して今後継続的なマイクロシミュレーション分析が行えるよう、その基礎としてエージェント型(agent-base)のマイクロシミュレーションモデルに必要な標本を生成するシステムを開発した。これはパネル調査データの管理情報を活用して、シミュレーション分析に必要となる標本モデルを半自動的に生成するシステムであり、現行では C++によるシミュレーションモデルを作成することができる。システムは、本事業で構築を行ったデータマネジメントシステムの一環として開発されており、統合的に扱うことができるものである。

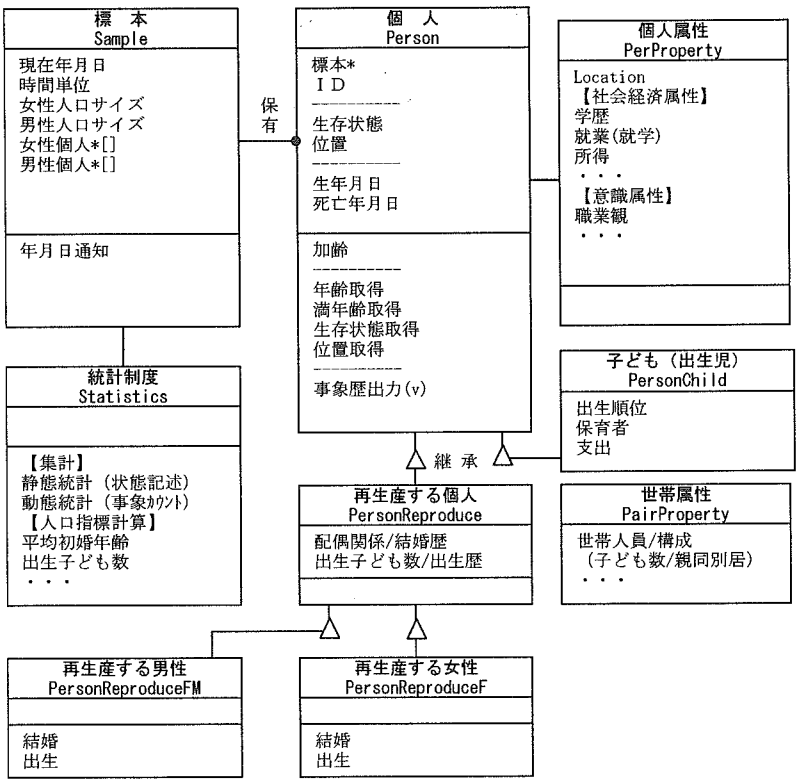
3．システムの開発とプロトタイプモデル

(1) マイクロシミュレーションモデルの概要

ここで想定する縦断型マイクロシミュレーションは、エージェント型(agent-base)のマイクロシミュレーションモデルを基礎とするものである。そこでは個人のモデルは、自律性を備えたオブジェクト、すなわちエージェントして実装される。図 1 には、本シミュレーションのベースモデルとなるプロトタイプモデルのクラス図を示した。これは観察単位（エージェント）の時間的変化・行動を継続的に発生するタイプのクラスの定義である（クラ

スとは、エージェントのシミュレーション言語上の定義のことである)。21 縦断調査の対象者に対応するエージェント・クラスを中心として、その属性や家族などの関係者、さらには標本集団とその統計的特性を集合的に計測、記録、出力する統計制度のエージェント・クラスを配置している。これらを基本とし、出生児調査、成年者調査など各調査ごとに、また分析テーマごとに、必要なエージェント・クラスを追加して分析モデルを構築することとなる。

図１　プロトタイプモデルのクラス図



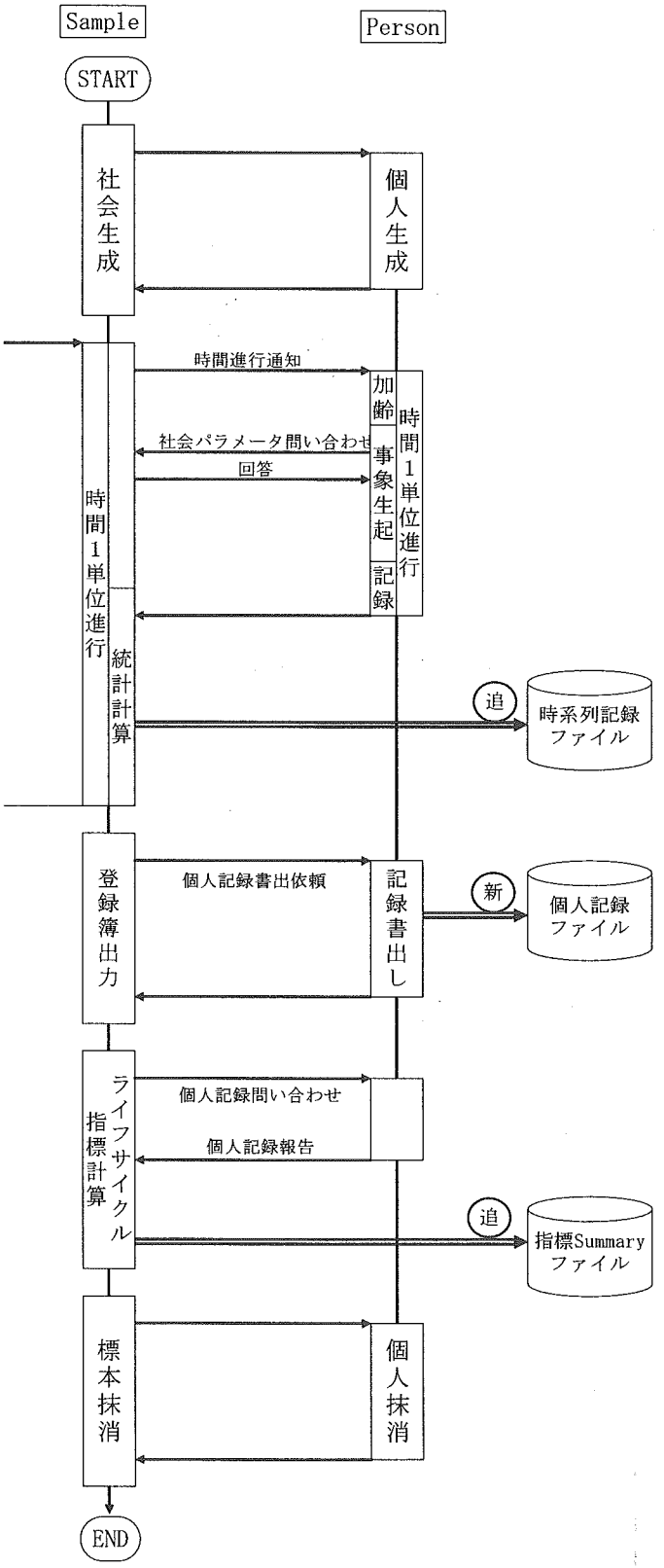
本システムでは、この個人クラスをパネル調査データ情報から自動的に生成することとする。これにより、作成されたクラスに対して、縦断調査の実データを実装することによって、シミュレーションの対象となる標本を生成することができる。この標本を、以下に示す経時的に過程の中におくことによって、個人ならびに集合単位（夫婦、家族、社会等）の行動が模擬され、シミュレーションが行われる。

(2) モデルのシーケンス

次に図２には、本シミュレーションの基本的なシーケンス図を示した。調査対象者の個人を順に生成することによって、標本を再現する。その際、個人には実際の調査から得られた各種の個人属性が割り当てられ、時間経過にしたがって検証するモデル（規則）によって行動が発生することになる。その間に、必要であれば、個人間においての相互作用が発生する。これら相互作用の過程は、通常の統計分析法では推定がほぼ不可能と考えられる領域である。出生等による新たな個人の参入については、新たなエージェントの発生として扱うことができる。これにより現実に近い状況を実現できるが、系の進行状況はきわめて複雑なものとなりうる。これを避けるためには、新たな個人を増やすのではなく新たな関係の発生と考え、属性の変化として扱うことも可能である。その方が状況の見通しはよくなると考えられる。

マイクロシミュレーションは、既存の統計手法の当該データ（21 世紀縦断調査データ）への適用妥当性の検討や、選択的脱落・不詳の効果の推定や将来的な帰結についての予見に用いることができる。そのためには、対象標本について詳細な統計指標を算出することが必要となる。統計指標の算出は、時間単位ごと行われるものと、ライフサイクル指標のように一定の期間終了後に算出されるものがある。いずれにせよ、それら指標はすべて記録され、標本に対する検討対象の統計手法の適用結果と詳細に比較されることによって、手法の妥当性が調べられることになる。

図２　プロトタイプモデルのシーケンス図



(3) シミュレーション用標本生成システムの開発

以上のプロトタイプで示された枠組みを持つシミュレーション分析に対して、21 世紀縦断調査データを用いた標本を生成するシステムを開発した。システムは、データマネジメントシステムとして開発されたコード表変換システムを利用し、シミュレーション言語（現行では C++）によって記述されたシミュレーションのフレーム（上記プロトタイプで示される）に対して、標本を自動生成して供給する。シミュレーションのフレームには、この標本を読み込む機能が用意されており、これらの連携によりシミュレーションの基礎部分が形成されることになる。さらにシミュレーションのフレームは、図２のモデルのシーケンスで示された経時変化がシミュレートされ、これに対して目的に応じた個人、夫婦、家族ならびに集団の振る舞いを定式化して与えることにより、シミュレーション分析が行われる。

図３には、本システムによって自動生成された標本記述の例（一部）を示した。この記述にしたがってシミュレーション言語は標本を読み込む。その際に、脱落や回答不詳に対しては、目的にあった補充 implementation が行われる。

図３ 標本記述の例（成年者調査データの一部）

```
//*****************************************************************************
//***   C++ Class :                                                         ***
//***   第１回成年者縦断調査（女性票）                                      ***
//***   [ test_code_adult_female01.xls (AF01) ]                             ***
//***   <Lexis>                                          2006/02/16         ***
//*****************************************************************************

[Data]
Name         = 第１回成年者縦断調査（女性票）
DataFileName = D:¥ProjectR¥Panel¥00_Data¥sample¥baby¥af01.dat

[Variable]
調査票番号<1,1>{ }
KEY番号（世帯情報）   地区番号<2,5>{
        [01002-47090]地区番号}
KEY番号（世帯情報）   単位区番号<7,2>{ }
KEY番号（世帯情報）   世帯番号<9,2>{ }
KEY番号（世帯情報）   該当者番号<11,1>{ }
KEY番号（世帯情報）   配偶者番号<12,1>{
        [1-6]1-6,
        [.]非該当}
出生年月　年<13,2>{
        [42-57]昭和42-57年,
        [.]不詳}
出生年月　月<15,2>{
        [01-12]01-12月,
        [.]不詳}
問１ 最終学歴<17,1>{
        [1]中学,
        [2]高校,
        [3]専門学校,
        [4]短大・高専,
        [5]大学,
        [6]大学院,
        [7]その他,
        [9]不詳,
        [.]不詳}
問１ 卒業・在学の別<18,1>{
        [1]卒業,
        [2]在学中（休学等を含む）,
        [9]不詳,
        [.]不詳}
```

まとめ

本研究では、諸外国の事例について検討を行った後、21 世紀縦断調査を基にしたマイクロシミュレーション分析を行うための基礎的なシステムの開発を行った。本システムは、パネルデータの管理情報を基に、シミュレーション分析に必要となる標本モデルをシミュレーション言語（現行では C++）と連携しながら半自動的に生成するシステムである。システムは、本事業で構築を行ったデータマネジメントシステムの一環として開発されており、統合的に扱うことができる。本システムによれば、縦断調査データに即した各種マイクロシミュレーション分析を比較的簡単に展開することができるが、本年の研究ではその基本機能を実現することができた。

マイクロシミュレーションは、既存の統計分析に止まらない分析手法として、結婚・出生・子育てなど分析対象となる事象の発生メカニズム、決定要因の解明や、制度・施策効果の評価を行う有力な手法となる。また、既存の統計モデルと合わせて用いることで、それらの信頼性を検証することができるので、パネル調査における統計分析の弱点ともいえる標本脱落や回答不詳・不整合の影響を評価することで、より信頼性の高い分析結果の獲得に資することが期待される。ここでは 21 縦断調査データから、シミュレーションの標本を自動生成するシステムを開発し、今後の分析の基礎を与えた。これにより、諸外国で行われているタイプの政策関連のマイクロシミュレーションをはじめ、今後、本縦断調査に即した多様なモデルが開発されることが期待される。

参照文献

Citro, C. F. and Hanushek, E. A. (eds.) 1991, *The Uses of Microsimulation Modelling. Vol. 1: Review and Recommendations*. National Academy Press, Washington, DC.

Harding, A.(ed.), 1991, *Microsimulation and Public Policy*, Contributions to Economic Analysis, vol.232, Elsevier, Amsterdam.

Lewis, G. H. and Michel, R. C. (eds.) 1990, *Microsimulation Techniques for Tax and Transfer Analysis*. Urban Institute Press, Washington, DC.

Ⅱ.　個別研究報告（データマネジメント）

5　統計処理の概要と課題についての検討

金子隆一

個人を対象として大規模に実施される縦断調査は、統計情報部において初めての経験であった、という以前に、官庁統計にとって初めての経験であった。そのため、調査企画、調査手法、実査、データ処理法、調査担当組織のあり方等について先例のない状況下では、従前の横断調査の経験に拠るしかなく、極端な場合は手探り的なあるいは対症療法的な対応によって、発生した問題やスケジュールをなんとかくぐり抜けてきている側面もあり、調査担当は、日々課題との格闘を通して方法論をひとつひとつ確かなものにしようとしている。以下は、主として縦断調査実施担当者からのヒアリングに基づいて、調査担当サイドでの現時点における方法論（その萌芽のようなものも含めて、つまり実現及び実現性の有無及び可否は問うていない）及びそこに至るまでの状況、経緯、考え方等を記述したものである。

1 統計処理の概要

（1）出生児縦断調査

（処理及びデータの特質）１月と７月生まれの客体について、誕生月の半年後に調査を行うため、１年に２回の調査を行って、それをひとつの結果公表にまとめている。

毎年、８月と２月に調査を行っているが、それぞれ同じ出生時期の、また全体としても同じ年齢の出生児を対象としているから、調査客体にまとまりはあると言えるものの、年二回の調査サイクルは処理工程管理の複雑さを招く調査方法であることも否めない。また調査時点が異なっているグループを共通の土俵上で分析する場合、季節差の評価と共に、あるイベントと個人履歴の時間差がある場合には、分析に組み込まれた時間量の評価が課題であると思われる。

（2）成年者縦断調査

（処理及びデータの特質）調査票が被調査者グループに対応して４種ある（正規コホート[調査開始時に２０歳から３４歳であった被調査者]及びその配偶者[正規コホートを除く]について、それぞれ男女用の調査票がある。）。調査票４種それぞれがレコード化されると共に、夫婦として組み合わされたふたつの調査票レコードから項目を抽出して世帯レコードを構築している。独身の調査対象が結婚した場合は、結婚相手も配偶者として調査対象としているが、彼らは中途から登場するため、他の調査対象との間に情報の構成上のアンバランスがある（アンバランスは、四つの被調査者グループごとに違っている調査項目量によってももたらされる。さらに、途中出場者については、サイクリックに調査する項目について、参加時点以前の情報が存在しないため、分析に用いるためのレコードの抽出に工夫が必要である）。

（３）縦断調査統計処理の特質

長期に亘って被調査者個人のデータセットを保持する必要があり、したがって長期に亘ってデータの整合性を確保する必要がある。整合性確保のための機能であるチェックやデータクリーニングの方針及びその処理に一貫性がなければ、整合性が確保できない。この整合性がなければ、数年分あるいは数十年分のデータを用いて分析する場合のデータの品質に問題が生じる。

２ 統計処理における課題

（１）パネルデータをどう扱うか

統計処理サイクルは一年周期であるが、そもそも、調査後一年以内の公表という目的に合わせてこのサイクルは定まっている。一般に、処理のアウトプットとしての公表物に何を求めるかによって統計処理サイクルは短くも長くもなるものである。パネルデータには、調査を重ねるに連れて、分析の対象となる「生きた」データが量を増やし、その度合いに応じてそれを処理する時間も増加するということが生じるから、必然的に処理のアウトプット（集計結果）作成に使える時間が毎年少なくなっていく。したがって、縦断調査の一年サイクルでの統計処理では、被調査者の様態の前回調査からの変化に集中して集計結果を作成する方向に向かうことになる。そのようにすれば、処理対象となるデータ量は、初回を除いてどの調査回をとってみても等しくなり、効率的に業務の時間が配分できるからである。しかし、調査の節目、例えば５年目、１０年目等にまとまった分析をする場合の時間配分の問題は残っている。

縦断調査の要諦のひとつは変化をいかに因果の関係性に組み込みこんで新たな発見をなしうるかである、と言いうる。そこでデータ処理のポイントはいかに変化を前回調査との比較において的確に捕らえるのか、ということになる。この場合の的確性は、被調査者の様態の全体を正しく捕らえられるかが問題となる。つまり、被調査者のある属性（意識を含む。以下同じ。）がデータ化されるタイミングと、その他の属性がデータ化されるタイミングに時差があっては分析にも時差による修正を加える必要が生じ、これはほとんど不可能だからである。かといって、この時差が無視されてよいものであるか否かは、事項ごとに、時間のもたらす効果が異なっているはずだから、それぞれ違った評価をしなければならないと思われる。

縦断調査のデータの理想的な取り扱いは、以上の諸点から、各調査回で可能な限り被調査者の全属性を調査し、それを個々の属性ごとに変化量と変化した時点と変化の方向について記録すること（履歴データの作成）であり、また、各調査回のスパンを可能な限り短くすることによって、記録されない変化を、より少なくすることである。

（２）統計処理時間の増加の回避

縦断調査では、データクリーニングに要する時間が毎年（毎調査回）増加している。

増加要因は、整合性違反（通常の単純なエラーということではなく、事項間の相互矛盾、すなわち当該調査回の内部比較における、また、以前調査との比較における相互矛盾を指す。）のチェック及び修正である。例えば、一回目と二回目の同じ項目データ

で矛盾があり、二回目に合わせて一回目データを修正したが、三回目調査で、一回目データと同じ値、またはまったく別の値が報告された場合があるとする。この状況が生まれるのは、すべての調査回データを保存しておき、常に全データでのチェックを行うからである。ところが、ある時点における被調査者の全属性があって、それが調査回ごとにアップデイトされるというふうになっていれば、被調査者のデータモデルとしての構造がただひとつ存在することになり、他方、それに対するトランザクションが発生するつどモデルは最新化される、という処理が妥当性を持つ。

理念的にはトランザクション（被調査者からの申し出：調査事項）は常にそのまま受け入れられなければならない。受け入れに際して、矛盾や言い間違いと思われる事項について、問いただすことも、確認することもできないから、構造内の理想の無矛盾性に向かって何をどれくらいやるのかが次のステップである。

なお、（1）の「履歴データ」とこの項の「最新化された被調査者のデータモデル」は、同じデータの集合体を別の切り口で表現したものに他ならない。

パネルデータは、最小レベルでは、ある値について、その値が変化を生じる項目の値であれば、変化量（物理的な変化量と質的な変化量）と変化が生じた時点をその項目の従属項として構成されればよい。（もちろん変化を生じない項目であれば値だけがあればよい。ただし、当然そのデータとしての製造日は記録されている。）

ここで生じうる整合性違反とは、変化を生じてはならない項目に変化が生じた場合だけである。これは不詳としてしか扱い得ない（もちろん本人に確認すればよいのであるが、その方法は調査のそもそもの企画時点で時間コスト、予算コスト及び調査方法上の問題から実施しないこととしている。）。その他の場合は、すべて変化として扱えばよいから、整合性違反の範囲は極めて限定できることになる。（不詳として扱う場合、全データを対象として処理するのであれば、一回目調査の項目の値を不詳とするのか、二回目のそれを不詳とするのか、それとも両方とも不詳として扱うのか、頭を悩ませなければならない。）

チェックを厳密に行った上での整合性違反の一例を挙げる。収入を伴う仕事をしていたと回答したのに、収入額についてはそれがないと回答したり（あるいは未記入であったり）、子どもがいないのに児童手当を受給していると回答したりした事例である。これらは、厳密さを事実による裏付けとして求めることができないから、どんなにチェックが厳密であっても単なる矛盾であり、もし不詳データを設けるのであれば、矛盾の関係にある二つのデータはともに不詳として扱われるべきである。

したがって、「時間的な前後がある整合性違反については、時間的に後に生じた事項を採用すべきであり、時間的な前後がない整合性違反は、双方を不詳とすべきである。」

しかし、仕事をしていたのに収入がない場合に、双方を不詳としたとすると、同時に、もしも職業や就業先の情報があれば、それらも不詳としなければならない。仕事をしていたとした方が救える情報が多いか

らそちらを採用する、ということはあり得るのであろうか？その場合、収入がないとしたのは何かの事情があることとして収入額だけを不詳とする方が合理的であると判断するのであろうか。事実の裏付けがあり得ないのであるから、判断基準をこのような功利によって満たしたとしてもやむを得ないと考える。ある人が質問に答える場合、より多く答えている設問の答のほうが、そうでない方に比べてより事実に基づいているという可能性は高く、収入額だけ記入しないのは、回答するのがいやだからなのだと考えたほうがよい。対面による即時の確認が不可能な場合においては、回答の事情を忖度するのは、事実と可能性を交換することであり、無益にして無駄である。ここでの例では、支払われない労働を受け持っている家族従業者の場合など、収入を記載しないことが仕事を持っていることと矛盾しないことがあるなど、さまざまな「可能性」がありうるため、正解に近づくための参照事例は・・・というふうに芋づるで関連のチェック項目が増殖し、同時にそれに費やす時間と労力も増え、処理しきれないボリュームはバックログとして積みあがっていくのみ、といった事態に陥りがちである。

したがって上述括弧書の記述は「時間的な前後がある整合性違反については、時間的に後に生じた事項を採用すべきであり、時間的な前後がない整合性違反は、双方を不詳とする。しかし、別のあるいはどちらか一方の情報を採用することによって、救済される情報がより多いならそうすべきである。」と書き換えられる。これが守られるためには、参照できる項目間の関係性が定義されていることが条件である。さもなければ、個々のチェックリストと格闘する時間が長引いて疲れ果てて生産性は墜落するであろう。また、関係性の定義は、チェックプログラムにおいて記述されるべきであることを付け加えておきたい。時間の制約があまりにも厳しい現場での文書作成の効率化として、最小単位としての、かつ、煩雑な仕様書作りの代替物、すなわち「記録体」として実際に稼動している（ソース）プログラムを活用することを考えたほうがよい。つまり、文法、記法的に厳密性かつ無誤謬性を有しているものとして、また、常に同一の成果を産出するものとしてのプログラムは、製品の質を保証する究極の生産管理文書としても役立たせることができるのである。

（3）欠損値への対応

外国の縦断調査の例では、(長時間の対面による、あるいは電話による）インタビュー、インタビューにおけるコンピュータ利用、といった方法を用いることによってデータの整合性違反は最小限にとどめられていると思われる。しかしそうであっても回答拒否のような事例が生じた場合、そのデータは欠損値として扱われるしかないであろう。また、長期の旅行等の理由によって、規定の期間内での調査が不可能であった場合では、当該調査回のデータがすべて欠損値になるであろう。このような非回答項目を含めて整合性違反等によるエラーデータへの対応は、二つの方法がある。ひとつは、別の推定された値によって補完すること。もうひとつは、最後まで欠損値として扱うことである。