データ・クリーニング時期別にみた
エラー検出傾向に関する基礎的分析 *[1]

菅澤貴之
（同志社大学）

保田時男
（関西大学）

【論文要旨】

2015 年 SSM 調査では、保田時男が提案する新たなデータ・クリーニング技法が採用された。保田が提案するデータ・クリーニング技法は、Fellegi and Holt（1976）の原則に従い、ケース単位でデータに修正を施していくというもので、作業プロセスが詳細に記録されるため、進捗状況の把握や作業内容の検証が容易である。

そこで本稿では、今回のクリーニングの過程で検出された edit エラー（edit ルール違反）に注目し、データ・クリーニングの観点から回収された調査データについて基礎的な分析を行った。まず、データ・クリーニング作業の過程を時系列に沿って詳細に紹介し、大きな混乱もなく作業が順調に進行したことを確認した。続いて、「①どのような edit ルールで異常ケースが頻発していたのか」、「②どういった属性のケースに edit エラーが集中していたのか」という 2 つの分析課題を設定し、edit エラーの検出傾向について分析を行った。分析の結果、職歴の多い回答者や高齢者層（特に男性）でエラーの検出数が増加することが判明した。ここから示唆される次回 2025 年 SSM 調査の課題は、高齢化に考慮し、どのように複雑化する経歴データを誤りなく収集していくかというものである。この難題を解決する手がかりが、CAI（computer assisted interview）であり、導入の可否について広範な議論が期待される。

キーワード：データ・クリーニング、回顧調査、経歴データ

１．はじめに

「社会階層と社会移動全国調査（以下、SSM 調査）」は 1955 年に開始され、以降、10 年間隔で実施され、2015 年 SSM 調査は第 7 回目の調査となる。今回の 2015 年 SSM 調査では、保田時男が提案する新たなデータ・クリーニング技法が採用された。

保田が提案するデータ・クリーニング技法は、Fellegi and Holt（1976）の原則に従い、ケース単位でデータに修正を施していくというもので、ポイントは次の 3 点に集約される*[2]。①まず、クリーニングの工程は、異常を検出する editing の段階と、検出された異常値を別の値に修正する imputation の段階に峻別され、これらの作業は完全に独立している。②次に、editing を実行するためには、異常値の範囲や組み合わせ（edit ルール）を論理式で明示しなければならない。この edit ルールは、範囲外の値の検出を行う「range edit」、枝分かれ（ス

[1] 本研究は、JSPS 科研費 JP25000001 の助成を受けたものです。

[2] データ・クリーニング技法の開発過程、クリーニング作業時に用いられた補助ツール等の詳細については、保田（2018）を参照のこと。

クリーニング）質問に対する回答の誤りを検出する「filter edit」、その他の異常の可能性が高い箇所について点検する「general edit」の 3 種類に区別される。③データの修正は検出された全ての edit エラー（edit ルール違反）を考慮し、ケース単位で修正を施す（保田 2011, 2012）。これらのポイントにくわえて、修正作業のプロセスが事細かに記録されるため、進捗状況の把握や作業内容の振り返り（検証）が容易という利点もある。

そこで本稿では、今回のクリーニングの過程で検出された edit エラーに注目し、データ・クリーニングの観点から回収された調査データについて基本的な検討を行う。また、データ・クリーニングやアフター・コーディングなどデータ作成の過程は、これまで、研究者間で伝聞により継承され、作業内容が公になることは稀であった。そのため、クリーニングなどの作業に新規に加わった若手研究者は、過去の経緯が分からず、判断に苦慮することもあった。こうした点を考慮すると、本稿をとおして、今回のデータ・クリーニング作業の経験が共有されることは一定の意義があると思われる。

本稿の構成は以下に示すとおりである。続く、第 2 節では 2015 年 SSM 調査データ・クリーニングの過程について紹介する。その後、第 3 節では分析に用いたデータと分析課題について示す。第 4 節、第 5 節では、データ・クリーニングの過程で検出された edit エラーの傾向を確認するために、基礎的な分析を行う。最後に、第 6 節では、分析結果を総括し、今回のクリーニング作業の結果、浮かび上がってきた SSM 調査に対する課題を述べる。

２.作業過程の紹介

ここでは、データ・クリーニング作業の過程について確認する。作業の大まかな流れは表 1 にまとめたとおりである。

まず、2015 年 10 月に保田、菅澤を含めた計 4 名が幹事 2 名のもとで edit ルールの作成を開始した。edit ルールの作成に先立ち、作業従事者は保田研究室でミーティングを開催し、クリーニング責任者である保田から edit ルールの概要や作成手順*[3]についてインストラクションを受けた。edit ルールの作成は各自で担当を分担し*[4]、11 月中旬にクリーニング責任者に作業結果を報告した。各自の作業結果を集約した結果、2986 個（range edit 845 個、filter edit 1170 個、general edit 971 個）の edit ルールが完成した。

2015 年 12 月 27 日から 29 日にかけて行われた第 1 次クリーニング合宿は、職業・産業コーディング合宿と並行して実施された。このため、第 1 次クリーニングでは、表 2 に示したとおり、職歴関係の edit ルール 707 個を保留し、2279 個について点検を行った。実際の作業は、点検の結果、edit エラーが検出されたそれぞれのケースについて、各ケースがもつ周辺

[3] edit ルールの論理式は SPSS シンタックスで記述される。
[4] range edit、filter edit については担当範囲を分担し、general edit については思いつく異常パターンを試験的に提案した。

情報を考慮しながら修正の有無や適切な対処方法を検討し、保田の開発した補助ツールを用いて修正シンタックスを作成することであった。修正作業はケース単位で行われたので、参加者が 50～100 ケース単位で分担した。作業内容は、修正の有無も含め修正シンタックスとともにケース単位でエクセルシートに記録されたため、クリーニング責任者は、合宿期間中でも作業内容の把握が可能である。3 日間の合宿期間内では、作業を完遂させることができなかったので、未修了ケースについては参加者で分担し、後日、作業結果をクリーニング責任者に報告した。

表１　データ・クリーニング作業の流れ

2015年10月	:	editルールの作成を開始（作業者名）
		range edit 84個　filter edit 1170個　general edit 971個を作成
2015年12月27日～29日	:	第1次クリーニング合宿（1日当たりの参加者は0名ほど）
		職業・産業コードは未入力
		未修了分は2月末〆切で分担
2016年8月5日～8日	:	第2次クリーニング合宿（1日当たりの参加者10名ほど）
		職歴関係を中心にクリーニング
2016年9月9日～12日	:	第2次追加クリーニング合宿（1日当たりの参加者10名ほど）
		第2回クリーニング合宿未修了ケースのクリーニング
2016年12月3日～2017年1月31日	:	最終クリーニング（保田、菅澤で作業）
		editエラーが1個以上検出されたケース数1691
		修正漏れの再修正がメイン
		教育・職歴関係の保留事項に関する最終判断
2017年2月27日	:	職歴クリーニングの完了した第次データ(v070)配布
2017年4月16日	:	職歴データワークショップ

表２　点検 edit ルール数

第1次

	range	filter	general	合計
①開始時	845	1170	971	2986
②保留事項（職歴関係等）	26	0	681	707
点検項目数（①-②）	819	1170	290	2279

第2次

	range	filter	general	合計
①開始時	892	1248	1478	3618
②許容されたedit	35	6	105	146
点検項目数（①-②）	857	1242	1373	3472

第 2 次クリーニング合宿は、2016 年 8 月 5 日から 8 日の 4 日間と 9 月 9 日から 12 日の 4 日間の 2 度に分けて実施された。第 2 次クリーニングでは、第 1 次クリーニングで保留された edit ルール 707 個を含めた 3618 個について点検を行った。第 2 次クリーニング用データ

には、職業・産業コーディングの結果が反映されていたため、修正作業の中心は職業関係項目であった。また、第 2 次クリーニングの点検 edit ルール数は、表 2 を見ても明らかなように、第 1 次クリーニング時と比べて大幅に増加している。これは、第 1 次クリーニングの終了後に、職業・教育・家族に関する設問のアフターコードがデータに反映された結果、新たな edit ルールが追加されたためである。さらに、edit ルールの一部については、クリーニング責任者の事後判断や参加者との議論の結果、合宿途中より許容することにした*[5]。こうした柔軟な対応が可能な点も、今回のクリーニング技法の特徴である。

第 1 次、第 2 次のクリーニングを通して、合宿の参加者は 1 日あたり平均 10 名ほどであった。参加者は、大学院生からシニア研究者までと多岐に渡ったが、研究歴と作業量に明確な関係性は認められず、参加者の 1 日あたりの作業量は 200〜300 ケースほどであった（1 日の作業時間は 7 時間ほど）。また、判断に迷ったケースに遭遇した場合は、参加者同士で相談しあいながら、適切な対処法を考え、情報を共有した。作業に関しても、合宿初日にクリーニング責任者による詳細なインストラクションを開催したため、大きな混乱もなく、順調に進めることができた。

合宿形式の労働集約型クリーニング作業は、2016 年 9 月 12 日をもって終了したが、2016 年 12 月に入り、最終的な確認作業を保田、菅澤の 2 名で行った。この時点で、edit エラーが 1 つでも検出されたのは 1691 ケースであり、全体の 1/5（21.6%）ほどを占めていた。ただし、作業は修正漏れ（修正内容を集約したエクセルシートへの修正シンタックスの記載忘れなど）への対処が中心であったことから、作業者が 2 人であっても、過大な負荷とはならず無理なく作業を進めることができた。さらに、これまでのクリーニング作業の中で修正判断が保留されたケース*[6]についても最終判断を決定し、このシステムによる一連のデータ・クリーニング作業が一旦終了した。

こうした過程を経て、データ・クリーニングの完了した第 3 次データ（v070）が 2 月下旬に完成した。その後、職歴データワークショップが 4 月 16 日に開催され、2015 年 SSM 調査研究会メンバー内で、今回のデータ・クリーニングの方法や作業内容について情報共有がはかられた。

---

[5] 許容された edit ルールの一例は以下のとおりである。
[ed3084]問 24(3)でずっと親と別居とあるが、問 42 で父か母と同居している
→質問文における「親の世帯で」という文言が曖昧であるため許容することにした。
また、合宿期間中に、参加者の提案により新たな edit ルールが追加されることもあった。

[6] 職業と教育関係の処理については、高度な専門性を必要とするため、職業関係を三輪哲氏、教育関係を中村高康氏に一任した。

３．使用データと分析課題

3.1　使用データの概要

以下に、データ・クリーニングの過程で生成された一連のデータを用いて、2015 年 SSM 調査におけるエラーの特徴について分析した結果を示す。分析に使用したデータの作業用ファイル名は、「SSM2015 クリーニング用 v030」、「SSM2015 クリーニング用 v060」、「SSM2015_v070_20170227」の 3 つである。各データの使用目的は以下のとおりである。

①SSM2015 クリーニング用 v030（第 1 次データ・クリーニング時の最新データ）
・表 2 に示した edit ルール 2279 個が記述された SPSS シンタックスプログラムを適用し、第 1 次クリーニング時の edit エラーを検出した。

②SSM2015 クリーニング用 v060（第 2 次データ・クリーニング時の最新データ）
・表 2 に示した edit ルール 3618 個が記述された SPSS シンタックスプログラムを適用し、第 2 次クリーニング時の edit エラーを検出した。

③SSM2015_v070_20170227
・第 4 節、第 5 節の分析で用いた個人属性・調査環境に関する変数を抽出した。

ID をもとに、上記 3 種類の調査データをマージし分析用データを作成した。

3.2　分析課題

本稿では、2015 年 SSM 調査で実施された新たなデータ・クリーニング技法について検討するため、以下に示す 2 つの分析課題を設定した。

分析課題（1）：どのような edit ルールで異常ケースが頻発していたのか
分析課題（2）：どういった属性のケースに edit エラー（異常値）が集中していたのか

分析課題（1）では edit ルールに注目し、分析課題（2）では、分析の視点をケースに変え、edit エラーが検出された傾向について基礎的な分析を行う。

４．分析課題（1）：edit ルール別にみた異常ケース数の検出傾向

ここでは、分析課題（1）について検討していく。表 3 は edit ルールごとに点検に引っかかったケース（以下、異常ケース）数を集計したものである。第 1 次クリーニングでは、1 つの edit ルールあたり平均 2.970 の異常ケースが検出され、第 2 次クリーニングにおいては

1 つの edit ルールあたり平均 2.092 の異常ケースが検出された。

次に、edit ルールの種別ごとに、異常ケースの検出傾向を付表 1 も参照しながら確認していく。まず、range edit で検出された異常ケース数は、第 1 次で平均 0.330、第 2 次で平均 0.104 であった。付表 1 にも明示されているように、第 1 次で点検した range edit の 93.5%（766 edit）、第 2 次においても 92.8%（795 edit）の range edit に異常ケースがまったく検出されなかった。異常ケースが検出された range edit ルールも、その多くは複数コードの未処理（処理待ち）*[7]が原因であった。

表３　edit ルールごとに集計した異常ケース数

第1次

	平均値	標準偏差	最小値	最大値
range	0.330	4.690	0	130
filter	2.654	12.787	0	206
general	11.700	27.459	0	250
全edit	2.970	14.144	0	250

第2次

	平均値	標準偏差	最小値	最大値
range	0.104	0.455	0	6
filter	0.643	3.068	0	51
general	4.642	20.416	0	508
全edit	2.092	13.136	0	508

（単位＝case）

続いて、filter edit について見てみると、異常ケース数の平均は、第 1 次で 2.654 であったが、第 2 次では 0.643 と大きく減少している。これは、第 1 次クリーニングの段階で、職業関係を除いた filter edit の大部分に適切な修正処理が施されたことを意味している。実際、異常ケースがまったく検出されなかった filter edit の割合は、第 1 次から第 2 次にかけて 10 ポイント以上増加している。他方、51 以上の異常ケースが検出された filter edit は、第 1 次の時点で 19 edit（1.6%）存在したが、第 2 次ではわずか 1 edit（0.1%）のみであった。

次に、general edit の集計結果を確認してみると、検出された異常ケース数の平均は、第 1 次クリーニング時で 11.700、第 2 次になると 4.642 まで減少するものの、その数は range edit と filter edit を遥かに凌駕している。付表 1 に示されているように、異常ケース数「0」の割合も、第 1 次では 44.8%（130 edit）、第 2 次でも 70.9%（974 edit）に留まる。

異常ケースが 41 以上検出された edit ルールは、第 1 次で全 2279edit 中 52 edit、第 2 次で全 3472edit 中 43 edit を数えたが、この内、general edit は、第 1 次で 55.8%（29 edit）、第 2

[7] データ・クリーニング作業と並行して家族、教育項目に関する複数コードの処理や「その他」回答のアフター・コーディングが実施された。

次では 97.7%（42 edit）を占めていた。

これらの結果から、第 1 次、第 2 次クリーニングを通して、異常ケースは general edit を中心に検出され、修正作業は general edit への対応に主眼が置かれていたことが判明した*[8]。

ところで、異常ケースが多発した edit ルールについて、もう少し詳しく内容を検討してみたい。本稿巻末の資料 1 には、異常ケースが 51 以上検出された edit ルールを掲載している。general edit の集計結果によると、第 1 次では、「父母学歴についての旧制・新制の混同」、「資産や世帯収入合計値の不一致」に対する修正処理が多かったようだ。職業関係項目の点検が中心であった第 2 次の集計結果を見ると、「現職と最終職で役職等の不一致」、「国鉄・公務員の規模不一致」、「初職が学生アルバイトの疑い」等の edit ルールで異常ケースが頻発したことが示されている*[9]。

さらに、第 1 次の filter edit の集計結果を見わたしてみると、現在の状況（現職）が学生または無職であるにも関わらず、有職者を対象とした留置票問 24 から問 26 について誤って回答したケースが多発していたことがわかる。また、これらの filter edit について、異常ケースの年齢構成を調べてみたところ、60 代以上が約 60%を占めていた*[10]。この結果は、リード文の文字を拡大する等、高齢層に配慮した調査票レイアウトが次回の SSM 調査では求められていることを暗示している。

最後に、異常ケースが頻発した edit ルール間で思いもよらない関係性が発見できるのかを確認するため、異常ケースが 10 ケース以上検出された edit ルールについて相関係数にもとづき組み合わせ表*[11]を作成してみた。集計の結果、edit ルールの組み合わせ（異常パターン）は、主問・副問や隣接する質問項目など、想定の範囲内に収まるものばかりであり、意外な質問項目同士の組み合わせは認められなかった。

---

[8] ただし、この結果が range edit や filter edit の重要性を否定するわけではないことは、念のため書き添えておく。多様な general edit に対応する修正をほどこすなかで、うっかりと range edit や filter edit に違反する修正値を入力してしまうことがままある。このミスに気付くためにも range edit や filter edit と確実に点検することは重要である。

[9] 職歴関係で頻発した edit ルールについては資料 2 に対処法も含めて記した。

[10] 例えば、[ed2573z：dq24_1 は非該当のはず（case＝64）]における異常ケースの年齢割合は以下のとおりである。
20 代：20.3%（case＝13）、30 代：7.8%（case＝5）、40 代：3.1%（case＝2）、50 代：4.7%（case＝3）、60 代：34.4%（case＝22）、70 代：29.7%（case＝19）
60 代以上の他には、学生が一定数を占める 20 代の割合が多い。また、ed2574z、ed2577z、ed2578z、ed2579z についても、検出された異常ケースの年齢割合に大きな相違はなかった。

[11] 組み合わせ表の作成・確認の手順は以下のとおりである。
①分析の対象を異常ケースが 10 ケース以上検出された edit ルールに限定した。この結果、分析対象となる edit ルールは第 1 次クリーニングで 131edit、第 2 次クリーニングで 174edit となった。
②クリーニング時期別に分析対象となった edit ルール間で相関係数を算出し、相関係数が 5%水準で有意であった組み合わせについて頻度順に表にまとめた。
③1 つずつの組み合わせを目視で確認したが、特段に意外な組み合わせは見られなかった。

５．分析課題（2）：属性・調査環境別にみた edit エラー検出傾向

5.1　分析で検討する変数

続いて、分析課題（2）について検討していく。本稿では、個人属性として性別、年齢、学歴、職歴段数（職歴数）の 4 項目、調査環境として実査時期、正規・予備票の相違、面接調査実施までの訪問回数、面接時間の 4 項目を取り上げ、各項目について edit エラーの総数を比較する。各変数（調査項目）の操作的定義は表 4 に示したとおりである。

表４　検討する変数（調査項目）リスト

変数名		
q1_1	性別	男性、女性
q1_2_5	年齢	20代＝20～29歳、30代＝30～39歳、40代＝40歳～49歳、50代＝50～59歳、60代＝60～69歳、70代＝70～80歳
edssmx	学歴	初等教育＝中学、中等教育＝高校・専門学校、高等教育＝短大・高専・大学・大学院
dansu	職歴段数	なし＝0、1職＝1、2職＝2、3職＝3、4職＝4、5職＝5、6職＝6、7職以上＝7～22
chosajiki	実査時期	第1期、第2期、第3期
yobi	正規・予備票	正規票、予備票
kaisuu	訪問回数	1回＝1、2回＝2、3回＝3、4回＝4、5回＝5、6回以上＝6～31
mensetu_4	面接時間	25分以内＝2-25、26-30分＝26-30、31-35分＝31-35、36-40分＝36-40、41-45分＝41-45、46-50分＝46-50、51-60分＝51-60、61分以上＝61-260

表５　分析に使用した変数の記述統計

		度数	%			度数	%
性別				実査時期			
	男性	3568	45.6		第1期	2882	36.9
	女性	4249	54.4		第2期	2570	32.9
年齢					第3期	2365	30.3
	20代	729	9.3	正規・予備票			
	30代	1157	14.8		正規	6911	88.4
	40代	1411	18.1		予備	907	11.6
	50代	1323	16.9	訪問回数			
	60代	1712	21.9		1回	1561	20.0
	70代	1485	19.0		2回	2089	26.8
学歴					3回	1622	20.8
	初等教育	1005	12.9		4回	1079	13.8
	中等教育	4175	53.5		5回	737	9.4
	高等教育	2631	33.7		6回以上	683	8.7
職歴段数					不明	38	0.5
	なし	209	2.7	面接時間			
	1職	668	8.5		25分以内	1559	20.0
	2職	1305	16.7		26-30分	1747	22.4
	3職	1392	17.8		31-35分	1098	14.1
	4職	1337	17.1		36-40分	1244	15.9
	5職	1073	13.7		41-45分	593	7.6
	6職	746	9.5		46-50分	564	7.2
	7職以上	1087	13.9		51-60分	518	6.6
					61分以上	452	5.8
					不明	34	0.4

5.2　記述的分析

まず edit エラー検出数について概観する。表 6 はケースごとに edit エラー検出数を集計した結果である。第 1 次クリーニングでは、1 ケースあたり平均 1.295 個の edit エラーが検出され、第 2 次クリーニングでは、1 ケースあたり平均 0.929 個の edit エラーが検出された。

さらに、edit ルールの種別ごとに比較すると、第 1 次、第 2 次クリーニングともに、general edit のエラー検出数が最も多い。この結果を見ても、修正作業の中心は general edit への対応であったことがわかる。

続いて、まったくエラーが検出されなかった、すなわち、edit エラー0 個のケース数は、第 1 次クリーニングで 3689 ケース（47.2%）、第 2 次クリーニングで 4169 ケース（53.3%）であった。さらに、検出された edit エラーが 1 個のみであったケースは、第 1 次クリーニングで 1797 ケース（23.0%）、第 2 次クリーニングで 1902 ケース（24.3%）であり、全体の 1/4 ほどは簡易な修正で対応可能であったことが示唆されている。他方、11 個以上の edit エラーが検出されたのは、第 1 次クリーニングで 60 ケース（0.8%）、第 2 次クリーニングでわずか 18 ケース（0.2%）であり、全体の 1%未満に留まっていた。

なお、edit エラー検出数の分布状況の詳細については、付表 2 に集計結果を整理してある。

表６　ケースごとに集計した edit エラー検出数

第1次

	平均値	標準偏差	最小値	最大値
range	0.042	0.250	0	7
filter	0.398	1.549	0	34
general	0.855	1.253	0	15
全edit	1.295	2.127	0	43

第2次

	平均値	標準偏差	最小値	最大値
range	0.011	0.166	0	7
filter	0.102	0.542	0	16
general	0.815	1.549	0	32
全edit	0.929	1.808	0	43

（単位＝edit）

表７　男女別にみたエラー検出数

		度数	平均値	標準偏差	最小値	最大値
第1次	男性	3563	1.319	2.076	0	43
	女性	4245	1.273	2.166	0	27
	合計	7808	1.294	2.126	0	43
第2次***	男性	3568	1.036	2.003	0	37
	女性	4249	0.839	1.621	0	43
	合計	7817	0.929	1.808	0	43

* p<0.05; ** p<0.01; *** p<0.001

続けて、属性別にエラー検出数の平均値を比較する。はじめに、男女別に比較した表 7 によると、第 1 次、第 2 次ともに男性のほうが、平均エラー検出数は多い。ただし、第 1 次クリーニングにおける平均エラー検出数の男女差は 0.046 であり統計的にも有意ではない。一方、第 2 次クリーニングの男女差は 0.197 であり、統計的にも有意差が認められた。

表８　年齢別にみたエラー検出数

		度数	平均値	標準偏差	最小値	最大値
第1次***	20代	729	1.089	2.090	0	27
	30代	1156	1.125	1.793	0	23
	40代	1409	1.204	2.260	0	43
	50代	1322	1.298	2.036	0	20
	60代	1710	1.424	2.150	0	24
	70代	1482	1.458	2.276	0	25
	合計	7808	1.294	2.126	0	43
第2次***	20代	729	0.573	1.082	0	12
	30代	1157	0.801	1.186	0	11
	40代	1411	0.898	1.987	0	43
	50代	1323	0.986	1.562	0	32
	60代	1712	1.058	1.896	0	37
	70代	1485	1.033	2.317	0	37
	合計	7817	0.929	1.808	0	43

* p<0.05; ** p<0.01; *** p<0.001

次に、年齢別に平均エラー検出数の相違を確認したところ、第 1 次、第 2 次の双方で、おおむね年齢が上がるごとにエラー検出数が増加する傾向にあった（表 8）。特に、第 2 次クリーニングでは、20 代の平均エラー検出数が他の年代と比べて少ないことが際立っている。この結果は、大学などに在学中で「職歴なし」のケースが 20 代前半では一定数を占めていることを反映しているとも解釈できる。

表９　学歴別にみたエラー検出数

		度数	平均値	標準偏差	最小値	最大値
第1次***	初等教育	1005	1.731	2.577	0	22
	中等教育	4169	1.284	2.179	0	43
	高等教育	2628	1.135	1.793	0	27
	合計	7802	1.291	2.123	0	43
第2次**	初等教育	1005	0.738	1.088	0	6
	中等教育	4175	0.944	1.973	0	43
	高等教育	2631	0.976	1.750	0	37
	合計	7811	0.928	1.808	0	43

* p<0.05; ** p<0.01; *** p<0.001

表 9 は、学歴別に平均エラー検出数を集計した結果である。第 1 次については、教育レベルが上がるほど、エラー検出数が減少傾向にあることが示されている。対して、第 2 次では、第 1 次で確認できた関係とは反対に、教育レベルが上がるほどエラー検出数が増加している。

ここで、再度、クリーニング作業の内容を時期別に確認すると、第 1 次クリーニング合宿では職業関係を除く調査項目「全般」について点検を実施し、第 2 次クリーニング合宿では「職業関係」項目について重点的に点検を実施し、検出された edit エラーへの修正方法の対応が作業の中心であった。この点を考慮すると、表 9 に示された第 1 次の結果は、学歴が高い方が調査項目全体に対する理解力も高い関係にあることを意味し、他方、第 2 次の結果は、学歴が高い方が職務や職歴の複雑性が増加する関係にあることを意味しているとも解釈できる。すなわち、第 1 次クリーニングで学歴は調査に対する理解力をあらわしているので、高学歴であるほどエラー検出数が減少し、第 2 次クリーニングで学歴は職務や職歴の複雑性をあらわしているので、高学歴であるほどエラー検出数が増加していると解釈できる。

実際、解釈の妥当性を検討するために学歴と調査理解度*[12]の関係を確認してみたところ、表 10 に示されているように、教育レベルが上昇するほど、面接調査の理解度が「非常に良い」の割合は顕著に増加していることがわかった。さらに、表 11 は、第 2 次クリーニング時に異常ケースが頻発した general edit 上位 5 位について、学歴別に edit エラーの検出率を集計した結果である。これによると、edit エラーの検出が職務や職歴の複雑性に起因すると想定される 3 つの general edit（ed4003、ed4004、ed4005*[13]）は、教育レベルが上昇するほどに検出率が増加傾向にあることが読み取れる。

表 10　学歴別にみた調査理解度

	非常に良い	どちらかといえば良い	あまり良くない	まったく良くない	合計
初等教育	247	571	169	8	995
	24.8%	57.4%	17.0%	0.8%	100.0%
中等教育	1685	2229	216	7	4137
	40.7%	53.9%	5.2%	0.2%	100.0%
高等教育	1574	992	46	3	2615
	60.2%	37.9%	1.8%	0.1%	100.0%
合計	3506	3792	431	18	7747
	45.3%	48.9%	5.6%	0.2%	100.0%

[12] 面接調査終了後に、調査員の視点から、対象者の面接調査への協力度、理解度等を記録している。

[13] edit ルールの詳細については、本稿巻末の資料 1 を参照のこと。

表 11　学歴別にみた edit エラー率（高頻度上位 5 位）

		ed4474	ed4003	ed4004	ed4005	ed3343
初等教育	(n=1005)	6.37%	1.59%	0.90%	0.70%	1.09%
中等教育	(n=4175)	6.54%	2.08%	1.82%	1.34%	1.84%
高等教育	(n=2631)	6.50%	2.58%	2.85%	2.66%	1.63%
合計	(n=7811)	6.50%	2.19%	2.05%	1.70%	1.68%

職歴段数（職歴数）と edit エラー検出数の関係をまとめた表 12 から、第 2 次では、職歴段数の増加に伴い、エラー検出数が増加傾向にあることが読み取れる。とりわけ、「職歴なし」のエラー検出数は少なく、この点からも、第 2 次クリーニング合宿の修正作業の中心が「職業関係」であったことがわかる。

表 12　職歴段数別にみたエラー検出数

		度数	平均値	標準偏差	最小値	最大値
第1次***	なし	209	1.593	3.115	0	27
	1職	668	1.000	1.876	0	20
	2職	1304	1.067	1.788	0	23
	3職	1392	1.216	1.868	0	21
	4職	1334	1.240	1.981	0	24
	5職	1068	1.489	2.248	0	22
	6職	746	1.477	2.300	0	25
	7職以上	1087	1.536	2.546	0	43
	合計	7808	1.294	2.126	0	43
第2次***	なし	209	0.378	0.788	0	4
	1職	668	0.714	1.228	0	12
	2職	1305	0.717	1.145	0	8
	3職	1392	0.881	1.254	0	12
	4職	1337	0.942	1.808	0	32
	5職	1073	1.158	2.970	0	37
	6職	746	1.064	2.063	0	43
	7職以上	1087	1.149	1.729	0	26
	合計	7817	0.929	1.808	0	43

* p<0.05; ** p<0.01; *** p<0.001

続いて、調査環境別に edit エラー検出数の傾向を確認していく。まず、実査期間と正規・予備票の相違による平均エラー検出数を見たものが表 13、表 14 である。実査期間ごとにエラー検出数を集計したものが表 13 であるが、これをみると、第 1 次、第 2 次ともに、実査期間の第 1 期がもっともエラー検出数が多く、第 2 期と第 3 期では大きな差は見られない。

特に、第 1 次クリーニングにおける実査期間第 1 期と第 2 期・第 3 期の差は大きく、統計的にも有意であった。多くの調査員は第 1 期～第 3 期まで続けて調査に関わっていた事から考えると、この結果は、調査員の SSM 調査に対する熟達度を反映していると解釈できる。実査期間を複数に区分したことは 2015 年 SSM 調査の特徴の一つであるが、回収率の向上だけ

でなく、限られた人数の調査員の熟達レベルを上げ良質なデータを回収する点においても、実査期間の複数化が有効に働いた可能性がある。

また、2015 年 SSM 調査でも予備票を用いたが、表 14 によると、第 1 次、第 2 次ともに、正規票、予備票のエラー検出数にほとんど差はなく、統計的にも有意差は認められなかった。

表 13　実査期間別にみたエラー検出数

		度数	平均値	標準偏差	最小値	最大値
第1次***	第1期	2881	1.476	2.196	0	27
	第2期	2564	1.172	2.098	0	43
	第3期	2363	1.203	2.053	0	21
	合計	7808	1.294	2.126	0	43
第2次	第1期	2882	0.967	1.706	0	43
	第2期	2570	0.902	1.980	0	37
	第3期	2365	0.912	1.731	0	34
	合計	7817	0.929	1.808	0	43

* p<0.05; ** p<0.01; *** p<0.001

表 14　正規・予備票別にみたエラー検出数

		度数	平均値	標準偏差	最小値	最大値
第1次	正規	6907	1.297	2.144	0	43
	予備	902	1.280	1.996	0	15
	合計	7809	1.295	2.127	0	43
第2次	正規	6910	0.920	1.657	0	43
	予備	907	1.000	2.693	0	37
	合計	7817	0.929	1.808	0	43

* p<0.05; ** p<0.01; *** p<0.001

最後に、面接調査に至るまでの訪問回数と面接調査に要した時間の差異が、edit エラー検出数に違いをもたらしていたのかを確認する。訪問回数別に平均エラー検出数をまとめた表 15 から、第 1 次では、1 回目から 5 回目までは訪問回数が増加するたびにエラー検出数が線形的に増加することが示されている。すなわち、早期に調査対象者に接触できた場合には、エラー検出数が少ない。一方、第 2 次では、第 1 次で示された関係性は認められず、明確な傾向を読み取ることはできない。

面接調査に要した時間とエラー検出数の関係については、第 1 次、第 2 次共通の傾向として、面接時間とエラー検出数の間には正の相関を発見できる。さらに、集計結果をまとめた表 16 には、「61 分以上」の時間を要した場合には、明確にエラー検出数が増加することが示されている。2015 年 SSM 調査における面接調査の平均所要時間は 37.6 分であった。面接時間が長いことによって回答者や調査員の疲労が増し、エラー数の増加につながっている可能性はある。一方で、職歴が複雑なことや回答者の理解度が低いことによって面接時間が長くなっていることも当然考えられ、面接時間とエラー検出数の関係がもつ意味は単純ではない。

それでも面接時間がエラー確率の一つの指標になることは明らかである。この点を踏まえると、平均時間を大幅に超過した場合には、調査票の点検を通常よりも入念に行う特別なステップを設ける、といった方策は誤りの早期発見に有効な手段となる可能性がある。

表 15　訪問回数別にみたエラー検出数

		度数	平均値	標準偏差	最小値	最大値
第1次*	1回	1561	1.198	1.769	0	15
	2回	2089	1.251	2.137	0	43
	3回	1622	1.325	2.010	0	22
	4回	1079	1.363	2.344	0	25
	5回	737	1.475	2.557	0	27
	6回以上	683	1.305	2.272	0	24
	不明	38	0.789	1.166	0	4
	合計	7809	1.295	2.127	0	43
第2次	1回	1560	0.926	1.565	0	34
	2回	2089	0.878	1.295	0	12
	3回	1622	0.921	1.779	0	43
	4回	1079	0.935	1.620	0	26
	5回	737	0.859	1.284	0	10
	6回以上	683	0.820	1.238	0	9
	不明	38	1.184	1.642	0	8
	合計	7808	0.899	1.503	0	43

* p<0.05; ** p<0.01; *** p<0.001

表 16　面接時間別にみたエラー検出数

		度数	平均値	標準偏差	最小値	最大値
第1次***	25分以内	1559	1.155	2.141	0	27
	26-30分	1747	1.143	1.744	0	17
	31-35分	1098	1.158	1.641	0	12
	36-40分	1244	1.298	2.098	0	24
	41-45分	593	1.567	2.969	0	43
	46-50分	564	1.498	2.207	0	17
	51-60分	518	1.562	2.535	0	22
	61分以上	452	1.774	2.478	0	21
	不明	34	1.265	1.831	0	8
	合計	7809	1.295	2.127	0	43
第2次***	25分以内	1559	0.693	1.294	0	25
	26-30分	1747	0.845	1.756	0	43
	31-35分	1097	0.868	1.209	0	9
	36-40分	1244	0.885	1.298	0	12
	41-45分	593	1.069	1.554	0	15
	46-50分	564	1.071	1.402	0	10
	51-60分	518	1.021	1.459	0	11
	61分以上	452	1.356	2.148	0	26
	不明	34	0.853	0.925	0	3
	合計	7808	0.899	1.503	0	43

* p<0.05; ** p<0.01; *** p<0.001

5.3　多変量解析による検討

5.2 では個人属性である性別、年齢、学歴、職歴段数（職歴数）、調査環境である実査時期、正規・予備票の違い、訪問回数、面接時間と edit エラー検出数の関係について個別に検討した。その結果、明確な関係が確認できた項目もあったが、関係性が不明瞭な項目もあった。

そこで、ここでは多変量解析を用いて、各項目を統制し、どういった項目が edit エラー検出数を規定していたのかを明らかにしたいと思う。従属変数は 1 ケースあたりの edit エラー検出数で、独立変数はこれまで個別に検討してきた 8 項目*[14]である。

付表 2 にも示されているように、従属変数である edit エラー検出数はゼロ事象（0 回）が第 1 次、第 2 次クリーニングともに半数近くを占めているカウントデータである。このようなゼロ事象が多くを占めるカウントデータの推定はポアソン回帰により行う。ポアソン回帰では、従属変数の平均と分散が等しいという仮定のもので推定を行っている。

ところが、事前の分析の結果、edit エラー検出数は第 1 次、第 2 次どちらも、分散が平均を上回ることが判明した。したがって、今回は、平均と分散が等しいという制約を緩めた負の二項回帰を採用し推定を行った。なお、「訪問回数」と「面接時間」に注目し、分析モデルを構築している。表 17 は第 1 次クリーニングの推定結果、表 18 は第 2 次クリーニングの推定結果をまとめたものである。

まず、第 1 次クリーニングの結果について確認していく。model 1 では、訪問回数と面接時間を独立変数として投入している。表 17 によると、訪問回数、面接時間いずれも正の効果を示し、統計的にも有意である。この結果は、先ほどの表 15、表 16 とも整合的である。model 2 は、実査時期と正規・予備票の相違を投入している。訪問回数、面接時間を統制しても、実査時期第 1 期は、第 2 期、第 3 期と比べて edit エラー検出数が増加することが示され、表 13 の結果とも一致する。さらに、表 14 と同様、正規・予備票の違いは edit エラー検出数に影響を与えていない。

model 3 では性別、model 4 では年齢を追加した。model 3 の結果から、調査環境を統制した場合でも、性別は edit エラー検出数に効果をもたないことが確認できた。年齢を独立変数に追加した model 4 からは、20 代と比べて 60 代、70 代では edit エラー検出数が増加することが示されている。この結果をみても高齢層に配慮した調査票設計が求められていることがわかる。

---

[14] 訪問回数、面接時間、職歴段数については、以下のとおりリコード処理を施したうえ、分析を行った。

訪問回数：　1 回＝1、2 回＝2、3 回＝3、4 回＝4、5 回＝5、6 回以上＝6
面接時間：　25 分以内＝1、26-30 分＝2、31-35 分＝3、36-40 分＝4、41-45 分＝5、46-50 分＝6、51-60 分＝7、61 分以上＝8
職歴段数：　なし＝0、1 職＝1、2 職＝2、3 職＝3、4 職＝4、5 職＝5、6 職＝6、7 職以上＝7

表 17　多変量解析（負の 2 項回帰）の結果【第 1 次】

	model 1	model 2	model 3	model 4	model 5	model 6	model 7
訪問回数	0.0350 **	0.0287 **	0.0283 *	0.0374 ***	0.0385 ***	0.0390 ***	0.0391 ***
面接時間	0.0651 ***	0.0605 ***	0.0610 ***	0.0512 ***	0.0488 ***	0.0419 ***	0.0421 ***
実査時期　（基準:第1期）							
第2期		-0.1923 ***	-0.1937 ***	-0.1963 ***	-0.2104 ***	-0.2087 ***	-0.2088 ***
第3期		-0.1700 ***	-0.1708 ***	-0.1793 ***	-0.1955 ***	-0.1930 ***	-0.1944 ***
正規・予備　（基準:正規票）							
予備票		-0.0008	-0.0019	0.0065	0.0028	0.0034	0.0042
性別　（基準:女性）							
男性			0.0529	0.0449	0.0590	0.0660	0.1128
年齢　（基準:20代）							
30代				-0.0237	-0.0406	-0.0870	-0.0365
40代				0.0292	0.0036	-0.0605 *	-0.0266
50代				0.1078	0.0864	0.0154	0.0428
60代				0.1861 **	0.0988	0.0223	0.0085
70代				0.1854 **	0.0358	-0.0307	0.0035
学歴　（基準:初等教育）							
中等教育					-0.2840 ***	-0.2826 ***	-0.2809 ***
高等教育					-0.3928 ***	-0.3805 ***	-0.3803 ***
職歴段数						0.0315 **	0.0311 **
（基準:男性×20代）							
男性×30代							-0.1105
男性×40代							-0.0742
男性×50代							-0.0585
男性×60代							0.0256
男性×70代							-0.0722
定数項	-0.0809	0.0640	0.0399	-0.0460	0.2989 **	0.2480 **	0.2263 *
lnalpha							
定数項	0.3333 ***	0.3249 ***	0.3242 ***	0.3173 ***	0.3034 ***	0.3001 ***	0.2996 ***
対数尤度	-12045.225	-12031.446	-12030.213	-12018.817	-11996.377	-11991.093	-11990.093
aic	24098.45	24076.893	24076.426	24063.634	24022.755	24014.187	24022.187
bic	24126.263	24125.566	24132.052	24154.026	24127.054	24125.439	24168.205
n				7733			

本人学歴を追加した model 5 の結果から、表 9 で認められた関係と同様、学歴は edit エラー検出数に負の効果を与えていることが確認できる。また、年齢係数の変化に注目してみると、model 4 で確認された 60 代、70 代の効果は大幅に減少し、統計的にも有意ではない。この結果を見る限り、同一の学歴であれば、年齢による影響は消失する。高齢者にエラーが多い理由は、単純に（聴力や体力の低下などの）身体的な問題に起因するわけではない、という可能性がある。

続いて、model 6 では職歴段数（職歴数）を独立変数に追加している。表 17 から、調査環境や性別、年齢、学歴を統制した場合でも、edit エラー検出数に対して職歴段数は有意に正の効果をもつことが明示されている。model 7 では、性別と年齢の交互作用について検討してみたが、いずれの年齢層においても有意な効果は確認できなかった。

次に、職歴関係に修正作業の力点が置かれた第 2 次クリーニングの結果について確認する。訪問回数と面接時間の影響をみた model 1 には、第 1 次クリーニングとは異なり面接時間のみ有意差が示されている。面接時間と edit エラー検出数の間には正の関係が認められ、これは表 16 の結果とも一致している。model 2 に目を向けると、第 1 次の結果と同様、正規票と予備票の違いは、edit エラー検出数に対して影響をもっていない。しかし、実査時期は、第

表 18　多変量解析（負の 2 項回帰）の結果【第 2 次】

	model 1	model 2	model 3	model 4	model 5	model 6	model 7
訪問回数	-0.0112	-0.0150	-0.0169	-0.0136	-0.0140	-0.0142	-0.0138
面接時間	0.0764 ***	0.0741 ***	0.0750 ***	0.0647 ***	0.0650 ***	0.0496 ***	0.0498 ***
実査時期　（基準:第1期）							
第2期		-0.1154 **	-0.1203 **	-0.1288 **	-0.1184 **	-0.1244 **	-0.1250 **
第3期		-0.0608	-0.0680	-0.0744	-0.0646	-0.0718	-0.0728
正規・予備　（基準:正規票）							
予備票		-0.0890	-0.0885	-0.0866	-0.0799	-0.0815	-0.0809
性別　（基準:女性）							
男性			0.1947 ***	0.1834 ***	0.1664 ***	0.1796 ***	-0.1454
年齢　（基準:20代）							
30代				0.2790 ***	0.2852 ***	0.1916 *	0.0887
40代				0.3445 ***	0.3565 ***	0.2239 **	0.1217
50代				0.4281 ***	0.4406 ***	0.3001 ***	0.1359
60代				0.4312 ***	0.4951 ***	0.3351 ***	0.0675
70代				0.3450 ***	0.4659 ***	0.3196 ***	0.2288 *
学歴　（基準:初等教育）							
中等教育					0.2974 ***	0.2916 ***	0.2929 ***
高等教育					0.3821 ***	0.3976 ***	0.3959 ***
職歴段数						0.0682 ***	0.0670 ***
（基準:男性×20代）							
男性×30代							0.2501
男性×40代							0.2453
男性×50代							0.3833 **
男性×60代							0.5672 ***
男性×70代							0.2309
定数項	-0.3530 ***	-0.2692 ***	-0.3568 ***	-0.6632 ***	-0.9976 ***	-1.0933 ***	-0.9530 ***
lnalpha							
定数項	0.0865 *	0.0823 *	0.0682	0.0519	0.0342	0.0142	0.0051
対数尤度	-10098.039	-10092.76	-10076.264	-10055.286	-10036.199	-10011.963	-10000.643
aic	20204.078	20199.519	20168.527	20136.572	20102.399	20055.927	20043.286
bic	20231.891	20248.192	20224.153	20226.965	20206.698	20167.179	20189.305
n				7733			

* p<0.05; ** p<0.01; *** p<0.001

1 次クリーニングの結果（表 17）とは違い、第 1 期と第 2 期の間にのみ負の関係が認められた。

性別の効果を確認した model 3 の結果を見ると、男性は女性に比べて edit エラー検出数が増加することが示されている。この関係は、先にみた表 7 の結果とも一致している。さらに、年齢を独立変数に追加した model 4 から、基準カテゴリーとして設定した 20 代と比べて、30 代以降の全世代で、edit エラーの検出数が有意に増加することがわかる。また、面接時間に着目すると、model 3 と比較して効果が減少していることが確認できる。これは、面接時間が edit エラー検出数に与える効果は、年齢を一つの原因としていることを示唆している。

続けて、model 5 の結果から、edit エラー検出数に対して、学歴は有意に正の効果をもつことが確認された。このことは、先に確認した表 9 に示されている結果とも符合する。さらに、model 6 の結果から、第 1 次クリーニングの結果と同様に、職歴段数と edit エラー検出数の間には、有意に正の関係が認められる。くわえて、面接時間の効果が減少していることも確認できる。このことから、面接時間が長時間化すると edit エラー検出数が増加することの一因は、聴取する職歴数が多いためであると推察される。model 6 には年齢効果の減少も示されているが、この結果もまた、職歴段数が年齢と edit エラー検出数を媒介していることを示

している。先に見た model 4 の結果から、面接時間は年齢を共通の原因として edit エラー検出数と関係していることが明らかとなった。したがって、これらの結果を踏まえると、面接時間が長時間化すると edit エラー検出数が増加することの一端は、年齢が上がるほど職歴数が増加し、職歴の聴取に時間を要するためであると推測される。

最後に、性別と年齢の交互作用を投入した model 7 から、男性×50 代、男性×60 代で有意に正の効果が確認できた。この結果は、キャリアの円熟期を迎える 50 代男性と定年退職を迎える 60 代男性は、他の年齢層の男性と比べて顕著に edit エラー検出数が増加することを示している。50 代男性と 60 代男性が調査対象者である場合には、慎重に面接調査を実施するよう調査員にインストラクションした方がよいだろう。また、model 7 からは、女性は男性の場合とは異なり、70 代のみ他の年齢層と比較して edit エラーの検出数が有意に増加することもわかった。

ここまで多変量解析の結果を読み解いてきたが、個人属性、調査環境が及ぼす効果について再確認できたほか、各項目（変数）間の関連性についても明らかにすることができた。次節では、これまでの分析結果をまとめ、SSM 調査の課題について考察する。

６．要約と課題

6.1　分析結果の要約

本稿では、2015 年 SSM 調査で新たに採用したデータ・クリーニング技法について検討するため、2 つの分析課題を提示し、クリーニング合宿を実施した時期ごとに分析を進めた。ここでは、再度、分析の結果、明らかとなった知見を整理する。

分析課題（1）では edit ルールに着目し、「どのような edit ルールで異常ケースが頻発していたのか」を検討した。異常ケースの検出傾向について、edit ルールの種別ごとに比較検討した結果、第 1 次、第 2 次クリーニング共通の傾向として、異常ケースは general edit を中心に検出されていることが判明した。このことから、修正作業の中心は general edit への対応に割かれていたことがデータからも裏付けられた。さらに、異常ケースが頻発した edit ルールの分析から、高齢層を意識した調査票作りが必要とされていることも垣間見えた。他方、edit ルール間で同時に発生しやすいエラーのパターンも分析したが、とくに予想外の組み合わせで発生しやすい異常といったものは発見されなかった。

分析課題（2）では分析の視点をケースに移し、「どういった属性のケースで edit エラーが集中していたのか」を検討した。具体的には、個人属性、調査環境別に edit エラーの検出数を比較した。男女差については、第 2 次に限り統計的な有意が認められ、男性の方が女性に比べて edit エラー検出数が多かった。年齢については、第 1 次、第 2 次クリーニングともに、総じて年齢が上がるほど edit エラー検出数が増加していた。くわえて、第 2 次クリーニング時に限定されるものの、キャリアが円熟期に達する 50 代男性と定年退職を迎える 60 代男性

は、他の年齢層の男性と比較して edit エラー検出数が増加していることがわかった。学歴については、クリーニング時期によって、まったく正反対の傾向を示していた。第 1 次では、教育レベルが上がるほど edit エラー検出数は減少していた。他方、第 2 次では、教育レベルが上がるほど edit エラー検出数が増加していた。この結果は、第 1 次クリーニング時に、学歴は「調査に対する理解力」を反映しており、一方、第 2 次クリーニング時に、学歴は「職務や職歴の複雑性」を反映していると解釈すれば、理解できる。職歴段数（職歴数）については、「職業関係」に力点を置き点検を実施した第 2 次クリーニングにおいて edit エラー検出数との間に正の相関が認められた。

調査環境については、実査時期、正規・予備票の違い、面接調査に至るまでの訪問回数、面接調査に要した時間という 4 項目について検討した。まず、実査期間については、第 1 次、第 2 次で共通した傾向として、実査期間の第 1 期において edit エラーの検出数がもっとも多く、第 2 期と第 3 期では顕著な差は確認されなかった。この結果は、SSM 調査に対する調査員の熟達度（調査への慣れ）をあらわしているとも解釈できる。また、正規票と予備票の相違によって、edit エラー検出数に差異は認められなかった。訪問回数については、第 1 次クリーニング時に限定されるものの、早期に調査対象者に接触できた場合には、edit エラー検出数が明確に減少することが確認できた。面接調査に要した時間については、第 1 次、第 2 次の双方で、面接時間と edit エラー検出数の間で正の相関が認められ、なかでも、面接調査に「61 分以上」要した場合には、edit エラー検出数が明らかに増加していた。さらに、多変量解析の結果から、面接時間の長時間化が edit エラー検出数の増加をもたらすことの一因は、職歴の聴取に時間を割く必要があるためと推察された。

6.2　SSM 調査の課題

分析の結果、職歴の多い回答者や高齢層（特に男性）で edit エラーの検出数が増加していることが見受けられた。ここから示唆される今後の SSM 調査の課題は、高齢化を考慮し、どのように複雑化する経歴データを誤りなく収集していくかというものである。

この難題を解決するヒントは、CAI（computer assisted interview）の導入にあるかもしれない。CAI とは、データ収集時にコンピュータの支援を得る手法一般を指している。欧米では 1970 年代より CATI（computer assisted telephone interview）の導入が進み、現在では広く活用されている。面接調査にコンピュータなどを持ち込み、調査員が回答を現場で入力したり、枝分かれ（スクリーニング）質問に対して適切な飛び先を自動的に提示する手法である CAPI（computer assisted personal interview）も 1980 年代から導入されている（杉野他 2015）。このように、欧米では、社会調査プロジェクト、特に大規模パネルデータの収集において、CAI の導入は一般化している。他方、日本では、SSP 調査プロジェクト等の少数の例外を除いて CAI の導入はほとんど進んでおらず、依然として、紙媒体の伝統的な調査手法である PAPI

（paper and pencil interview）が中心である。

経歴データの収集過程において、CAI が有力な支援ツールとなることは明らかである。職歴を聴取する際に、オリンピック等の象徴的なイベント映像をタブレット PC 上で提示すれば、調査対象者が記憶を思いかえす手助けとなる。調査票の多言語対応や文字サイズの拡大も容易である。さらに、あらかじめ edit ルールをプログラミングしタブレット PC に組み込んでおけば、論理的な不一致や矛盾点をその場で調査対象者に確認することが可能となり、データ・クリーニングの物理的・金銭的コストの削減にもつながる。また、CAI を導入することで、実査時の調査対象者の行動を記録したパラデータを獲得することもできる。パラデータの入手は、これまで以上に調査デザインの詳細な検証を可能とする。

しかしながら、日本の調査会社が雇用する調査員が比較的高齢であることや、個人情報の電子化に対する日本社会独特の抵抗感などが CAI 導入の妨げとなっていると思われる。ただ、「社会調査の困難」が声高く叫ばれる現在、高齢化に配慮し、複雑化する経歴データを正確に収集するには、CAI の導入を避けて通ることは難しい。調査員、調査対象者の双方に過度な負担をかけることなく良質なデータを収集するためにも、CAI 導入の可否について本格的な議論が期待される。

［文献］

Fellegi, I.P., and D. Holt. 1976. “A Systematic Approach to Automatic Edit and Imputation,” *Journal of the American Statistical Association* 71 (No. 353): 17-35.

歸山亜紀・小林大祐・平沢和司．2015.「コンピュータ支援調査におけるモード効果の検証―実験的デザインにもとづく PAPI，CAPI，CASI の比較―」『理論と方法』30(1): 273-292.

杉野勇・俵希實・轟亮．2015.「モード比較研究の解くべき課題」『理論と方法』30(1): 253-272.

保田時男．2011.「NFRJ-08Panel における調査票の設計：研究課題とクリーニングを視野に」『家族社会学研究』23(1): 89-95.

保田時男．2012.「パネルデータの収集と管理をめぐる方法論的な課題」『理論と方法』27(1): 85-98.

保田時男. 2018.「複雑な社会調査におけるデータ・クリーニング技法の開発」保田時男編『2015 年 SSM 調査報告書 1　調査方法・概要』2015 年 SSM 調査研究会: 177-200.

付表 1：edit ルールごとに集計した異常ケース数の分布

第1次

	0	1-10	11-20	21-30	31-40	41-50	51以上	合計
range	766	48	3	1	0	0	1	819
	93.5%	5.9%	0.4%	0.1%	0.0%	0.0%	0.1%	100.0%
filter	890	223	19	12	4	3	19	1170
	76.1%	19.1%	1.6%	1.0%	0.3%	0.3%	1.6%	100.0%
general	130	96	18	10	7	12	17	290
	44.8%	33.1%	6.2%	3.4%	2.4%	4.1%	5.9%	100.0%
全edit	1786	367	40	23	11	15	37	2279
	78.4%	16.1%	1.8%	1.0%	0.5%	0.7%	1.6%	100.0%

第2次

	0	1-10	11-20	21-30	31-40	41-50	51以上	合計
range	795	62	0	0	0	0	0	857
	92.8%	7.2%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
filter	1081	139	16	2	3	0	1	1242
	87.0%	11.2%	1.3%	0.2%	0.2%	0.0%	0.1%	100.0%
general	974	278	42	26	11	14	28	1373
	70.9%	20.2%	3.1%	1.9%	0.8%	1.0%	2.0%	100.0%
全edit	2850	479	58	28	14	14	29	3472
	82.1%	13.8%	1.7%	0.8%	0.4%	0.4%	0.8%	100.0%

（度数＝edit）

付表 2：ケースごとに集計した edit エラー検出数の分布

第1次

	0	1	2	3	4	5	6-10	11以上	合計
range	7534	238	29	2	5	0	1	0	7809
	96.5%	3.0%	0.4%	0.0%	0.1%	0.0%	0.0%	0.0%	100.0%
filter	6703	617	92	48	185	42	83	39	7809
	85.8%	7.9%	1.2%	0.6%	2.4%	0.5%	1.1%	0.5%	100.0%
general	4227	1858	972	417	189	74	70	2	7809
	54.1%	23.8%	12.4%	5.3%	2.4%	0.9%	0.9%	0.0%	100.0%
全edit	3689	1797	1050	493	317	166	237	60	7809
	47.2%	23.0%	13.4%	6.3%	4.1%	2.1%	3.0%	0.8%	100.0%

第2次

	0	1	2	3	4	5	6-10	11以上	合計
range	7753	55	4	0	3	0	2	0	7817
	99.2%	0.7%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%
filter	7372	287	70	23	45	13	6	1	7817
	94.3%	3.7%	0.9%	0.3%	0.6%	0.2%	0.1%	0.0%	100.0%
general	4348	1920	921	374	139	56	46	13	7817
	55.6%	24.6%	11.8%	4.8%	1.8%	0.7%	0.6%	0.2%	100.0%
全edit	4169	1902	952	414	191	99	72	18	7817
	53.3%	24.3%	12.2%	5.3%	2.4%	1.3%	0.9%	0.2%	100.0%

（度数＝case）

付表 3：男女ごとにみた edit エラー検出数の分布

第1次

	0	1	2	3	4	5	6-10	11以上	合計
男性	1624	828	499	245	159	71	116	21	3563
	45.6%	23.2%	14.0%	6.9%	4.5%	2.0%	3.3%	0.6%	100.0%
女性	2065	969	551	248	158	95	120	39	4245
	48.6%	22.8%	13.0%	5.8%	3.7%	2.2%	2.8%	0.9%	100.0%
合計	3689	1797	1050	493	317	166	236	60	7808
	47.2%	23.0%	13.4%	6.3%	4.1%	2.1%	3.0%	0.8%	100.0%

第2次

	0	1	2	3	4	5	6-10	11以上	合計
男性	1822	877	444	211	103	50	49	12	3568
	51.1%	24.6%	12.4%	5.9%	2.9%	1.4%	1.4%	0.3%	100.0%
女性	2347	1025	508	203	88	49	23	6	4249
	55.2%	24.1%	12.0%	4.8%	2.1%	1.2%	0.5%	0.1%	100.0%
合計	4169	1902	952	414	191	99	72	18	7817
	53.3%	24.3%	12.2%	5.3%	2.4%	1.3%	0.9%	0.2%	100.0%

付表 4：年齢ごとにみた edit エラー検出数の分布

第1次

	0	1	2	3	4	5	6-10	11以上	合計
20代	398	163	77	31	20	14	21	5	729
	54.6%	22.4%	10.6%	4.3%	2.7%	1.9%	2.9%	0.7%	100.0%
30代	597	239	145	66	59	27	18	5	1156
	51.6%	20.7%	12.5%	5.7%	5.1%	2.3%	1.6%	0.4%	100.0%
40代	697	342	165	88	53	17	34	13	1409
	49.5%	24.3%	11.7%	6.2%	3.8%	1.2%	2.4%	0.9%	100.0%
50代	608	299	205	81	50	36	35	8	1322
	46.0%	22.6%	15.5%	6.1%	3.8%	2.7%	2.6%	0.6%	100.0%
60代	717	433	253	121	73	38	60	15	1710
	41.9%	25.3%	14.8%	7.1%	4.3%	2.2%	3.5%	0.9%	100.0%
70代	672	321	205	106	62	34	68	14	1482
	45.3%	21.7%	13.8%	7.2%	4.2%	2.3%	4.6%	0.9%	100.0%
合計	3689	1797	1050	493	317	166	236	60	7808
	47.2%	23.0%	13.4%	6.3%	4.1%	2.1%	3.0%	0.8%	100.0%

第2次

	0	1	2	3	4	5	6-10	11以上	合計
20代	481	152	57	23	9	4	2	1	729
	66.0%	20.9%	7.8%	3.2%	1.2%	0.5%	0.3%	0.1%	100.0%
30代	641	279	139	53	27	11	6	1	1157
	55.4%	24.1%	12.0%	4.6%	2.3%	1.0%	0.5%	0.1%	100.0%
40代	764	362	162	60	31	14	14	4	1411
	54.1%	25.7%	11.5%	4.3%	2.2%	1.0%	1.0%	0.3%	100.0%
50代	665	309	208	71	32	24	13	1	1323
	50.3%	23.4%	15.7%	5.4%	2.4%	1.8%	1.0%	0.1%	100.0%
60代	863	418	202	117	56	28	24	4	1712
	50.4%	24.4%	11.8%	6.8%	3.3%	1.6%	1.4%	0.2%	100.0%
70代	755	382	184	90	36	18	13	7	1485
	50.8%	25.7%	12.4%	6.1%	2.4%	1.2%	0.9%	0.5%	100.0%
合計	4169	1902	952	414	191	99	72	18	7817
	53.3%	24.3%	12.2%	5.3%	2.4%	1.3%	0.9%	0.2%	100.0%

付表 5：学歴ごとにみた edit エラー検出数の分布

第1次

	0	1	2	3	4	5	6-10	11以上	合計
初等教育	405	204	174	70	56	29	50	17	1005
	40.3%	20.3%	17.3%	7.0%	5.6%	2.9%	5.0%	1.7%	100.0%
中等教育	2013	944	530	260	172	89	131	30	4169
	48.3%	22.6%	12.7%	6.2%	4.1%	2.1%	3.1%	0.7%	100.0%
高等教育	1271	648	346	160	89	47	55	12	2628
	48.4%	24.7%	13.2%	6.1%	3.4%	1.8%	2.1%	0.5%	100.0%
合計	3689	1796	1050	490	317	165	236	59	7802
	47.3%	23.0%	13.5%	6.3%	4.1%	2.1%	3.0%	0.8%	100.0%

第2次

	0	1	2	3	4	5	6-10	11以上	合計
初等教育	575	244	105	49	20	11	1	0	1005
	57.2%	24.3%	10.4%	4.9%	2.0%	1.1%	0.1%	0.0%	100.0%
中等教育	2229	1004	530	217	95	45	44	11	4175
	53.4%	24.0%	12.7%	5.2%	2.3%	1.1%	1.1%	0.3%	100.0%
高等教育	1364	651	317	147	76	42	27	7	2631
	51.8%	24.7%	12.0%	5.6%	2.9%	1.6%	1.0%	0.3%	100.0%
合計	4168	1899	952	413	191	98	72	18	7811
	53.4%	24.3%	12.2%	5.3%	2.4%	1.3%	0.9%	0.2%	100.0%

付表 6：職歴段数ごとにみた edit エラー検出数の分布

第1次

	0	1	2	3	4	5	6-10	11以上	合計
なし	99	53	19	15	4	4	9	6	209
	47.4%	25.4%	9.1%	7.2%	1.9%	1.9%	4.3%	2.9%	100.0%
1職	376	133	79	31	27	8	10	4	668
	56.3%	19.9%	11.8%	4.6%	4.0%	1.2%	1.5%	0.6%	100.0%
2職	664	302	181	67	39	24	20	7	1304
	50.9%	23.2%	13.9%	5.1%	3.0%	1.8%	1.5%	0.5%	100.0%
3職	667	307	196	105	51	25	32	9	1392
	47.9%	22.1%	14.1%	7.5%	3.7%	1.8%	2.3%	0.6%	100.0%
4職	632	319	173	79	68	19	36	8	1334
	47.4%	23.9%	13.0%	5.9%	5.1%	1.4%	2.7%	0.6%	100.0%
5職	460	245	161	78	35	30	49	10	1068
	43.1%	22.9%	15.1%	7.3%	3.3%	2.8%	4.6%	0.9%	100.0%
6職	337	162	96	51	36	28	30	6	746
	45.2%	21.7%	12.9%	6.8%	4.8%	3.8%	4.0%	0.8%	100.0%
7職以上	454	276	145	67	57	28	50	10	1087
	41.8%	25.4%	13.3%	6.2%	5.2%	2.6%	4.6%	0.9%	100.0%
合計	3689	1797	1050	493	317	166	236	60	7808
	47.2%	23.0%	13.4%	6.3%	4.1%	2.1%	3.0%	0.8%	100.0%

第2次

	0	1	2	3	4	5	6-10	11以上	合計
なし	159	30	13	5	2	0	0	0	209
	76.1%	14.4%	6.2%	2.4%	1.0%	0.0%	0.0%	0.0%	100.0%
1職	410	135	73	29	12	5	3	1	668
	61.4%	20.2%	10.9%	4.3%	1.8%	0.7%	0.4%	0.1%	100.0%
2職	770	311	129	52	20	12	11	0	1305
	59.0%	23.8%	9.9%	4.0%	1.5%	0.9%	0.8%	0.0%	100.0%
3職	725	350	182	79	30	15	10	1	1392
	52.1%	25.1%	13.1%	5.7%	2.2%	1.1%	0.7%	0.1%	100.0%
4職	710	328	163	68	34	16	14	4	1337
	53.1%	24.5%	12.2%	5.1%	2.5%	1.2%	1.0%	0.3%	100.0%
5職	536	285	126	60	28	18	13	7	1073
	50.0%	26.6%	11.7%	5.6%	2.6%	1.7%	1.2%	0.7%	100.0%
6職	369	177	107	45	26	14	6	2	746
	49.5%	23.7%	14.3%	6.0%	3.5%	1.9%	0.8%	0.3%	100.0%
7職以上	490	286	159	76	39	19	15	3	1087
	45.1%	26.3%	14.6%	7.0%	3.6%	1.7%	1.4%	0.3%	100.0%
合計	4169	1902	952	414	191	99	72	18	7817

付表 7：実査期間ごとにみた edit エラー検出数の分布

第1次

	0	1	2	3	4	5	6-10	11以上	合計
第1期	1212	689	431	204	129	74	121	21	2881
	42.1%	23.9%	15.0%	7.1%	4.5%	2.6%	4.2%	0.7%	100.0%
第2期	1290	582	325	149	86	57	57	18	2564
	50.3%	22.7%	12.7%	5.8%	3.4%	2.2%	2.2%	0.7%	100.0%
第3期	1187	526	294	140	102	35	58	21	2363
	50.2%	22.3%	12.4%	5.9%	4.3%	1.5%	2.5%	0.9%	100.0%
合計	3689	1797	1050	493	317	166	236	60	7808
	47.2%	23.0%	13.4%	6.3%	4.1%	2.1%	3.0%	0.8%	100.0%

第2次

	0	1	2	3	4	5	6-10	11以上	合計
第1期	1503	688	367	167	76	45	29	7	2882
	52.2%	23.9%	12.7%	5.8%	2.6%	1.6%	1.0%	0.2%	100.0%
第2期	1393	635	310	129	53	22	21	7	2570
	54.2%	24.7%	12.1%	5.0%	2.1%	0.9%	0.8%	0.3%	100.0%
第3期	1273	579	275	118	62	32	22	4	2365
	53.8%	24.5%	11.6%	5.0%	2.6%	1.4%	0.9%	0.2%	100.0%
合計	4169	1902	952	414	191	99	72	18	7817
	53.3%	24.3%	12.2%	5.3%	2.4%	1.3%	0.9%	0.2%	100.0%

付表 8：正規・予備票ごとにみた edit エラー検出数の分布

第1次

	0	1	2	3	4	5	6-10	11以上	合計
正規	3238	1624	930	428	281	143	211	52	6907
	46.9%	23.5%	13.5%	6.2%	4.1%	2.1%	3.1%	0.8%	100.0%
予備	451	173	120	65	36	23	26	8	902
	50.0%	19.2%	13.3%	7.2%	4.0%	2.5%	2.9%	0.9%	100.0%
合計	3689	1797	1050	493	317	166	237	60	7809
	47.2%	23.0%	13.4%	6.3%	4.1%	2.1%	3.0%	0.8%	100.0%

第2次

	0	1	2	3	4	5	6-10	11以上	合計
正規	3663	1698	843	374	167	89	63	13	6910
	53.0%	24.6%	12.2%	5.4%	2.4%	1.3%	0.9%	0.2%	100.0%
予備	506	204	109	40	24	10	9	5	907
	55.8%	22.5%	12.0%	4.4%	2.6%	1.1%	1.0%	0.6%	100.0%
合計	4169	1902	952	414	191	99	72	18	7817
	53.3%	24.3%	12.2%	5.3%	2.4%	1.3%	0.9%	0.2%	100.0%

付表 9：訪問回数ごとにみた edit エラー検出数の分布

第1次

	0	1	2	3	4	5	6-10	11以上	合計
1回	739	360	231	93	62	30	40	6	1561
	47.3%	23.1%	14.8%	6.0%	4.0%	1.9%	2.6%	0.4%	100.0%
2回	1006	484	273	125	79	49	60	13	2089
	48.2%	23.2%	13.1%	6.0%	3.8%	2.3%	2.9%	0.6%	100.0%
3回	738	373	230	116	70	32	49	14	1622
	45.5%	23.0%	14.2%	7.2%	4.3%	2.0%	3.0%	0.9%	100.0%
4回	514	244	138	71	38	18	47	9	1079
	47.6%	22.6%	12.8%	6.6%	3.5%	1.7%	4.4%	0.8%	100.0%
5回	343	156	99	51	39	14	27	8	737
	46.5%	21.2%	13.4%	6.9%	5.3%	1.9%	3.7%	1.1%	100.0%
6回以上	328	169	78	34	27	23	14	10	683
	48.0%	24.7%	11.4%	5.0%	4.0%	3.4%	2.0%	1.5%	100.0%
不明	21	11	1	3	2	0	0	0	38
	55.3%	28.9%	2.6%	7.9%	5.3%	0.0%	0.0%	0.0%	100.0%
合計	3689	1797	1050	493	317	166	237	60	7809
	47.2%	23.0%	13.4%	6.3%	4.1%	2.1%	3.0%	0.8%	100.0%

第2次

	0	1	2	3	4	5	6-10	11以上	合計
1回	834	356	199	94	40	20	14	3	1560
	53.5%	22.8%	12.8%	6.0%	2.6%	1.3%	0.9%	0.2%	100.0%
2回	1100	552	228	102	57	30	19	1	2089
	52.7%	26.4%	10.9%	4.9%	2.7%	1.4%	0.9%	0.0%	100.0%
3回	862	380	225	88	39	10	15	3	1622
	53.1%	23.4%	13.9%	5.4%	2.4%	0.6%	0.9%	0.2%	100.0%
4回	579	252	135	58	25	17	9	4	1079
	53.7%	23.4%	12.5%	5.4%	2.3%	1.6%	0.8%	0.4%	100.0%
5回	396	183	91	36	14	5	12	0	737
	53.7%	24.8%	12.3%	4.9%	1.9%	0.7%	1.6%	0.0%	100.0%
6回以上	381	167	69	32	16	16	2	0	683
	55.8%	24.5%	10.1%	4.7%	2.3%	2.3%	0.3%	0.0%	100.0%
不明	16	12	4	4	0	1	1	0	38
	42.1%	31.6%	10.5%	10.5%	0.0%	2.6%	2.6%	0.0%	100.0%
合計	4168	1902	951	414	191	99	72	11	7808
	53.4%	24.4%	12.2%	5.3%	2.4%	1.3%	0.9%	0.1%	100.0%

付表 10：面接時間ごとにみた edit エラー検出数の分布

第1次

	0	1	2	3	4	5	6-10	11以上	合計
25分以内	792	383	177	79	52	25	36	15	1559
	50.8%	24.6%	11.4%	5.1%	3.3%	1.6%	2.3%	1.0%	100.0%
26-30分	845	423	223	119	62	30	35	10	1747
	48.4%	24.2%	12.8%	6.8%	3.5%	1.7%	2.0%	0.6%	100.0%
31-35分	535	242	157	58	51	28	26	1	1098
	48.7%	22.0%	14.3%	5.3%	4.6%	2.6%	2.4%	0.1%	100.0%
36-40分	587	276	177	81	48	28	40	7	1244
	47.2%	22.2%	14.2%	6.5%	3.9%	2.3%	3.2%	0.6%	100.0%
41-45分	272	142	62	41	19	16	34	7	593
	45.9%	23.9%	10.5%	6.9%	3.2%	2.7%	5.7%	1.2%	100.0%
46-50分	243	116	96	42	26	14	21	6	564
	43.1%	20.6%	17.0%	7.4%	4.6%	2.5%	3.7%	1.1%	100.0%
51-60分	227	113	77	34	23	14	22	8	518
	43.8%	21.8%	14.9%	6.6%	4.4%	2.7%	4.2%	1.5%	100.0%
61分以上	173	93	75	39	34	11	21	6	452
	38.3%	20.6%	16.6%	8.6%	7.5%	2.4%	4.6%	1.3%	100.0%
不明	15	9	6	0	2	0	2	0	34
	44.1%	26.5%	17.6%	0.0%	5.9%	0.0%	5.9%	0.0%	100.0%
合計	3689	1797	1050	493	317	166	237	60	7809
	47.2%	23.0%	13.4%	6.3%	4.1%	2.1%	3.0%	0.8%	100.0%

第2次

	0	1	2	3	4	5	6-10	11以上	合計
25分以内	947	363	151	45	26	17	8	2	1559
	60.7%	23.3%	9.7%	2.9%	1.7%	1.1%	0.5%	0.1%	100.0%
26-30分	979	407	205	85	37	20	12	2	1747
	56.0%	23.3%	11.7%	4.9%	2.1%	1.1%	0.7%	0.1%	100.0%
31-35分	582	262	140	68	31	7	7	0	1097
	53.1%	23.9%	12.8%	6.2%	2.8%	0.6%	0.6%	0.0%	100.0%
36-40分	652	318	152	68	25	16	12	1	1244
	52.4%	25.6%	12.2%	5.5%	2.0%	1.3%	1.0%	0.1%	100.0%
41-45分	286	142	89	43	14	7	10	2	593
	48.2%	23.9%	15.0%	7.3%	2.4%	1.2%	1.7%	0.3%	100.0%
46-50分	258	154	72	46	19	7	8	0	564
	45.7%	27.3%	12.8%	8.2%	3.4%	1.2%	1.4%	0.0%	100.0%
51-60分	255	132	67	29	19	8	7	1	518
	49.2%	25.5%	12.9%	5.6%	3.7%	1.5%	1.4%	0.2%	100.0%
61分以上	193	116	66	29	20	17	8	3	452
	42.7%	25.7%	14.6%	6.4%	4.4%	3.8%	1.8%	0.7%	100.0%
不明	16	8	9	1	0	0	0	0	34
	47.1%	23.5%	26.5%	2.9%	0.0%	0.0%	0.0%	0.0%	100.0%
合計	4168	1902	951	414	191	99	72	11	7808
	53.4%	24.4%	12.2%	5.3%	2.4%	1.3%	0.9%	0.1%	100.0%

資料 1

異常ケース数が 51 以上検出された edit ルール一覧

第 1 次

1.range edit（該当 edit ルール数＝1）
① [ed1811]dq30_1 の値が範囲外（case＝130）

2.filter edit（該当 edit ルール数＝19）
① [ed2581x]dq35_2 は非該当ではないはず（case＝206）
② [ed2580x]dq30_1 は非該当ではないはず（case＝134）
③ [ed2152z]q50 は非該当のはず（case＝134）
④ [ed2094x]q23_1_b_1 は非該当ではないはず（case＝129）
⑤ [ed2095x]q23_1_b_2 は非該当ではないはず（case＝114）
⑥ [ed2096x]q23_1_c_1 は非該当ではないはず（case＝99）
⑦ [ed2097x]q23_1_c_2 は非該当ではないはず（case＝99）
⑧ [ed2060z]q20_1 は非該当のはず（case＝99）
⑨ [ed2093z]q21_3_e は非該当のはず（case＝78）
⑩ [ed2573z]dq24_1 は非該当のはず（case＝64）
⑪ [ed2574z]dq24_2 は非該当のはず（case＝63）
⑫ [ed2099x]q23_2_b_2 は非該当ではないはず（case＝60）
⑬ [ed2579z]dq26 は非該当のはず（case＝59）
⑭ [ed2098x]q23_2_b_1 は非該当ではないはず（case＝58）
⑮ [ed2577z]dq25_1 は非該当のはず（case＝57）
⑯ [ed2578z]dq25_2 は非該当のはず（case＝55）
⑰ [ed2056x]q19_a は非該当ではないはず（case＝51）
⑱ [ed2058x]q19_c は非該当ではないはず（case＝51）
⑲ [ed2059x]q19_d は非該当ではないはず（case＝51）

3.general edit（該当 edit ルール数＝17）
① [ed3148]問 49 で「働いて得た収入」が 1%以上あるが、現在無職（case＝250）
② [ed3080]問 24(1)で「ずっと親と同居している」とあるが、問 42 で父とも母とも同居していない（case＝161）
③ [ed3027]問 18 で○が付いていない学校種が問 20a で回答されている（case＝133）
④ [ed3029]問 20 の学校 1 で、通常の最低年齢よりも早く入学している（case＝122）
⑤ [ed3022]問 19d で高校を卒業していないのに問 18 で専門学校に通っている（case＝115）
⑥ [ed3008]問 4 で労働組合に入っているが、問 2（現職）e でそぐわない役職を答えている（課長以上）（case＝99）
⑦ [ed3961]問 27 で結婚時の職歴番号について、職歴の在職年齢と合わない（case＝96）
⑧ [ed3145]問 42(2)から考えられる同居家族の最大人数よりも問 42(1)の回答が大きい（case＝95）
⑨ [ed3205]留問 33 で金融資産と不動産の合計の最低額よりも（ウ）の合計回答が下になっている（case＝88）
⑩ [ed3943]問 24(1)で「ずっと親と同居している」とあるが、問 23 で父も母も死亡している（case＝88）
⑪ [ed3066]問 22 で父学歴が旧制学校だが、年齢的に考えにくい（case＝79）
⑫ [ed3227]問 21(2)（父の主職）で、従業上の地位が経営者だが、役職が社長でない（case＝76）
⑬ [ed3217]問 21(1)（15 歳時父職）で、従業上の地位が経営者だが、役職が社長でない（case＝75）
⑭ [ed3013]問 10 で兄弟姉妹数の合計が合わない（case＝71）
⑮ [ed3067]問 22 で母学歴が旧制学校だが、年齢的に考えにくい（case＝70）
⑯ [ed3151]問 48 の本人収入と問 50 の配偶者収入の最低合計収入よりも問 51 の世帯収入が少ない（case＝55）
⑰ [ed3146]問 44 で坪数と平米の両方に回答があるが、対応が合わない（平米＝坪数×3.0～3.3 の範囲から逸脱）（case＝51）

第 2 次

1.range edit（該当 edit ルール数＝0）

2.filter edit（該当 edit ルール数＝1）
① [ed2087x]q21_2_d は非該当ではないはず（case＝51）

3.general edit（該当 edit ルール数＝28）

① [ed4474]結婚時の職歴（問 27）の開始年齢（問 9-XX(c)7）と、結婚年齢（問 26）が一致している。問 27 では結婚直前の職歴番号を答えなければならないが、間違いがなさそうか要確認※特に、クリーニングによる職歴の削除で段数がずれた場合には修正漏れの可能性大（case＝508）
② [ed4003]問 9-2（職歴 2）で、従業先、従業上の地位、仕事内容、役職が前職とそっくり同じ※そっくり同じ職歴段数は認めていない（case＝171）
③ [ed4004]問 9-2（職歴 3）で、従業先、従業上の地位、仕事内容、役職が前職とそっくり同じ※そっくり同じ職歴段数は認めていない（case＝160）
④ [ed4005]問 9-2（職歴 4）で、従業先、従業上の地位、仕事内容、役職が前職とそっくり同じ※そっくり同じ職歴段数は認めていない（case＝133）
⑤ [ed3343]問 9-3（職歴 3）が「現職にあたる」とあるが、従業上の地位、職業、役職のいずれかが一致しない（case＝131）
⑥ [ed3344]問 9-3（職歴 3）が「現職にあたる」とあるが、産業か規模のいずれかが一致しない※厳密には変化することもありえるが、形式上合わせているはず（case＝129）
⑦ [ed3321]問 9-2（職歴 2）が「現職にあたる」とあるが、従業上の地位、職業、役職のいずれかが一致しない（case＝117）
⑧ [ed4376]職歴（初職・現職含む）の中に、「郵便」「郵政」の記述がある。あるいは産業コードが「郵便局」。規模を確認※2007 年 10 月 1 日からは 1000 人以上。それ以前は官公庁（case＝117）
⑨ [ed3366]問 9-4（職歴 4）が「現職にあたる」とあるが、産業か規模のいずれかが一致しない※厳密には変化することもありえるが、形式上合わせているはず（case＝114）
⑩ [ed3365]問 9-4（職歴 4）が「現職にあたる」とあるが、従業上の地位、職業、役職のいずれかが一致しない（case＝110）
⑪ [ed4001]問 8（初職）が学生バイトの疑い※初職開始年齢が学卒よりも早く、かつパート／家従／DK（case＝108）
⑫ [ed3387]問 9-5（職歴 5）が「現職にあたる」とあるが、従業上の地位、職業、役職のいずれかが一致しない（case＝98）
⑬ [ed4006]問 9-2（職歴 5）で、従業先、従業上の地位、仕事内容、役職が前職とそっくり同じ※そっくり同じ職歴段数は認めていない（case＝94）
⑭ [ed3388]問 9-5（職歴 5）が「現職にあたる」とあるが、産業か規模のいずれかが一致しない※厳密には変化することもありえるが、形式上合わせているはず（case＝90）
⑮ [ed4478]問 8 で現職と初職が同じ勤め先とあるが、初職の入職経路が現職と一致しない（case＝90）
⑯ [ed3302]問 8（初職）が「現職にあたる」とあるが、従業上の地位、職業、役職のいずれかが一致しない（case＝87）
⑰ [ed3233]問 21(2)（父の主職）で、国鉄や公社だが規模が 1000 人以上になっていない※官公庁は誤り（case＝85）
⑱ [ed3223]問 21(1)（15 歳時父職）で、国鉄や公社だが規模が 1000 人以上になっていない※官公庁は誤り（case＝83）
⑲ [ed4007]問 9-2（職歴 6）で、従業先、従業上の地位、仕事内容、役職が前職とそっくり同じ※そっくり同じ職歴段数は認めていない（case＝79）
⑳ [ed4371]問 21_2（父主職）で、「郵便」「郵政」の記述がある。あるいは産業コードが「郵便局」。規模を確認※2007 年 10 月 1 日からは 1000 人以上。それ以前は官公庁（case＝78）
㉑ [ed4475]初婚時の職歴（問 35）の開始年齢（問 9-XX(c)7）と、初婚年齢（問 34）が一致している。問 28 では結婚直前の職歴番号を答えなければならないが、間違いなさそうか要確認※特に、クリーニングによる職歴の削除で段数がずれた場合には修正漏れの可能性大（case＝77）
㉒ [ed3335]問 9-3（職歴 3）で、警察や消防、自衛隊、市役所等だが規模が官公庁になっていない※産業・職業のコードと記述内容を参照（case＝72）
㉓ [ed4370]問 21_1（15 歳時父職）で、「郵便」「郵政」の記述がある。あるいは産業コードが「郵便局」。規模を確認※2007 年 10 月 1 日からは 1000 人以上。それ以前は官公庁（case＝72）
㉔ [ed3322]問 9-2（職歴 2）が「現職にあたる」とあるが、産業か規模のいずれかが一致しない※厳密には変化することもありえるが、形式上合わせているはず（case＝63）
㉕ [ed3235]問 21(2)（父の主職）で、警察や消防、自衛隊、市役所等だが規模が官公庁になっていない※産業・職業のコードと記述内容を参照（case＝58）
㉖ [ed3145]問 42(2)から考えられる同居家族の最大人数よりも問 42(1)の回答が大きい※留問 13 の子との同居も参照。問 10 の兄弟姉妹数も参照（case＝57）
㉗ [ed4008]問 9-2（職歴 7）で、従業先、従業上の地位、仕事内容、役職が前職とそっくり同じ※そっくり同じ職歴段数は認めていない（case＝56）
㉘ [ed3409]問 9-6（職歴 6）が「現職にあたる」とあるが、従業上の地位、職業、役職のいずれかが一致しない（case＝54）

資料 2

職歴関係で頻発した edit ルールの対応策に関するまとめ

ここでは、主に第 2 次クリーニング時に頻発した edit ルールについて、修正対応策も含めて紹介する。なお、本文でも述べているように、実際のクリーニング作業は、各ケースがもつ周辺情報を考慮したうえで、修正の有無や適切な対処方法を判断している。そのため、以下に示す修正対応策は、あくまでも典型例である。ケースバイケースバイで修正判断が異なるため、一律に処理されているわけではない。この点に留意すること。

① 最終職と現職の不一致

edit ルール：（職歴〇）が「現職にあたる」とあるが、従業上の地位、職業、役職のいずれかが一致しない

（職歴〇）が「現職にあたる」とあるが、産業か規模のいずれかが一致しない

・職歴の最終職と現職で、従業上の地位、規模、職業（仕事内容）、役職、産業のいずれかが一致しない。

対応：原則的に現職の情報を優先し、最終職を修正しているが、周辺情報を考慮し、以下に示すように柔軟に対応した。なお、職業（仕事内容）と産業の不一致に関しては判断を保留し、三輪哲氏に最終処理を一任した。

職歴段数を追加した事例

例）現職：課長、最終職：係長

→調査員の記入漏れと判断し、現職の情報をもとに職歴段数を追加した。

不一致を許容した事例

例）現職：1000 人以上、最終職：500～900 人

→最終職を長期継続していた場合には、規模の不一致を許容した。

② 事務職公務員の部署間異動

edit ルール：（職歴〇）で、従業先、従業上の地位、仕事内容、役職が前職とそっくり同じ

・上記の edit ルールで異常値（edit エラー）が検出された場合は、ほぼ事務職公務員の部署間異動であった（民間企業勤務でも少数ケース確認されている）。すなわち、部署のみが変化（例えば、庶務課から総務課への異動）し、地位や役職など、その他の事項に変化が認められないにも関わらず、職歴段数が 1 段追加されている。

対応：周辺情報から明らかに部署間異動のみと判断できる場合には、職歴段数を削除し統合した（公立学校勤務の教員で学校間のみの異動も同様に対応した）。

③ 退職した職場に再就職した際の規模の不一致

edit ルール：（職歴〇）で、以前と同じ従業先番号が付いているが、産業か規模のいずれかが一致しない

・一度、退職した職場に、無職期間を挟んで再就職、または、いくつかの職歴を経て再就職した場合に、従業先番号が同一であるにも関わらず、規模が一致しない。

対応：従業先番号が同一の場合、原則的に規模、産業の不一致は認められないが、離職期間中に事業規模が拡大（縮小）したと判断し、ほとんどのケースで修正は行っていない。すなわち、規模の不一致を許容した。なお、edit エラーが検出されたケースの多くは女性であり、結婚、出産を契機に退職し、子どもが就学年齢に達した頃に、元の職場にパート・アルバイトとして再就職（復帰）していた。

④ 「派遣会社」勤務の場合の派遣先異動

edit ルール：（職歴〇）で、前職から派遣のまま従業先番号が新しくなっている

・職歴データ上、1 年から 3 年程度の短期間で派遣会社を転職している（ように見える）。

対応：「派遣会社」勤務の場合は、派遣先が変更となっても従業先番号を追加しない。したがって、派遣会社そのものを変えていないと判断できる場合には、前後の職歴情報をもとに従業先番号を修正した。ただし、原票に「離職理由」が明記されている場合など、周辺情報から「別の派遣会社」に転職したと明確に判断できる場合には、従業先番号が変更されていても修正は行っていない。

⑤ 経営者/自営業主と役職の不一致

edit ルール：（職歴〇）で、従業上の地位が経営者だが、役職が社長でない

対応：役職なしの場合は社長に修正した。ただし、課長等の中間役職の場合には、原票の記載内容や周辺情報をもとに役職を社長に変更するか、もしくは、従業上の地位を経営者（自営業主）から常時雇用に修正した。なお、小規模飲食店主など一部のケースでは、従業上の地位を経営者から自営業主に変更している。

edit ルール：（職歴〇）で、従業上の地位が自営だが、役職が中間役職になっている

対応：原則として課長等の中間役職の場合は役職なしに修正した。ただし、原票の記載内容や周辺情報をもとに、従業上の地位を常時雇用に修正している場合もある。

周辺情報をもとに従業上の地位を自営業主から常時雇用に修正した事例

配偶者職：自営業主、部長、仕事内容「親が社長で跡つぎとして管理職で営業もする」

→仕事内容の記述から管理職の一般従業員（常時雇用）として勤務していることが明らかな

ので、このような場合には、従業上の地位を自営業主ではなく常時雇用に変更したほうが適切と判断し修正した。

⑥ 規模「1 人」で家族従業/被用者

edit ルール：（職歴〇）で、規模が「1 人」だが、家族従業である

（職歴〇）で、規模が「1 人」だが、被用者である

対応：自営業主（経営者）が高齢・入院等の理由により働くことが不可能なため、現状として、家族従業者（被用者）1 人になっているという可能性を排除できないので、明確な根拠がない場合は修正を行っていない。ただし、企業規模ではなく事業所規模を誤答していると判断できる場合には、規模を無回答（DK）に修正している。なお、規模が「1 人」で職業（仕事内容）が「地方議員」の場合は、従業上の地位を経営者、規模を官公庁に修正している。

類似の edit ルール：（職歴〇）で、規模が「1 人」だが、役職が中間役職になっている

対応：規定に従い社長または役職なしに修正しているが、原票の記載内容や周辺情報をもとに、規模を 1 人から無回答（DK）などに修正している場合も多い。

⑦ 初職開始年齢が学卒年齢よりも早い

edit ルール：（初職）が学生バイトの疑い

・初職の開始年齢が学卒時の年齢よりも早く、かつ、従業上の地位が「パート／家族従業者/無回答（DK)」の場合には、学生時代のアルバイトを初職と誤答している可能性が高い。

対応：卒業と同時にアルバイトを辞めている場合は、職歴とはみなさず削除した。

学生時代のアルバイトを卒業後も続けている場合には、初職とみなし修正は行っていない。

学生アルバイトにしなかった事例

専門学校在学時から行っていたコンビニでのアルバイトを卒業後も継続している。

⑧ 国鉄・郵政民営化に伴う従業先の変化

edit ルール：（職歴〇）で、「JR」の記述があり、就業時期に民営化（1987 年 4 月 1 日）の時点を含んでいる。

（職歴〇）で、「郵便」「郵政」の記述があるか産業コードが「郵便局」で、就業時期に民営化（2007 年 10 月 1 日）の時点を含んでいる。

・国鉄は 1987 年 4 月 1 日に分割民営化されたので、1987 年前後で別の従業先に変化する（職歴段数を追加する)。同様に、郵便局も 2007 年 10 月 1 日に民営化されたので、別の従業先

に変化しているのかを確認する。

対応：民営化前後で職歴段数追加の有無を確認し、必要があれば職歴段数を追加した（民営化の前後で職歴段数が適切に追加されているケースは少なかった）。

類似の edit ルール：（15 歳時父職）で、「郵便」「郵政」の記述がある。あるいは産業コードが「郵便局」

対応：就業時期が民営化後と判断できた場合には、規模が官公庁であれば 1000 人以上に修正している。

⑨ 「国鉄」勤務で規模が官公庁

edit ルール：（15 歳時父職/父の主職）国鉄や公社だが規模が 1000 人以上になっていない

対応：国鉄は民営化前も「特殊法人」であり公務員ではなかったので、規模を官公庁と回答していた場合には 1000 人以上に修正している。

⑩ 官公庁勤務の有無を確認

edit ルール：（職歴○）で、警察や消防、自衛隊、市役所等だが規模が官公庁になっていない
（職歴○）で、「国立」「公立」「県立」「市立」等の記述があるが、規模が「官公庁」でない。

対応：職業（仕事内容）の記述を確認し、官公庁勤務であると明確に判断できる場合には、規模を官公庁に修正している。なお、大多数のケースで、規模を無回答（DK）または勤務先である役所の従業員数を回答していたので、官公庁に修正している。

⑪ 管理職の職場労組への加入について

edit ルール：問 4 で労働組合に入っているが、問 2（現職）e でそぐわない役職を答えている（課長以上）

・一般的に、課長以上の管理職は職場の労働組合に加入する権利がない。

対応：クリーニング合宿時に、参加者の指摘により、「職場の労働組合に一定レベル以下の管理職が加入している場合もある」ことが判明したため、明らかな誤りと判断できるケース以外は職場内の労働組合の加入の有無については修正を行っていない。

A Basic Analysis on the Trends of Detecting Editing Errors by Focusing on the Terms of Data Cleaning Operations *

Takayuki SUGASAWA (Doshisha University)
Tokio YASUDA (Kansai University)

A new data cleaning technique proposed by Tokio Yasuda was utilized in SSM 2015 survey. The data-cleaning technique proposed by Yasuda conforms to the Fellegi and Holt paradigm. The data are imputed on a case-by-case basis, and because the work process is recorded in detail, it is easy to monitor the progress and verify the contents of the operations.

This paper is focused on the editing errors detected in the process of data cleaning conducted by us; further, we performed a basic analysis of the survey data collected from the perspective of data cleaning. First, the detailed process of data cleaning was introduced, according to the time series, and it was confirmed that the operations progressed smoothly and without much complication. Subsequently, we analyzed the trends in the detection of editing errors. The analysis revealed that the number of detected errors increased in respondents who had extensive work experience or were elderly (particularly males). The question posed here is how to collect complicated data without errors on personal history while taking the aging population into account; this needs to be tackled in the next survey. To find an answer to this difficult problem, debates on the validity of implementing CAI (Computer-Assisted Interview) are warranted.

Keywords: retrospective survey, personal history data, data cleaning

* The study was supported by JSPS KAKENHI Grant Number JP25000001.