製表技術参考資料 31

ISSN　1348－7345
ISSN－L 1348－7329

政府統計データのエディティングに関する国際的動向：

選択的エディティングの理論とソフトウェア

NSTAC

Working Paper No.31

平成 28 年 3 月

独立行政法人　統計センター

製表技術参考資料は、独立行政法人 統計センターの職員がその業務に関連して行った製表技術に関する研究の結果を紹介するためのものである。

ただし、本資料に示された見解は、執筆者の個人的見解である。

目　次

政府統計データのエディティングに関する国際的動向：選択的エディティングの理論とソフトウェア*

高橋　将宜**

要　旨

近年、コンピュータ技術の向上に伴って、公的統計のあり方は、従来のプロセス中心アプローチからデータ中心アプローチへと転換しつつある。その目的は、質を維持しながら業務を効率化することである。古今東西の公的統計において、データの品質を向上させるためにデータエディティングが活用されてきたが、人手によるデータエディティングは、非常に多くの労力を必要とする。データ中心アプローチという点から、データを重点的に審査し、高い質を維持しながらも効率的にエディティングを実行できる手法が望まれている。

そこで、本稿は、国連欧州経済委員会の統計データエディティングに関するワークセッションを中心として、公的統計におけるデータの審査と修正方法に関してサーベイを行ったものである。具体的には、エディティング、エラー、エディット規則といった概念の定義を調査し、伝統的なミクロエディティングをはじめ、マクロエディティング、双方向的な対話エディティング、自動エディティングなどの近代的な手法について調査をした。

さらに、国際的に高い評価を得ている選択的エディティングに関して、その理論と方法を直感的に説明し、具体例を用いてその有用性を例証した。スウェーデン統計局の開発した選択的エディティング専用ソフトウェア SELEKT を入手し、別冊として使用マニュアルも用意した。

本稿には、付録として、2014 年と 2015 年の統計データエディティングに関するワークセッションにおいて報告された全論文の日本語要旨を掲載し、統計データエディティングに関する最新の国際的動向を概観できるように配慮した。

---

* 本稿は、第 142 回研究報告会(総務省統計研修所、2016 年 3 月 17 日)における資料を増補・改訂したものである。本稿を執筆するにあたり、スウェーデン統計局の Magnus Ohlsson 氏には、SELEKT ソフトウェアを無償提供いただいた。また、2015 年の UNECE 統計データエディティングに関するワークセッションの参加者には、選択的エディティングについて多くの見解を示していただいた。ここに感謝の意を表したい。ただし、本稿にあり得るべき誤りはすべて執筆者に属する。本稿の内容は、執筆者の個人的見解を示すものであり、機関の見解を示すものではない。

** （独）統計センター統計情報・技術部統計技術研究課上級研究員（東洋大学経済学部非常勤講師）

政府統計データのエディティングに関する国際的動向：選択的エディティングの理論とソフトウェア

高橋　将宜

序論（研究の背景と目的）

独立行政法人統計センターは、国勢調査や経済センサスなど国の基幹的な統計調査によって収集した調査票を集計し、統計の作成を行っている。統計は、調査票の受付・入力、自由記入欄の符号化、クリーンデータの作成、結果表の作成・審査という複数の手順を踏んで作成される。この中でも、クリーンデータの作成と結果表の作成・審査に関わる部分は、諸外国において統計的データエディティングと呼ばれている部分に該当する。あらゆる調査・観測データにはエラーが付き物であり、高い品質を維持するためにはデータエディティングは欠かせない作業である。

一方、近年、財政健全化を目標とした行財政改革が進められており、公的統計制度の一翼を担う統計センターにおいても、業務プロセスの改善を通じて効率性を追求している。伝統的な人手によるデータエディティングは、非常に時間と労力がかかるものであり、コンピュータを利用してエディティングを効率化することが求められている。

そこで、本稿は、諸外国におけるエディティング手法をサーベイし、製表業務プロセスの効率化に資する材料を提供することを目的としている。具体的には、諸外国においてマクロエディティング[1]と選択的エディティング[2]と呼ばれている手法について、調査を行った。

1 エディティングとエラー

1.1 エディティングの定義

Granquist (1997, p.382)によれば、エディティングとは、「エディット規則を用いて、人手または自動的に、データ収集やデータ処理によって生じたエラーを識別し調整する手法」である。また、de Waal (2013, p.474)によれば、統計的データエディティングとは、「観測データ内のエラーを検出し、訂正するための手法」である。いずれの定義に従うとしても、エラーを探し出し、正しい値となるように修正しようという試みとして理解することができる。

ここから、エディティングの主たる目的はデータの修正であるという印象を受けるが、Granquist (1997, p.383)は、以下のとおりエディティングの 3 つの目的を指摘している。エディティングの目的とは、第一に、データ品質に関する情報を提供することである。つまり、

[1] Macro editing
[2] Selective editing（セレクティブエディティング）

あらゆる調査・観測データは、必ずエラーを含んでいるものであり、データ内にどれだけのエラーが含まれているのかを示すことが重要である。例えば、標本データは、母集団データと一致するわけではなく、標本誤差が含まれており、あたかも誤差がないかのごとくに分析を行うことは不適切であるように、調査誤差がどれぐらいあるのかを示すことは非常に重要なことである。

第二に、エディティングの目的とは、調査を将来的に改善するための基礎的な情報を提供することである。つまり、今期データにエラーが含まれている場合、どのような理由でエラーが発生したのか、原因を突き止めることによって調査設計を見直すことができ、あらかじめエラーの発生を予防することができる。

そして、最後に、エディティングの目的とは、当該データをきれいにすることである。つまり、データのクリーニングは、それ自体が最終的な目的ではなく、むしろ第一と第二の大きな目的を達成するための手段なのである。

統計的データエディティングは、先行研究においてあまり知られていない分野であるが、関連した分野として欠測値補定があり、こちらは学術的にも研究が行われている[3]。例えば、表 1 のようなデータがあるとしよう。企業 B の売上原価は空欄となっている。このようなセルのことを欠測値と呼ぶが、これは一見して明らかにエラーと認定できる。統計データエディティングでは、このような欠測値の処理に加えて、記入されているものの明らかな間違いや統計的に間違いであると推定される値の審査や訂正を行おうとするものである。

表 1：欠測データの例

企業	費用総額	売上原価	給与総額
A	100	60	40
B	80		20
C	90	80	80

1.2 エラーの種類とエディット規則

エラーには、大きく分けて 2 種類のエラーがある。1 つ目は、致命的エラー[4]と呼ばれるものである。一貫していない回答、無効な回答、項目無回答といった明らかなエラーであり、非統計的なエラーである。2 つ目のエラーは、疑わしいエラー[5]である。データ内の他の情報と大きく乖離している値や、事前に知られている情報と大きく異なっている値のことであり、統計的なエラーである(Norberg et al., 2010, p.11)。エラーの検出は、エディット規

[3] 欠測値補定については、高橋, 伊藤(2013)、高橋, 伊藤(2014)、高橋, 阿部, 野呂(2015)において詳しく取り上げているので、参照されたい。
[4] Fatal error
[5] Suspected error

則[6]によって行われる。エディット規則とは、各々のレコードにおける変数の値に関して、許容範囲を定義するものである(de Waal et al., 2011, p.10)。

ハードエディット規則[7]は、あるレコードの値が妥当なものである場合、必ず満たされるものである。致命的エラーの検出は、主にハードエディット規則によって行う。もし変数 1 と変数 2 の合計値が変数 3 と一致するなら、バランスエディットは t = 変数 3 − 変数 2 − 変数 1 という試験変数[8]に基づくこととになる。この場合、致命的エディットとは、t = 0 である。もし t ≠ 0 であれば、エディット規則が満たされていないこととなり、変数 1、変数 2、変数 3 のすべてのデータ項目がエラーとして疑われることになる(Norberg et al., 2010, p.12)。例えば、表 2 のように、費用総額 = 売上原価 + 給与総額という式が成り立つとしよう。その場合、企業 C のデータは、費用総額 90 < 売上原価 80 + 給与総額 80 となり、ハードエディット規則に適合しない。よって、企業 C のレコードのどこかにエラーが含まれていることが論理的に導き出される。このようなハードエディット規則は、論理式に基づくため、比較的簡単にコンピュータによって自動化できる。

表 2：ハードエディット規則の例

企業	費用総額	売上原価	給与総額
A	100	60	40
B	80	60	20
C	90	80	80

ソフトエディット規則[9]は、ありそうにない値や外れている値を検出するもので、疑わしいエラーの検出に用いられる。ただし、この検出は、論理的必然性に基づくわけではなく、統計的蓋然性に基づくものである。例えば、表 3 では、2016 年において企業 B の売上高/従業員数は 900/9 = 100 であるのに対して、2015 年において企業 B の売上高/従業員数は 80/9 = 8.889 である。このように、ある企業における従業員 1 人あたりの売上高が前年の値の 10 倍を超える場合、エラーではないかと疑うといった検出方法がソフトエディット規則にあたる。すなわち、ソフトエディット規則を満たさないことは、必ずしもエラーであることを論理的に保証はしないものの、さらなる検証を進めるきっかけとなる(de Waal et al., 2011, p.11)。

[6] エディット規則とは、edit rule の訳語である。なお、英語では、edit rule を省略して単に edit と言うこともある。
[7] Fatal edit とも言う。
[8] Test variable
[9] Query edit とも言う。

表 3：ソフトエディット規則の例

企業	売上高 2016	従業員数 2016	売上高 2015	従業員数 2015
A	110	11	100	11
B	900	9	80	9
C	100	10	90	10

なお、他にも体系的エラー[10]とランダムエラー[11]という種類のエラーにも注意が必要である。これらのエラーについては、高橋(2013, pp.10-12)において詳しく取り上げたので、そちらを参考にされたい。

2 様々なエディティング手法

2.1 ミクロエディティング

Granquist (1997, pp.382-383)は、ミクロエディティングを「個票データレコードの妥当性や一貫性を保証するプロセス」と定義している。そもそも、あるデータに対して 1 つひとつの値が正しいかどうかを確認しようと考えるのが直感的な対応方法であり、1980 年以前において伝統的に行われていたエディティングである。つまり、元々はミクロエディティングという特別な手法があったわけではなく、1980 年以前においてエディティングと言った場合には、「会計士的な視点」によって、データ内のすべてのエラーを検出し「訂正」することを目標としていた。特に、今日的な意味合いではミクロエディティングとは、人手によって 1 つずつの値をユニットレベルですべて確認し、照会を通じて訂正しようとするものである(Norberg et al., 2010, p.3, p.11)。

しかし、現在では、このような手法は非効率的であることが知られている。つまり、最終集計結果にほとんど影響を及ぼさない多数のエラーの訂正に多額の予算を注ぎ込むことは、現代社会では正当化されないと考えられている。また、このような伝統的ミクロエディティングは、時として、オーバー・エディティングの問題を引き起こす。すなわち、ありそうにないが正しいデータを、間違っているがよりありそうな値に変更してしまう問題である(de Waal, 2013, p.476)。

このような背景において、回答者への照会を行う主な目的はエラーの原因と報告者のキャパシティを知ることが目的であって、個別のケースをクリーンにすることではないと Granquist (1997, p.385)は述べている。

[10] Systematic error
[11] Random error

2.2 マクロエディティング

ミクロエディティングは個別のデータを単位としてエラーに対処するのに対して、マクロエディティングは集計値を単位としてエラーに対処する手法である。マクロエディティングは、1980 年代に提唱されたエディティング手法であり、Granquist (1991)によれば、主だった手法だけでも 6 つの方法が確認されている[12]。本稿では、その中でも最もベーシックな 2 種類を取り上げる。

1 つ目は、集計値に基づく手法である。これは、すべての公的統計機関が公表前に通常行っている類の審査を形式化・体系化したもので、公表予定の結果表数値を前回の表の数値と比較して、妥当かどうかを確認するものである(de Waal et al., 2011, pp.209-210)。

2 つ目は、分布に基づく手法である。利用可能なデータを用いて対象となる変数の分布を推定し、すべての個別データを分布と比較して検討する。一般的には、平均値や中央値といった中心傾向を表す代表値や、標準偏差や四分位偏差といったばらつきを表す代表値を算出する。分布に対して普通ではないとみなされる値を持つレコードは、さらなる検証の候補として検出される(de Waal et al., 2011, pp.210-212)。以後、本稿におけるマクロエディティングは、このタイプのものを意味する。

	A	B
1	id	身長
2	1	173.8
3	2	161.8
4	3	171.7
5	4	170.8
6	5	182.6
7	6	174.8
8	7	162.2
9	8	169.2
10	9	205.4
11	10	172.9
12	11	165.1
13	12	170.1
14	13	162.6
15	14	166.6
16	15	177.3
17	16	160.2
18	17	173.1
19	18	170.9
20	19	170.9
21	20	163.3
22		

図 1：マクロエディティングの例

[12] なお、かつて、マクロエディティングは選択的エディティングの一種と見なされていたことがあったが、現在では、選択的エディティングという用語は、エラー検出のプロセスにおける優先付けを自動化する手法として、別の手法と認識されている(de Waal, 2013, p.476)。

例えば、図 1 のように 20 人の身長に関するデータがあるとしよう。中央値は 171cm、第 1 四分位値は 165cm、第 3 四分位値は 173cm、よって四分位偏差(IQR)は 8cm である。中央値から 1.5×四分位偏差の範囲外にある値は、データの分布に対して普通ではないとみなされ、今回の場合は、ID9 の値がさらなる検証の対象として検出される。

2.3 双方向的な対話エディティング

双方向的な対話エディティング[13]は、コンピュータによる補助を受けながら、エラーデータの訂正を行うものである。経済調査を担当する職員であったり、人口調査を担当する職員であったり、個別の分野に精通している専門職員は、当該の分野に関して広範な知識を持っている。エディティングにおいては、こういった専門職員の知識を可能な限り活用すべきである。具体的な手法は、伝統的な人手によるミクロエディティングと同じで、回答者への照会、回答者のデータを前期データと比較、回答者のデータを他の似通ったデータと比較し、コンピュータによる補助を受けながら、専門家が人手により訂正するものである。今日では、双方向的な対話エディティングは、データエディティングの標準的な手法として活用されている(de Waal et al., 2011, p.15)。

具体的には、コンピュータにより一貫性審査[14]を行い、エディット規則を満たしていないレコードをリストアップする。その後、各調査の専門職員が人手により直接的にデータを修正する。修正が行われた場合には、コンピュータによって即座に一貫性審査が行われ、修正によってエディット規則が満たされるようになったかどうかを確認する(de Waal et al., 2011, p.213)。

2.4 自動エディティング

自動エディティング[15]のプロセスは、通常、2 つのステップを用いて自動化される。まず、エラー特定ステップにおいてエラーの検出を行う。次に、補定ステップにおいて、エラーデータを補定値に置き換える。エラー特定ステップは、決定論的なハードエディット規則に基づく手法、統計モデルによるソフトエディット規則に基づく手法、数理的最適化問題に基づく手法に分類できる。決定論的なハードエディット規則に基づく手法では、あるレコードの値に一貫性があるかどうかを確認し、一貫性のないレコードにはエラーが含まれていると判断する。統計モデルによるソフトエディット規則に基づく手法は、外れ値検出法に基づいて、エラーの大部分と異なる傾向を示す値をエラーの候補と判断して検出する(de Waal et al., 2011, pp.57-58)。数理的最適化問題に基づく手法については、Arbués et al. (2015)を参照されたい。補定に関しては、高橋, 伊藤(2013)、高橋, 伊藤(2014)及び高橋, 阿部, 野呂

[13] Interactive editing
[14] Consistency check
[15] Automatic editing

(2015)を参照されたい。

3 選択的エディティング

3.1 選択的エディティングの理論と方法（直感的な説明）

データエディティングにおいて、すべてのデータを細部にわたるまで審査・訂正する必要がないことは、1950 年代頃から指摘され始めていた(de Waal, 2013, p.477)。しかし、当時はまだコンピュータが汎用的に活用できず、ミクロエディティングの手法が主流であった。実際に選択的エディティングの議論が活発になったのは、1990 年代に入ってからのことである。

選択的エディティングは、影響力の高いエラーや外れ値を検出する方法に関する包括的な用語である。つまり、選択的エディティングと一言で表しても、諸外国において実装されている手法には様々なものがある(高橋, 2012, pp.5-6; 高橋, 2013, pp.49-52)。今日において選択的エディティングと呼ばれている手法に共通する重要な点として、Latouche & Berthelot (1992)の提唱したスコア関数を活用している点を指摘できる。最も外れた値から始めて、推定値の変化が見られなくなった段階で修正をストップすることが、大原則である(Granquist, 1997, p.384)。

一般論として、選択的エディティングでは、エラー特定手法を用いて、あるレコードがエラーである可能性と最終集計値への影響度の 2 つを計算しスコアを算出することで、優先的にエディティングするべきエラー候補を抽出する。その上で、双方向的な対話エディティングを行ってエラーを訂正する。このようにすることで、限られた時間と予算の制約の中で、最終集計値の品質に重大な影響を及ぼすエラーをもれなく修正することができる(de Waal et al., 2011, p.16)。

この手法は、de Waal (2013, p.479)が指摘するとおり、「常識に基づく比較的単純な手法」である。以下では、3 人の所持金に関する小さなデータを用いて、直感的に説明する。状況設定として、3 つの値の合計値を集計することを目的とし、予算の都合上、人手による審査と訂正ができるのは 1 つの値だけとする。当然ながら、真値は、実際には不明である。

表 4：例示用データ

人名	実測値	真値
鈴木	10 円	1 円
佐藤	10000 円	1000 円
田中	5000 円	5000 円
合計	15010 円	6001 円

3 つのうち 1 つのみ審査と訂正が行えるので、何らかの基準を設けて優先付けを行う必要

性がある。そこで、どの値がエラーである可能性（確率）が高いかに注目する方法が考えられるだろう。よりエラーである可能性が高いものから順番にエディティングしていく方法である。ここでは、どのようにして「エラー確率」を計算するかについて深く考えず、とりあえず表 5 のようにエラー確率が分かったとして話を進める[16]。すると、鈴木さんの値は、エラー確率が 0.99 で最も高く、この値は極めて高い確率でエラーだと疑われる。よって、エラー確率に着目してエディティングを行うならば、鈴木さんの値 10 円を審査し、真の値である 1 円に修正する。その結果、合計値は、15010 円から 15001 円に修正される。

表 5：エラー確率

人名	値	確率
鈴木	10 円	0.99
佐藤	10000 円	0.90
田中	5000 円	0.05
合計	15010 円	

別の基準として、エラーの影響度に注目する方法が考えられる。つまり、ある値がエラーだった場合に、その値が合計値に与える影響の大きい順番にエディティングしていく方法である。仮にエラーが真値の 10 倍のマグニチュードで発生しているとしよう。これは、一般的に、数字の記入や入力の際に桁を間違える行為と同じである。表 6 のとおり、1 つずつの値が 10 倍の大きさだった場合に与えるエラーを見ていくと、田中さんの値の影響度が最も高いことが分かる。

表 6：影響度

人名	値	正誤
鈴木	1 円	正
佐藤	1000 円	正
田中	5000 円	正
合計	6001 円	

人名	値	正誤
鈴木	10 円	誤
佐藤	1000 円	正
田中	5000 円	正
合計	6010 円	影響：小

人名	値	正誤
鈴木	1 円	正
佐藤	10000 円	誤
田中	5000 円	正
合計	15001 円	影響：中

人名	値	正誤
鈴木	1 円	正
佐藤	1000 円	正
田中	50000 円	誤
合計	51001 円	影響：大

つまり、田中さんの値は、もしエラーだった場合、最終集計値に与える影響度が大きいので、田中さんの値 5000 円を審査し、真の値である 5000 円に修正する。その結果、合計値

[16] 後ほど、具体的に数式を用いながら理論的な議論をする。ここでは、選択的エディティングの直感的なメカニズムに注目して話を進める。

は、15010 円から 15010 円に修正される[17]。

ここまで見てきたとおり、エラー確率とエラーの影響度を別々に測定した場合、エラーである可能性は高いが影響の低いもの、エラーとしての影響度は高いもののエラーではないものなどを訂正し、効率的なエディティングが行えないことが分かる。

そこで、選択的エディティングでは、エラー確率とエラーの影響度を同時に考慮し、エラーの確率が高く、かつ、影響度が高い値を優先的に修正する。われわれの例では、佐藤さんの値は、エラーである可能性が高く、かつ、影響度が高いため、この値を優先的に審査・訂正するべきである。その結果、最終集計値は、15010 円から 6010 円となり、真値の 6001 円とほぼ変わらない状態になるのである。

表 7：影響度とエラー確率

人名	値	確率
鈴木	10 円	0.99
佐藤	**10000 円**	**0.90**
田中	5000 円	0.05
合計	15010 円	

3.2 選択的エディティングの理論と方法（メカニズム）

ここまでの議論を定式化[18]すると、式(1)のローカルスコアとなる。ローカルスコアは、影響度とリスクの積として定義し、この値が大きなものから順番に修正を行っていく。なお、ここで、i はユニット、j は変数を表す。

$$\text{ローカルスコア}_{ij} = \text{リスク}_{ij} \times \text{影響度}_{ij} \tag{1}$$

リスク[19]は、潜在的なエラーの確からしさを測定するもので、前節において「エラー確率」と呼んでいた概念に相当する。リスクは、式(2)のように、「観測値」と「期待される値[20]」との差の絶対値の比率として推定する。ここで、y_{ij}はユニット i における変数 j の値であり、$\hat{y}_{ij}$はその「期待される値」である。

$$\text{リスク}_{ij} = \frac{|y_{ij} - \hat{y}_{ij}|}{|\hat{y}_{ij}|} \tag{2}$$

---

17 ただし、田中さんの値は実際にはエラーではないため、集計値に変化はない。

18 本節の内容については、de Waal (2013, pp.479-481)も参照されたい。

19 スウェーデン統計局では、この概念を suspicion と呼んでいる。

20 Anticipated value の訳語である。一般的に、「期待される値」は、補助変数の関数としてモデリングされる。例えば、補助変数を説明変数として用いた回帰分析における被説明変数の予測値などである。他にも、前期データの値、税務データなど外部データの値を用いるコールドデックが使用されることもある。補助変数とモデルパラメータの推定値は、通常、エディット済みの前期データなどから入手することが多い。

影響度は、式(3)のように、対象となる変数の合計推定値に対する相対的な影響度を測定するものである。ここで、w_iはユニットｉのウェイト[21]を表す。

$$影響度_{ij} = w_i|\hat{y}_{ij}| \tag{3}$$

このようにして算出したローカルスコアを、レコード全体に関してエディティングの必要性を測るために統合したものをグローバルスコアと呼ぶ。なお、スコアを統合するには、ローカルスコアが同等のスケールで評価される必要があるため、合計値で割ったり、標準偏差で割ったり、何らかの標準化を実施した上で統合する必要がある。

4 エディティングの具体例

表 8 は、EDINET データをもとに作成したシミュレーションデータである。単位は 100 万円である。

表 8：シミュレーションによる事業所・企業データ

	A	B	C	D	E	F
1	id	費用(前期)	売上(前期)	誤差項	売上(今期)	乱数
2	1	27400.49067	40921.00828	69704.13609	110625.1444	0.027039
3	2	80086.10281	74670.41412	−49815.08501	24855.32911	0.071322
4	3	3476.370136	7958.409338	133348.271	141306.6804	0.222663
5	4	4433.392543	8177.589124	67163.72606	75341.31519	0.235786
6	5	24955.85143	31915.33263	52343.82019	84259.15282	0.276254
7	6	2924.239538	2613.263162	1541.59352	4154.856681	0.286691
8	7	41777.42789	32849.70652	−19.77582542	32829.9307	0.373028
9	8	1702.770423	2862.240836	−1410.503501	1451.737335	0.457533
10	9	14621.7907	20154.24431	1793.973752	21948.21806	0.495956
11	10	1346.977425	3459.9163	533.2958608	3993.212161	0.527787
12	11	4227.930458	6149.407496	907.3755791	7056.783075	0.553545
13	12	16099.97939	51722.87453	−151.6252723	51571.24926	0.586413
14	13	13370.195	23977.29404	−593.4600722	23383.83397	0.590228
15	14	17115.66307	22040.67811	−368.4385774	21672.23953	0.663289
16	15	103204.2744	130775.6946	695.8453014	131471.5399	0.686544
17	16	18424.09189	62838.42573	1320.186129	64158.61186	0.715567
18	17	204526.0213	232923.5693	−1339.258233	231584.3111	0.870846
19	18	321029.7606	317924.2386	−1416.319719	316507.9189	0.913877
20	19	3246.702452	7363.488519	467.8213372	7831.309856	0.951537
21	20	5530.359431	14222.57536	−1137.527761	13085.0476	0.972472
22						
23	平均値	45475.01958	54776.01855	13678.40254	68454.42109	
24	標準偏差	81144.27087	82816.69008	38829.98115	83553.31563	

[21] ウェイトは、包含確率と無回答に対する補正なので、データ収集が完了し、無回答の推定が行われた段階にならなければ使用できない。そこで、プロキシーが必要となり、一般的にはデザインウェイトによって代替する。つまり、包含確率の逆数によって包含確率のみを補正するものである。なお、もし標本抽出が単純無作為抽出ならば、包含確率は１であり、ウェイトは無視できる。

前期データは修正済みのデータ（エラーなし）、今期データは修正前のデータ（エラーあり）を表しているものとする。われわれの目的は、今期の売上高の平均値（合計値）を算出することである。このデータの基本統計量は、表 9 に示すとおりである。

表 9：基本統計量

	費用(前期)	売上(前期)	売上(今期)
平均	45475.0196	54776.02	68454.42
標準誤差	18144.4106	18518.37	18683.09
中央値（メジアン）	15360.885	23008.99	28842.63
最頻値（モード）	#N/A	#N/A	#N/A
標準偏差	81144.2709	82816.69	83553.32
分散	6584392695	6.86E+09	6.98E+09
尖度	7.20384456	5.497339	3.355984
歪度	2.67462278	2.397156	1.848145
範囲	319682.783	315311	315056.2
最小	1346.97743	2613.263	1451.737
最大	321029.761	317924.2	316507.9
合計	909500.392	1095520	1369088
標本数	20	20	20

図 2 は費用(前期)の分布である。図 3 は売上(前期)の分布である。図 4 は売上(今期)の分布である。いずれのデータも、経済データによくあるように、右に歪んだ分布である。また、単変量の分布を見るだけでは、どこにエラーがあるのか判別することは困難である。

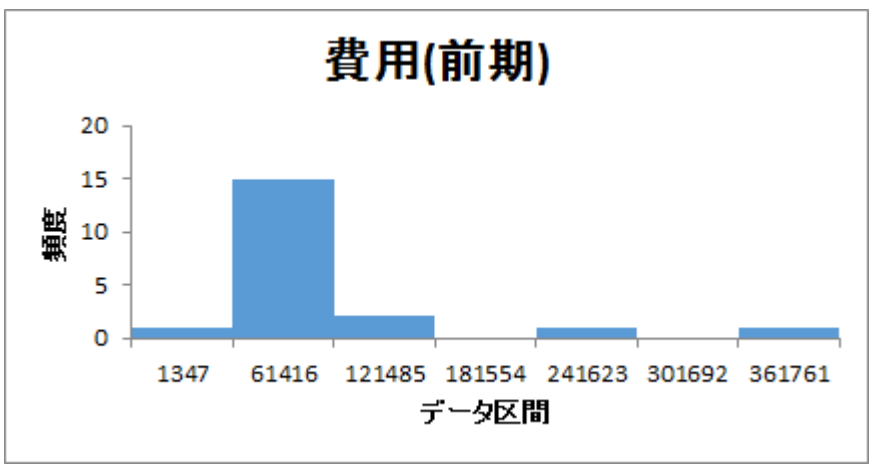


図 2：費用(前期)のヒストグラム

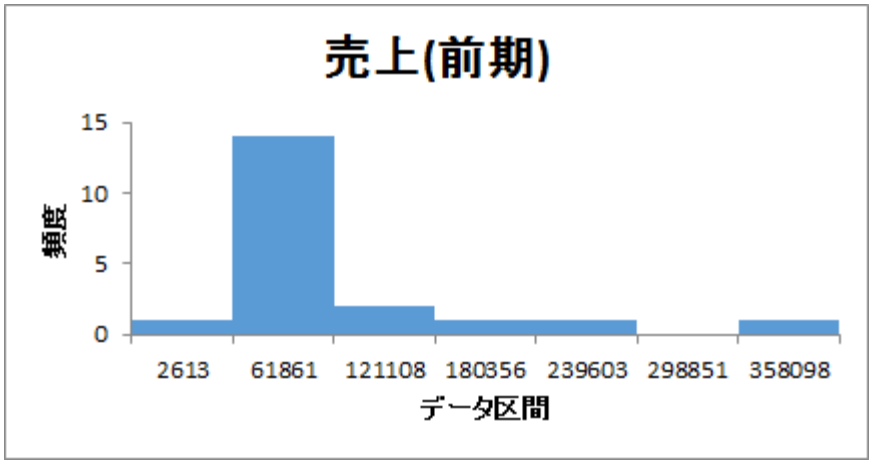


図 3：売上(前期)のヒストグラム

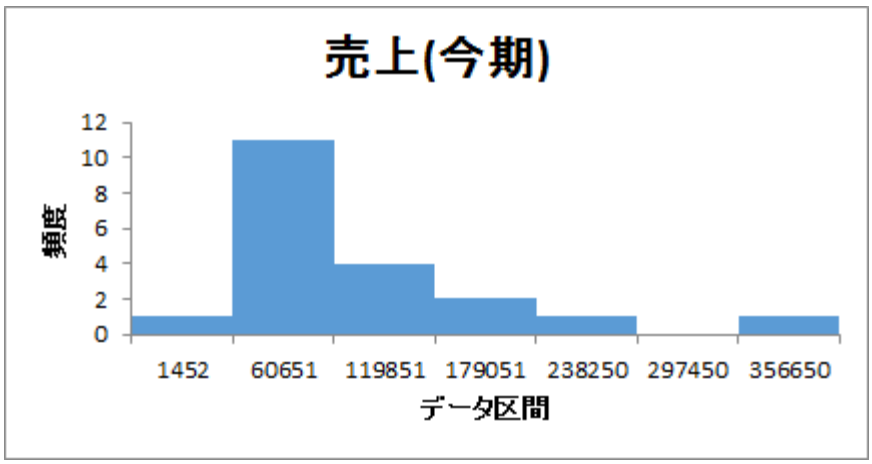


図 4：：売上(今期)のヒストグラム

4.1 ランダムエディティング

本節では、人手によるエディティングを模した結果を示す。実際には、人手によるエディティングにおいても何らかの基準を用いてエディティングを行っていると考えられる。例えば、調査は都道府県単位で行われるため、調査票は都道府県ごとに送られてくる。この場合、人手によるエディティングを行う順番は、早く調査の終わった都道府県から行うこととなる。しかし、その順番は、エラーの重要性とは必ずしも関係がない。そこで、本節では、具体的なエディティングの優先付けを行う理由が不明な場合を想定して、乱数によってでたらめな順番でエディティングをした場合について例証する。

表 10：人手によるエディティングの例

	A	B	C	D	E	F	G
1	id	売上(前期)	売上(今期)	乱数		訂正数	
2	8	2862.240836	1451.737335	0.058015687		0	13678.4
3	20	14222.57536	13085.0476	0.064058351		1	13748.93
4	4	8177.589124	75341.31519	0.069673757		2	13805.8
5	5	31915.33263	84259.15282	0.132602924		3	10447.62
6	13	23977.29404	23383.83397	0.237372967		4	7830.427
7	16	62838.42573	64158.61186	0.357005524		5	7860.1
8	11	6149.407496	7056.783075	0.392986847		6	7794.09
9	6	2613.263162	4154.856681	0.461561937		7	7748.722
10	18	317924.2386	316507.9189	0.470961638		8	7671.642
11	14	22040.67811	21672.23953	0.501266518		9	7742.458
12	12	51722.87453	51571.24926	0.624897		10	7760.88
13	10	3459.9163	3993.212161	0.637226478		11	7768.461
14	17	232923.5693	231584.3111	0.650746178		12	7741.796
15	7	32849.70652	32829.9307	0.655720695		13	7808.759
16	1	40921.00828	110625.1444	0.675374615		14	7809.748
17	19	7363.488519	7831.309856	0.810968352		15	4324.541
18	2	74670.41412	24855.32911	0.887356182		16	4301.15
19	15	130775.6946	131471.5399	0.918607135		17	6791.905
20	9	20154.24431	21948.21806	0.921445357		18	6757.112
21	3	7958.409338	141306.6804	0.955107273		19	6667.414
22						20	0
23	平均	54776.01855	68454.42109	0.524147771			

その結果は、図 5 に示すとおりである。エラーを取り除くには、20 個すべてのデータを確認しなければならないことが分かる。使用した乱数を変更した結果が図 6 である。ランダムにエディティングを実行した場合、偶発的に効率よく行える場合もあれば、効率が非常に悪い場合もある。おおむね、すべてのデータを人手により審査・修正しない限り、データ品質は十分ではない。

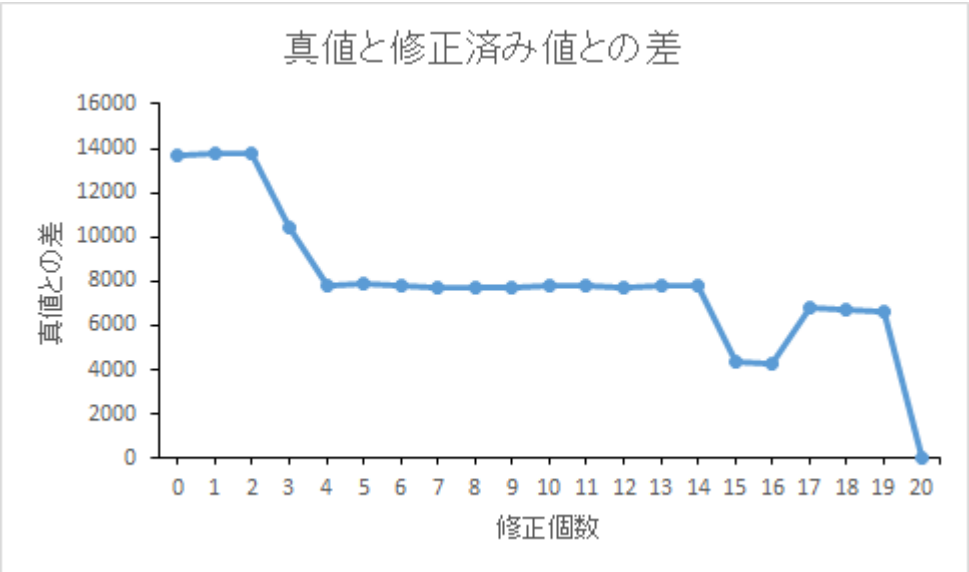


図 5：人手によるエディティングの効率性

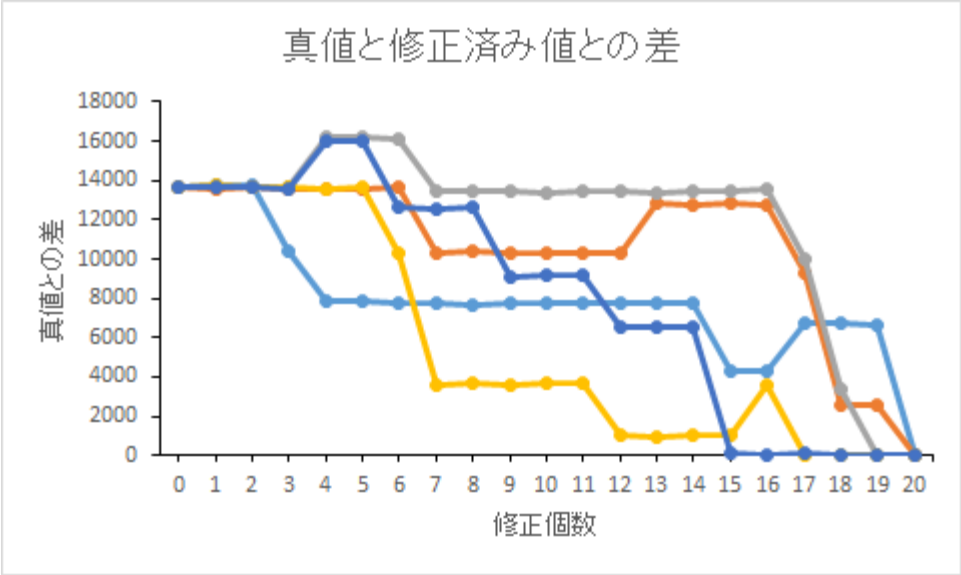


図 6：人手によるエディティングの効率性（複数のケース）

そこで、何らかの基準を設けてエディティングを行っていく必要がある。次の節では、マクロエディティングを用いた例を示す。

4.2 マクロエディティング

箱ひげ図を利用して、マクロエディティングを行う。中央値±1.5×IQR を超える値は、異常な値として検出することとする。これは、一種のエラーの影響度のみに注目したエディティングと言える。

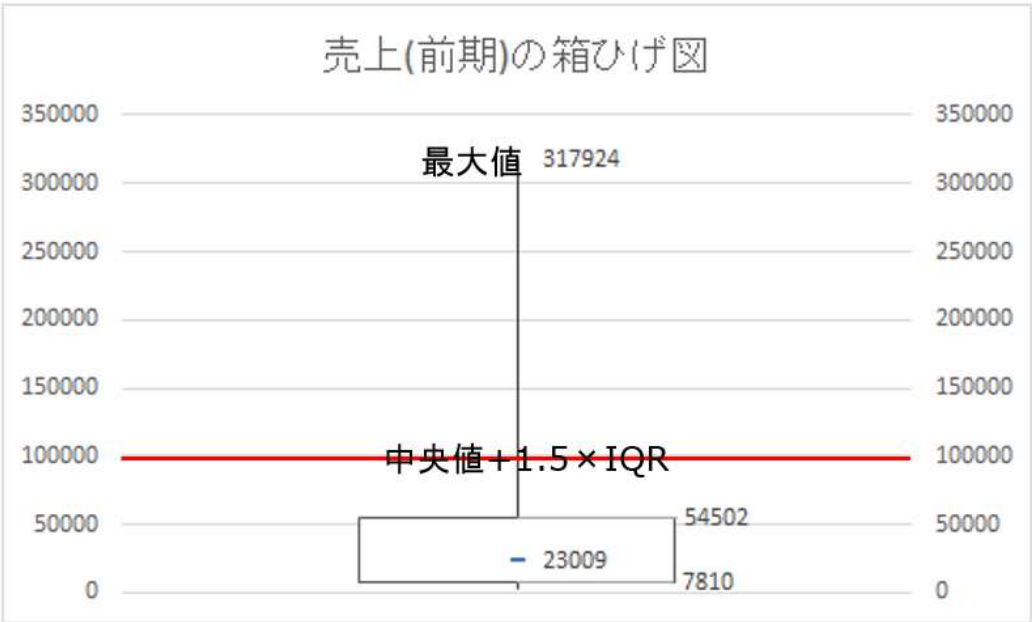


図 7：箱ひげ図によるマクロエディティング

表 11：マクロエディティングの例

	A	B	C	D	E	F	G	H
1	id	売上(前期)	売上(今期)	1.5IQR上限	1.5IQR下限		訂正数	
2	18	317924.2386	316507.9189	-223460.808	363537.058		0	13678.4
3	17	232923.5693	231584.3111	-138537.200	278613.450		1	13749.22
4	3	7958.409338	141306.6804	-48259.570	188335.819		2	13816.18
5	15	130775.6946	131471.5399	-38424.429	178500.679		3	7148.768
6	1	40921.00828	110625.1444	-17578.034	157654.283		4	7113.976
7	5	31915.33263	84259.15282	8787.958	131288.292		5	3628.769
8	4	8177.589124	75341.31519	17705.796	122370.454		6	1011.578
9	16	62838.42573	64158.61186	28888.499	111187.751		7	2346.608
10	12	51722.87453	51571.24926	41475.862	98600.388		8	2412.618
11	7	32849.70652	32829.9307	60217.180	79859.069		9	2405.037
12	2	74670.41412	24855.32911	68191.782	71884.468		10	2404.048
13	13	23977.29404	23383.83397	69663.277	70412.973		11	86.70651
14	9	20154.24431	21948.21806	71098.893	68977.357		12	116.3795
15	14	22040.67811	21672.23953	71374.871	68701.378		13	26.68082
16	20	14222.57536	13085.0476	79962.063	60114.186		14	45.10275
17	19	7363.488519	7831.309856	85215.801	54860.449		15	101.9791
18	11	6149.407496	7056.783075	85990.328	54085.922		16	78.58807
19	6	2613.263162	4154.856681	88892.254	51183.995		17	33.21929
20	10	3459.9163	3993.212161	89053.899	51022.351		18	43.86038
21	8	2862.240836	1451.737335	91595.374	48480.876		19	70.52518
22							20	0
23	平均	54776.01855	68454.42109	24592.63978	115483.5598			

この値に応じて並べ替える。

ランダムエディティングと比較すると非常に効率的な結果だが、エラーの可能性（リスク）は無視しているので、無駄な作業が発生している。

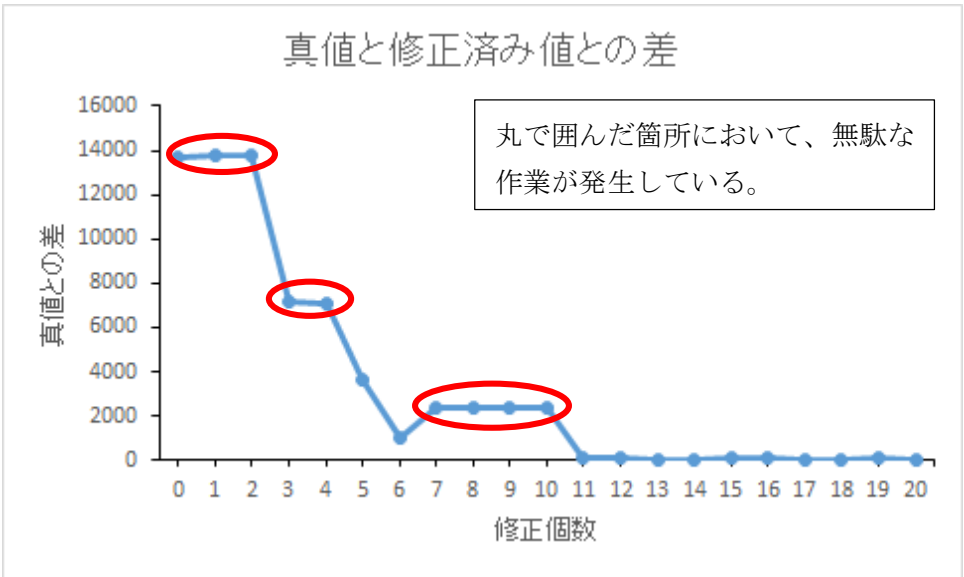


図 8：マクロエディティングによる効率性

4.3 選択的エディティング

原理的には、横軸に前期の売上高を取り、縦軸に今期の売上高を取ることで、外れている値をエラーとして検出し、効率よくエディティングを行うことができる。しかし、今期の売上高のデータは、エディティングを実行する際には、収集中なので、データが揃っていない。すなわち、このような形でエディティングを実施することは、机上の空論である。

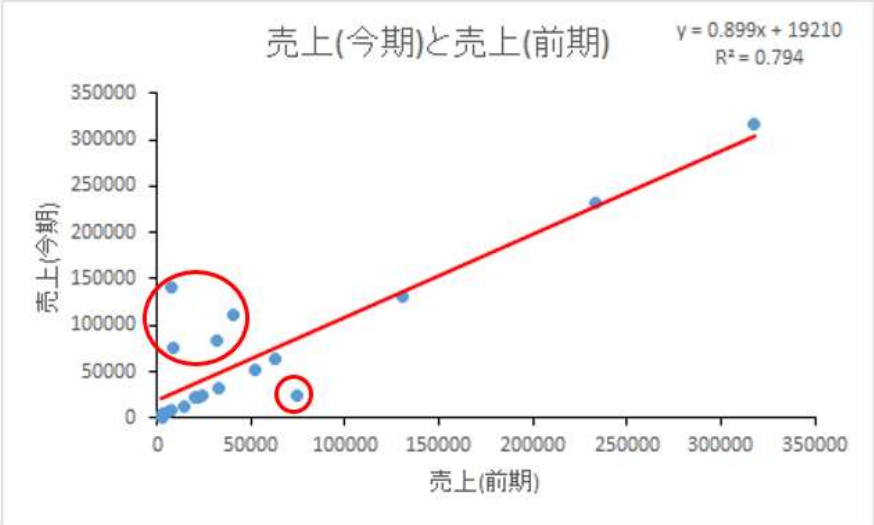


図 9：理論上の選択的エディティング

そこで、前期のデータをプロキシーとして用いて、回帰パラメータの推定を行う。なお、ここでは簡単のため二変数モデルを用いているが、二変数で十分でない場合は、重回帰モデルなど、複数の説明変数を用いるべきである。また、使用するモデルは、回帰モデルに限定されるものではない。優れた「期待される値」を算出できる手法であれば、どのような統計モデルを組み込むかは、エディティング担当者の責任である。

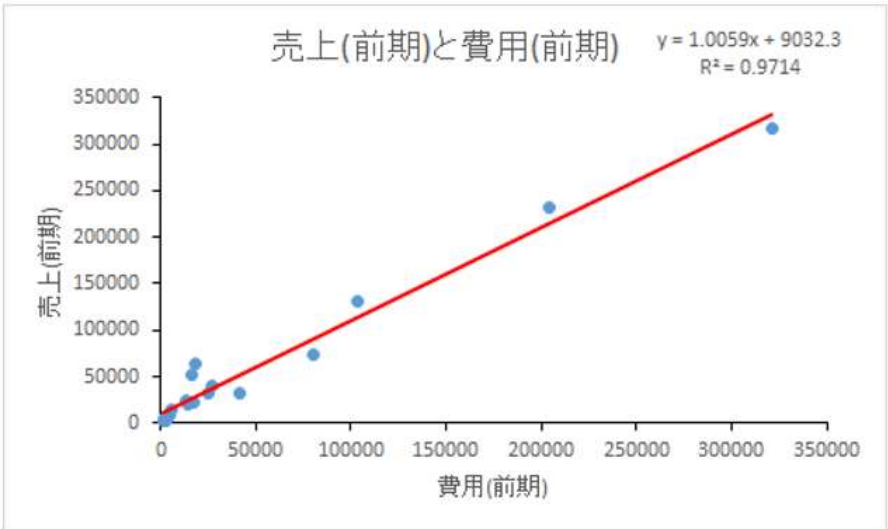


図 10：プロキシーを用いた選択的エディティング

ここでは、仮に、売上高(前期)を被説明変数とし、費用(前期)を説明変数として、切片のない単回帰モデルによってパラメータを推定したとする。すなわち、推定式は以下のとおりである。売上高の予測値 = 0 + 1.055*費用

表 12：モデルパラメータの推定

回帰統計				
重相関 R	0.986900894			
重決定 R2	0.973973375			
補正 R2	0.921341796			
標準誤差	16146.42219			
観測数	20			
分散分析表				
	自由度	変動	分散	観測された分散比
回帰	1	1.85368E+11	1.85368E+11	711.0216713
残差	19	4953432044	260706949.7	
合計	20	1.90322E+11		
	係数	標準誤差	t	P-値
切片	0	#N/A	#N/A	#N/A
費用(前期)	1.055258484	0.039574682	26.66498962	1.6162E-16

表 13 は、表 12 において算出したパラメータを用いて、式(1)、(2)、(3)によってスコア(Score)、リスク(Risk)、影響度(Influence)を算出したものである。スコアの大きさに応じて並べ替えてエディティングを行う。

表 13：選択的エディティング

	A	B	C	D	E	F	G	H	I	J
1	id	費用(前期)	売上(前期)	売上(今期)	Influence	Risk	Score		訂正数	
2	3	3476.370136	7958.409338	141306.6804	3668	37.519	137638		0	13678.4
3	1	27400.49067	40921.00828	110625.1444	28915	2.826	81711		1	7010.989
4	4	4433.392543	8177.589124	75341.31519	4678	15.104	70663		2	3525.782
5	2	80086.10281	74670.41412	24855.32911	84512	0.706	59656		3	167.5959
6	5	24955.85143	31915.33263	84259.15282	26335	2.200	57924		4	2658.35
7	16	18424.09189	62838.42573	64158.61186	19442	2.300	44716		5	41.15913
8	12	16099.97939	51722.87453	51571.24926	16990	2.035	34582		6	24.85018
9	15	103204.2744	130775.6946	131471.5399	108907	0.207	22564		7	17.26892
10	18	321029.7606	317924.2386	316507.9189	338769	0.066	22261		8	52.06118
11	17	204526.0213	232923.5693	231584.3111	215828	0.073	15756		9	18.7548
12	7	41777.42789	32849.70652	32829.9307	44086	0.255	11256		10	85.71772
13	13	13370.195	23977.29404	23383.83397	14109	0.657	9275		11	86.70651
14	20	5530.359431	14222.57536	13085.0476	5836	1.242	7249		12	116.3795
15	9	14621.7907	20154.24431	21948.21806	15430	0.422	6518		13	173.2559
16	19	3246.702452	7363.488519	7831.309856	3426	1.286	4405		14	83.55721
17	14	17115.66307	22040.67811	21672.23953	18061	0.200	3611		15	60.16614
18	11	4227.930458	6149.407496	7056.783075	4462	0.582	2595		16	78.58807
19	10	1346.977425	3459.9163	3993.212161	1421	1.809	2572		17	33.21929
20	6	2924.239538	2613.263162	4154.856681	3086	0.346	1069		18	6.554501
21	8	1702.770423	2862.240836	1451.737335	1797	0.192	345		19	70.52518
22									20	0
23	平均値	45475.01958	54776.01855	68454.42109	47987.90021	3.501429645	29818.40588			
24	標準偏差	81144.27087	82816.69008	83553.31563	85628.18023	8.64845025	35944.93775			
25	切片	0								
26	回帰係数	1.055258484								

図 11 は、上記のスコアに従ってエラーを優先的にエディティングした際の効率性を示している。極めて正確、かつ、効率的にエラーを除去できている様子が分かる。

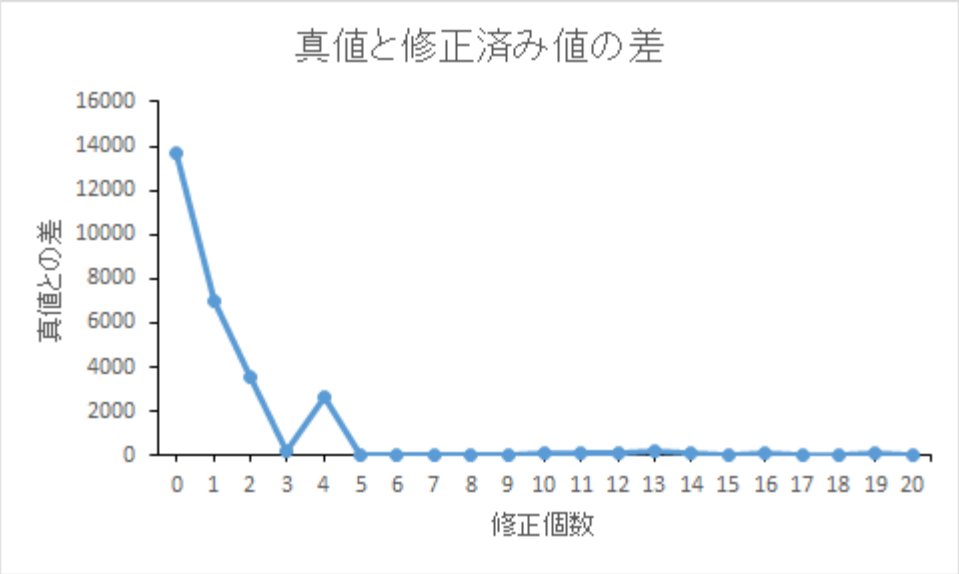


図 11：選択的エディティングの効率性

4.4 エディティング手法の比較

図 12 は、以上の結果を 1 つの図にまとめたものである。ランダムな人手によるエディティングよりもマクロエディティングの方が、さらに、マクロエディティングよりも選択的エディティングの方が優れていることが示されている。

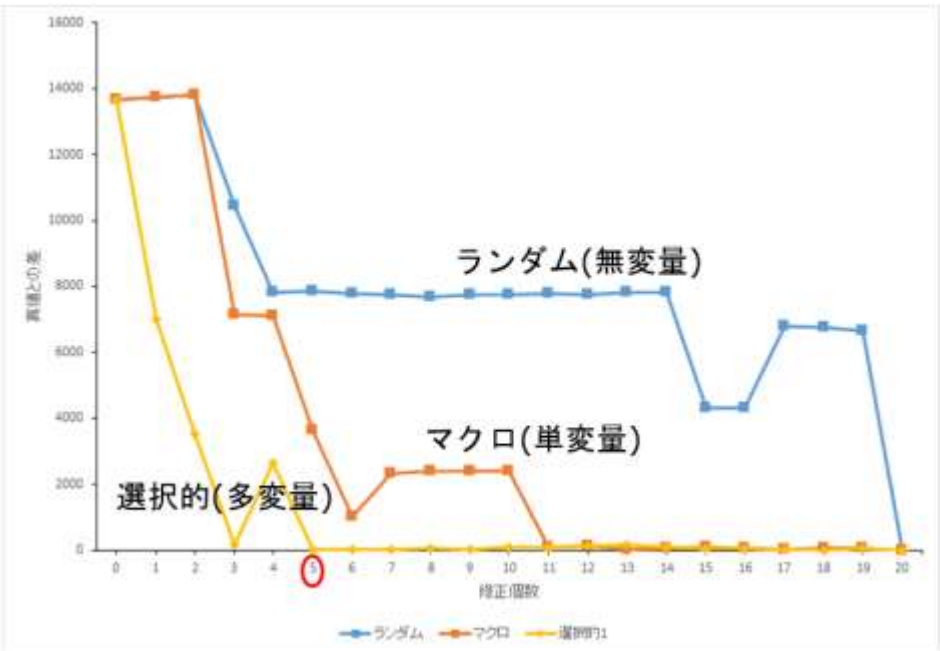


図 12：3 手法の効率性

図 13 は、4 つの異なるモデルによってパラメータを推定し、選択的エディティングを行った結果である。パラメータをどのように推定するかによって、選択的エディティングの間にも多少の差が生じることがある。どのモデルが適切であるかは、調査データの特徴に応じて決定するべき事項である。

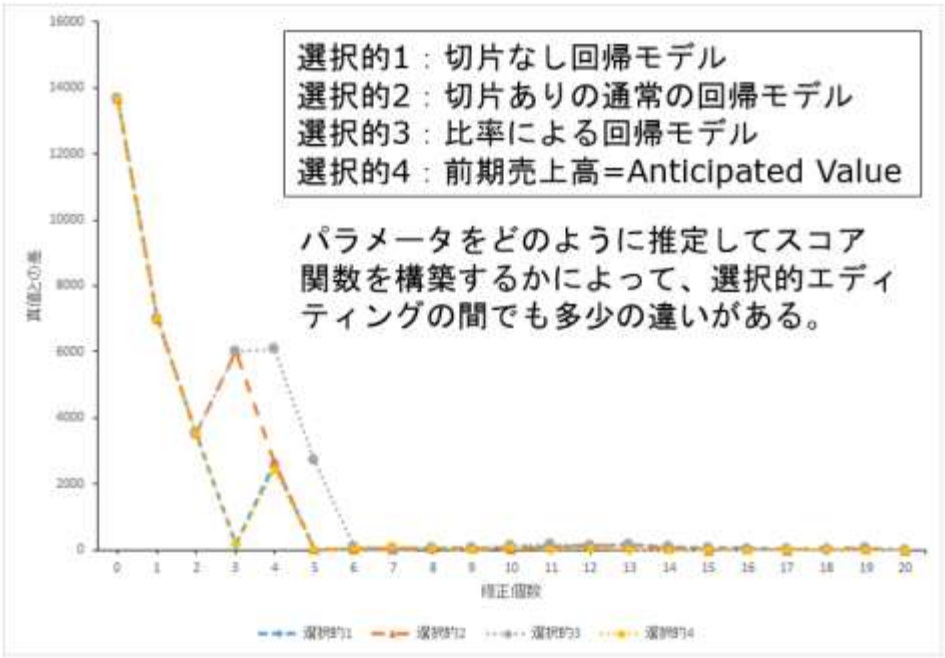


図 13：異なる選択的エディティングモデルの比較

5 実際の運用の際に気をつけるべきこと

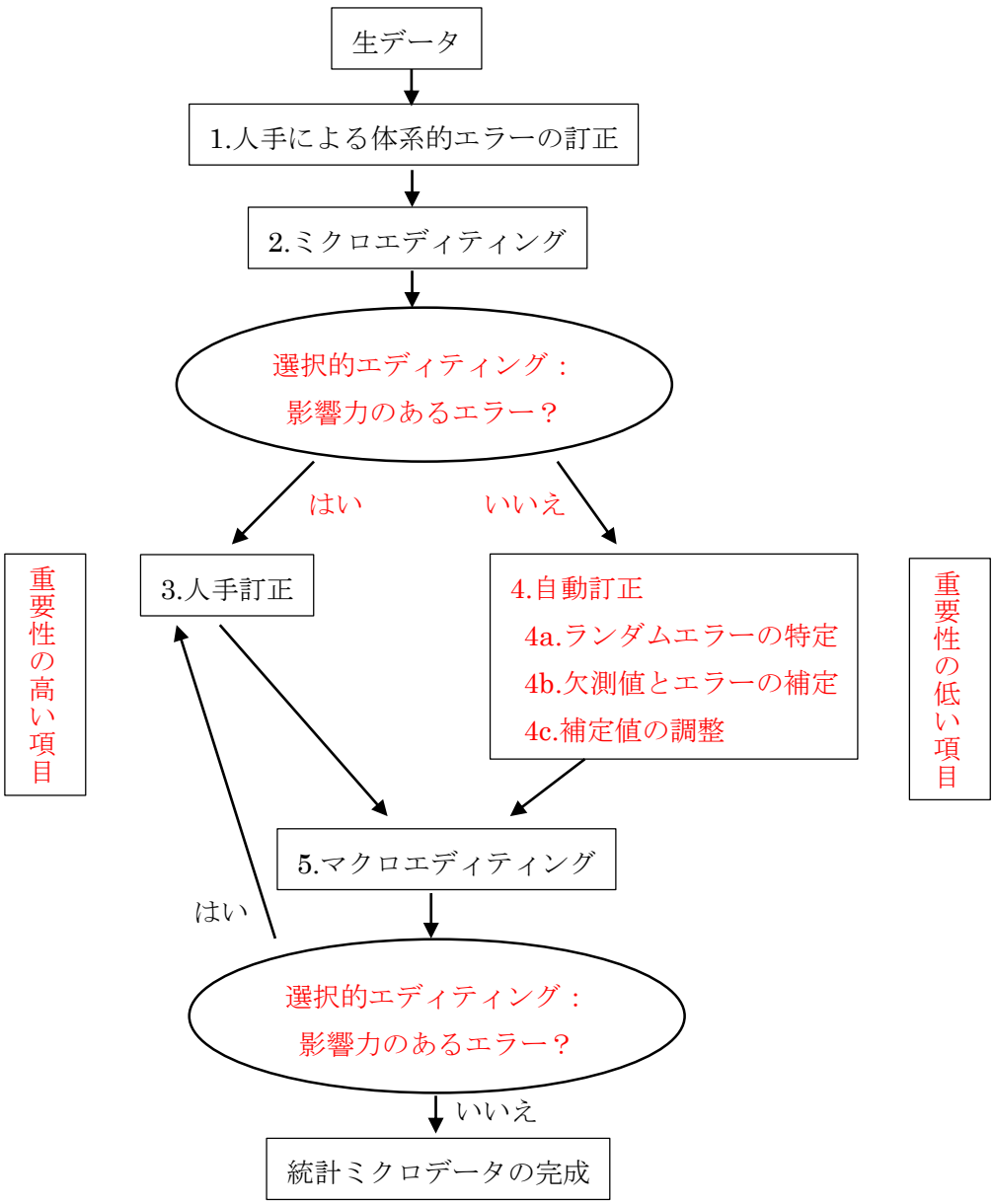
選択的エディティングを実行するには、モデルのパラメータとスコアの閾値の 2 つの情報が必要である。これらの情報を当該の今期データから算出する場合、調査プロセスがすべて終了するまでエディティングを実行することができなくなる。もし、調査プロセスのすべてが終了してからエディティングを開始するという手順を踏んだ場合、公表時期の大幅な遅れにつながる。よって、データを収集しながら、エディティングも平行して実行していく必要がある。

そのためには、過去のデータを活用して、モデルパラメータを事前に推定する。また、過去のデータを活用したシミュレーション研究によって閾値を事前に設定する。なお、閾値をいくつに設定するかによって、どれだけの個数のエラーを訂正するかが決まるため、この情報は極めて重要である。よって、恒常的に研究を実施し、設定した閾値が適切かどうかを追跡研究するべきである。また、実際にエディティングを行う際にも、事前に設定した閾値の妥当性に関して注視する必要がある。

このように、事前にモデルパラメータと閾値を設定することによって、データ全体に関してスコアを比較せずとも、データが入手されるたびに選択的エディティングを実行していくことができる。

6 選択的エディティングの手順の流れ

選択的エディティングの大まかな流れを図示する。ステップ 2 のミクロエディティングとステップ 5 のマクロエディティングにおいて、影響力のあるエラーを特定する作業のことを選択的エディティングと呼ぶ。これは、飽くまでも 1 つの例に過ぎない。他の類型については、Di Zio et al. (2015, pp.33-38)を参照されたい。



出典：de Waal et al. (2011, p.18)を修正して作成

7 SELEKT ソフトウェアについて

本稿で示した選択的エディティングを実装したソフトウェア SELEKT をスウェーデン統計局から入手した。SELEKT は、非売品の SAS マクロであり、スウェーデン統計局において、2004 年から 2014 年まで 11 種類の調査で実装され、10%～60%の費用削減を達成した実績がある。また、フィンランド統計局、英国国家統計局、カナダ統計局、ニュージーランド統計局など、諸外国にも貸し出して実装されており、国際的な評価も高い。使用方法については、本稿の別冊「SELEKT 1.3 のユーザーガイド」（統計センター内限）を参照されたい。

8 統計データエディティングに関するワークセッション

統計データエディティングに関するワークセッションは、国連欧州経済委員会(UNECE)と現地統計局の共催で開催され、2015 年のワークセッションは 1991 年の第 1 回から数えて 20 回目となる国際会議であり（1990 年代は毎年、2000 年代から 1 年半周期で開催)、UNECE とハンガリー中央統計局との共催で開催された。討議内容は、主に、データエディティングの革新的な手法や技術開発、統計の加工処理におけるデータエディティングの工程などについてであり、各国の統計機関が参集し、情報や意見の交換を行うものである。特に、選択的エディティングは、1990 年代初頭より、本会合にて提案・議論され、2000 年代に入って各国の統計機関において実務に導入されてきた実績がある。

このように、データエディティングに関して活発な議論の行われている本ワークセッションに参加し、最新の情報を収集すると同時に研究成果を発表し、各国統計局の研究員との意見交換を行い、交流を図った。とりわけ、業務プロセス改革に資すると思われる選択的エディティングに関する情報収集と意見交換を行った。なお、国連欧州経済委員会は、国連経済社会理事会の下部機関である地域経済委員会の一つとして 1947 年 3 月に設立され、事務局はジュネーブに所在している。我が国は加盟国ではないものの、国連加盟国としてオブザーバー参加が許されている。筆者は、2012 年 9 月、2014 年 4 月、2015 年 9 月のワークセッションに参加した。2012 年 9 月のワークセッションについては、高橋(2013)を参照されたい。本節では、2014 年 4 月と 2015 年 9 月のワークセッションについて報告する。

8.1 2014 年 4 月のワークセッション概要

第19回のワークセッションは、2014年4月28日から30日まで、フランスの首都パリで開催された。参加国は以下のとおり：アイルランド、イタリア、オーストリア、オランダ、カナダ、スイス、スウェーデン、スペイン、スロヴェニア、デンマーク、ドイツ、ノルウェイ、ハンガリー、フィンランド、フランス、メキシコ、モルドバ共和国、ロシア、英国、日本、米国。欧州委員会は欧州統計局(Eurostat)が代表した。また、ユーラシア経済委員会

(EURASEC)、国際労働機関 (ILO)、経済協力開発機構(OECD)の代表者も参加していた。出席者は約50人であり、ニュージーランドは論文を提出したものの、諸事情により欠席であった。討議事項は以下のとおりである。詳しい内容は、本稿の付録1を参照されたい。

(i) 選択的エディティング/マクロエディティング
(ii) 新たな手法
(iii)データエディティングの実施と関係者の協力
(iv) センサスデータ及び社会データのエディティング
(v) 国際協力及びソフトウェアとツール

8.2 2015 年 9 月のワークセッションの概要

第20回のワークセッションは、2015年9月14日から16日まで、ハンガリーの首都ブダペストで開催された。参加国は以下のとおり：イタリア、オーストリア、オランダ、カナダ、スイス、スウェーデン、スペイン、スロヴェニア、チリ、デンマーク、ドイツ、ニュージーランド、ノルウェイ、ハンガリー、フィンランド、ボスニア・ヘルツェゴヴィナ、ラトヴィア、リトアニア、ロシア、英国、日本、米国。また、ユーラシア経済委員会(EURASEC)及び世界保健機関(WHO)の代表者も参加していた。出席者は約60人であった。討議事項は以下のとおりである。詳しい内容は、本稿の付録2を参照されたい。

(i) 選択的エディティング及びマクロエディティング
(ii) エディティング及び補定に関する変更点の運用とサポート
(iii) ソフトウェアツールと国際協力
(iv) 評価とフィードバック
(v) 革新的手法及びデータ革命
(vi) 統計データエディティングの汎用的なプロセスの枠組みを構築する作業部会の報告

8.3 選択的エディティングに関する意見交換

質問 1：人口系調査（主に質的変数となるもの）への適用で具体的なよい例示があれば。
回答 1：質的変数の中でも、特に、「白人、黒人、その他」といった順序のない変数の場合、自然なメトリックがなく、選択的エディティングは適用できない。一方、1 週間の労働時間(0 時間～168 時間)という量的変数を 7 つのカテゴリーに分割した順序変数の場合、これまで検討をした事例がない。通常、エラーの影響度が分からなければ選択的エディティングを用いることは難しいため、カテゴリー1 とカテゴリー7 との間で、エラーの影響度が違うと言えるかどうかを考える必要がある。7 つのカテゴリーから元々の 1 時間単位といった自然なメトリックの回復が行えるならば、適用できるのではないか。（イタリア国家統計局：Marco Di Zio 博士、Ugo Guarnera

博士）

質問 2：理論で行う選択的エディティングの適用範囲（内容）以外で、（感覚的なところの）人手審査に引っかかるものはどの程度あるのか。また、その具体的な例など。

回答 2：選択的エディティングを実行する前に、体系的エラー(systematic error)の検出と訂正を人手によって行う。この割合は、調査ごとに異なるため、一概には言えない。（ノルウェイ統計局：Li-Chun Zhang 教授）

質問 3：一般的に、選択的エディティングを行ってから補定を行うが、統計センターでは、補定を行ってから選択的エディティングを行う流れを考えている。何か問題点はあるか？

回答 3a：欠測データによってスコアにどのぐらい影響力が出るかを検討するために、コールドデックの通常の変動の下限を用いて補定を行ってから選択的エディティングを行う方法もある。しかし、項目無回答のある変数からスコア関数への影響を取り除きたい場合もある。（スウェーデン統計局：Karin Lindgren 氏）

回答 3b：エディティングのプロセスをどのように構築するかは、各機関次第である。つまり、あるステップが最初にあり、別のステップが後に来るかに関して理由が説明できる必要がある。（イタリア国家統計局：Marco Di Zio 博士）

質問 4：もし体系的エラーを人手訂正せずに選択的エディティングを行った場合、選択的エディティングによって体系的エラーを適切に検出することは可能か？

回答 4a：どのような体系的エラーについて話をしているかによって答えが違ってくるので、汎用的な答えは存在しないだろう。もしエラーが測定単位エラー(unity measure error)なら、他の大多数のデータから明らかに分離した集団を形成しており、おそらく選択的エディティングによって測定単位エラーを処理することができるだろう。しかし、測定単位エラーがどのぐらいあるかに大きく依存する。もし体系的エラーが非常にわずかであり小さな値であるなら、そういった体系的エラーを処理する唯一の方法は、選択的エディティングを行う前に処理するしかないだろう。（イタリア国家統計局：Ugo Guarnera 博士）

回答 4b：選択的エディティングは、体系的エラーを検出するものではない。しかし、重要なことは、選択的エディティングによって時間の節約が達成でき、より重要な体系的エラーを人手訂正する時間が十分に確保できるのである。（スペイン国家統計局：Pedro Revilla 氏）

質問 5：選択的エディティングを適用させた際，集計途中で当初見込んだ内容からの調整や変更といったものがどの程度発生しているか。

回答 5：この件については、すでに刊行論文に記述があるので、下記の p.42 を参照されたい。「選択的エディティング関連の報告論文翻訳集：国連欧州経済委員会（UNECE）統計データエディティングに関するワークセッション」『製表技術関連資料集』no.11. 選択的エディティングを実装する際に必要となる閾値の算出方法に関する問題を扱っている。また、閾値の設定方法については、pp.88-98 も参照されたい。

8.4 次回ワークセッションについて

オランダ統計局のSander Scholtus氏の提案により、次回の統計データエディティングに関するワークセッションは、2017年春にオランダのハーグにて開催される予定となった。次回のワークセッションで討議される予定の事項は、以下のとおりである。掲載されている国名及び団体名は、2015年9月16日現在において、これらの議題に参加の意思を表明したものである（ただし、拘束力はない）。

1. 機械学習
   (i) ニュージーランド、フランス
2. 新たな手法
   (i) オランダ、イタリア
3. ソフトウェアツールの共有と CSPA
   (i) オランダ、ドイツ、オーストリア、英国、スロヴェニア、カナダ
   (ii) この議題では、ソフトウェアのデモや実装上の経験談の共有などが期待されている。
4. 新たなデータ情報源
   (i) カナダ、オランダ、米国、イタリア
   (ii) この議題では、ビッグデータや複数情報源の統合に関する報告が期待されている。
5. 国際的な協力体制と標準化
   (i) フィンランド、ドイツ、スロヴェニア、オランダ
   (ii) この議題では、VTL、GSDEMs、CSPA といった新たな標準的手法の実装などに関する報告が期待されている。
6. 2021 年センサス
   (i) ドイツ、ノルウェイ、イタリア、カナダ
7. 変化への対応
   (i) カナダ、デンマーク、ニュージーランド、米国、フィンランド

参考文献

[1] Arbués, I., Revilla, P, & Shaldaña, S.（高橋将宜訳）. (2015).「確率最適化問題としての選択的エディティング」, 『製表技術関連資料』no.11, pp.138-150.

[2] de Waal, T. (2013). “Selective Editing: A Quest for Efficiency and Data Quality,” *Journal of Official Statistics* 29 (4), pp.473-488.

[3] Di Zio, M. & Guarnera, U. (2013). “A Contamination Model for Selective Editing,” *Journal of Official Statistics* 29 (4), pp.539-555.

[4] Di Zio, M., Gros, E., Guarnera, U., Kolomiyets, T., Luzi, O., Oinonen, S., Ollila, P., Pannekoek, J., Pyy-Martikainen, M., Vale, S., & Zhang, L. (2015). “Generic Statistical Data Editing Models: GSDEMs (Version 0.5),” *Work Session on Statistical Data Editing, UNECE*, Budapest, Hungary, 14-16 September 2015.

[5] de Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken: A John Wiley & Sons, Inc.

[6] Granquist, L. (1991). “Macro-Editing- A Review of Some Methods for rationalizing the Editing of Survey Data,” *Statistical Journal of the United Nations Economic Commission for Europe* 8 (2), pp.137-154.

[7] Granquist, L. (1997). “The New View on Editing,” *International Statistical Review* 65 (3), pp.381-387.

[8] Latouche, M. & Berthelot, J.-M. (1992). “Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys,” *Journal of Official Statistics* 8 (3), pp.389-400.

[9] Norberg, A., Adolfsson, C., Arvidson, G., Gidlund, P., & Nordberg, L. (2010). *A General Methodology for Selective Data Editing*, version 1.0. Statistics Sweden.

[10] 高橋将宜. (2012).「諸外国のデータエディティング及び混淆正規分布モデルによる多変量外れ値検出法についての研究」, 『製表技術参考資料』no.17, pp.1-45.

[11] 高橋将宜, 伊藤孝之. (2013).「経済調査における売上高の欠測値補定方法について～多重代入法による精度の評価～」, 『統計研究彙報』第 70 号 no.2, pp.19-86.

[12] 高橋将宜. (2013).「諸外国における最新のデータエディティング事情～混淆正規分布モデルによる多変量外れ値検出法の検証～」, 『製表技術参考資料』no.23, pp.1-67.

[13] 高橋将宜, 伊藤孝之. (2014).「様々な多重代入法アルゴリズムの比較～大規模経済系データを用いた分析～」, 『統計研究彙報』第 71 号 no.3, pp.39-82.

[14] 高橋将宜, 阿部穂日, 野呂竜夫. (2015).「公的統計における欠測値補定の研究：多重代入法と単一代入法」, 『製表技術参考資料』no.30, pp.1-95.

付録 1：2014 年 UNECE ワークセッション報告論文概要

本付録では、2014年4月のUNECE統計データエディティングに関するワークセッションにて報告された全論文を日本語で簡潔に要約して紹介している。実際の全論文（英語）は、UNECEのウェブサイト[22]にて閲覧及びダウンロードが可能である。以下、WPはワーキングペーパー(Working Paper)の番号を表している。その後に英文タイトルを掲載し、括弧の中に著者名と国名を記し、その下に要旨を掲載している[23]。

WP.1 Provisional Agenda and Tentative Timetable (UNECE)

ワーキングペーパー1番は、報告論文ではなく、ワークセッションのタイムテーブルである。本ワークセッションは、フランスの首都パリにおいて、2014年4月28日（月）に開幕し、4月30日（水）に閉幕した。討議された事項は、以下の5つのトピックであった：(1) 選択的エディティング/マクロエディティング；(2) 新たな手法；(3) データエディティングの実施と関係者の協力；(4) センサスデータ及び社会データのエディティング；(5) 国際協力及びソフトウェアとツール。報告された論文の数は34（WP.2～WP.35）であった。

トピック（i）：選択的エディティング/マクロエディティング

WP.2 Score Functions under the Optimization Approach（Ignacio Arbués and Pedro Revilla, スペイン）

本報告は、選択的エディティングを定義する理論的なフレームワークの構築を目指している。通常の選択的エディティングと同様に、まず重要なデータと重要ではないデータとに分割する。重要なデータとは、影響力の高いエラーを含んでいる可能性が高く、人手により修正されるべきレコードから構成される。標本ユニットのどれを人手によるエディティングのために選択するべきかという決定に関する問題がある。この問題は、最適化問題として定式化することができる。この最適化の目的は、エディットするべき標本ユニットの数を最小化することであるが、選択されたユニットのみをエディティングすることと一定の範囲内のもののみをエディティングするという制約がある。そこで、2つのバージョンの汎用的な最適な問題として定式化している。一つ目は、もしユニットの選択に関して横断的な情報を用いることができないならば、確率的な最適化問題を導出する。二つ目は、もし横断的な情報を用いることができるならば、組み合わせによる最適化問題を導出する。一見すると、スコア関数と最適化は、非常に異なったアプローチのように思われるが、実際には、スコア関数は最適な問題に組み込まれたものと理解できる。

---

[22] http://www.unece.org/index.php?id=33757（2016 年 3 月 2 日アクセス）

[23] 論文の引用には、下記のフォーマットの使用を推奨する。著者名. (2014). “タイトル,” *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Paris, France, 28-30 April 2014.

WP.3 Maintenance of Selective Editing in ONS Business Surveys（Daniel Lewis, 英国）

英国国家統計局は、少数の変数で構成される月次調査から、多くの変数で構成される年次調査まで、多くの企業調査において選択的エディティングを使用している。選択的エディティングでは、あらかじめ設定した閾値を超えるスコアの企業をエディットする。ここで重要なことは、推定値の品質を維持するためには、閾値を定期的に検証する必要がある。これは、実際に行うには難しい問題である。というのも、選択的エディティングを用いるということは、データが検証されていない企業が存在するからである。選択的エディティングによって費用削減が達成でき、その削減された費用の一部を用いて、いくつかの企業を標本抽出して、実際に閾値が妥当であったかどうかを確認するべきである。メンテナンスでは、選択的エディティングの結果として発生するいかなる不測のデータ問題についても対処できるプロセス管理が必要であり、選択的エディティングを実装する際に生じる文化的な問題に対処することも重要である。

WP.4 Experiences from Selective Editing at Statistics Sweden（Anders Norberg, Karin Lindgren, and Can Tongur, スウェーデン）

スコア関数を用いた選択的エディティングを実装するために、スウェーデン統計局では、2007年から汎用的なITツールの開発に取り組んできた。外国貿易統計や賃金と給与構造調査といった調査において、汎用的エディティングツールSELEKTの試作版を成功裏に実装した。本稿では、スウェーデン統計局において、選択的エディティングを実装した際の経験談を紹介している。特に、SELEKTを用いる場合には、実装の初期段階において、以下のチェックリストを確認する必要があることが示されている。

・ミクロ的なエディティングに多くの費用がかけられており、費用削減の余地がある。
・主要な変数は、連続変数である。
・主要な出力結果は、ミクロデータの総計により構成されている。
・期待される値(anticipated value)が入手可能である。

WP.5 Use of Administrative Data for Selective Editing: the Case of Business Investments（Marco Di Zio, Paolo Forestieri, Ugo Guarnera, Massimiliano Iommi, and Antonio Regano, イタリア）

Rのプログラミング環境において開発した選択的エディティングツールであるSeleMixについての報告を行った。国民経済計算に関して、構造的企業調査のデータを用いて計算を行っている。近年では、イタリア国家統計局における構造的企業調査のプロセスにおいて、行政データも活用している。しかし、投資額に関しては、行政データが利用可能ではないため、構造的企業調査のみを通じて収集している。本稿では、投資変数に注目して、2010年の構造的企業調査に選択的エディティングを応用した結果を報告している。本報告で

は、SeleMixを構造的企業調査に応用し、評価を行った。妥当性検証の結果によると、モデルを構築する際に使用する補助情報（過去の調査票情報）と比較した場合に、もしエラーが特殊な傾向を示さない場合には、検出することが難しいことが分かった。

WP.6 Selective Editing Techniques and Seasonal Adjustment（Thomas Balcone, Antonia Bertin, Marie Cordier-Villoing, and Dominique Ladiray, フランス）

本稿は、短期経済統計の指標の月例生産に選択的エディティング手法を適用した結果について報告を行った。4年前から、フランスでは、短期統計の生産システムを大幅に変更している。本報告の目的は、この4年間で導入された新たな手法を紹介し共有することである。このシステムは、主に、GSBPMの枠組みで行われ、統計手法に関しては、EDIMBUSの手引書にしたがった。

WP.7 Using R-Indicators to Monitor Household Surveys and Prioritize Data Collection: An Application to the 2010 Household Wealth Survey in France（Thomas Merly-Alpa, フランス）

本稿では、無回答による偏りを可能な限り是正し、回答率の低下による精度の低下を最小限にするための世帯の選択手法について紹介を行った。調査環境の悪化に伴い、すべての主要なユニットに関して一葉に、回答率は低下する傾向にある。そこで、特定の世帯を抽出して重点的に無回答バイアスを減らし、精度を向上させる必要があるが、そのためには、優先化を行わなければならない。そのために、本稿では、標本の代表性を分析するためにR指標を用いた。本稿の目的は、2010年の世帯貯蓄調査データにこれらの手法を適用することである。はじめに、標本のR指標を計算し、優先化を行わない段階での調査がどのようになっているかを示すことができる。次に、地域ごとに調査員が不在となる場所をシミュレーションし、どのような優先化を行うかによって、フランスにおける貯蓄の分布の推定品質への影響を比較する。

WP.8 An Assessment of Automatic Editing via the Contamination Model and Multiple Imputation（Masayoshi Takahashi, 日本）

自動エディティングのプロセスは、通常、エラー特定ステップ（エラー検出）とエラー訂正ステップ（補定）の2つのプロセスから構成される。本稿の目的は、日本の経済調査のエディティングプロセスの一部を自動化する手法の提案である。そのために、経済センサス - 活動調査のデータを用いて検証を行った。まず、人工的なエラーを導入し、RパッケージSeleMixによる検出を行い、次に、MCMC、FCS、EMBといった3つの多重代入法アルゴリズムを用いて補定を行った。最後に、真値と比較して、これら2つのアルゴリズムのパフォーマンス比較を行った。

WP.9 Text Analysis Tools for Editing and Verification（Wendy L. Martinez, 米国）

公的統計におけるデータプロセスにおいて、テキストフィールドといった非構造データは、十分に活用されていない。本稿は、こういったフィールドから情報を抽出し利用するテキスト分析の手法を紹介している。自動車事故報告書のラベル付けや既存のコーディングの検証、データのエディットなどをどのようにして行えるか、自動車事故の報告書データを用いて例証している。本稿では、データエディティング手法として、テキスト分析を用いる目的について議論している。例えば、テキスト分析を用いることで、以下のようなことが可能となる。報告書に記述されている話やイベントの種類といった情報を用いることで、分類を行い、データに応用し、誤分類されたレコードがないかを調べたり、コーディングミスのあったレコードを再分類したりできる。

WP.10 Adjusting for Remaining Measurement Error after Selective Editing（Thomas Laitila and Anders Norberg, スウェーデン）

選択的エディティングは、いわゆる無作為抽出理論に基づいていないため、エディット済みデータから得られた結果を伝統的な統計手法によって一般化し、エディットされていない観測値を含む母集団に対する結果に当てはめることができない。すなわち、データ内に残存している測定誤差の影響により、選択的にエディットされたデータセットに基づく推定量には偏りが存在する。先行研究では、測定誤差に関して、ユニットを無作為抽出した選択的エディティング手法が提案されている。本稿は、選択的にエディットされたデータセットに残存するエラーによる推定量の偏りを修正する方法を提案する。エディット済みのデータにおいて観測された測定誤差が観測されたスコアと関連がある場合には、モデルベースの手法を用いる。推定モデルを用いて、エディットされなかったケースに残存する測定誤差を予測し、母集団レベルまで要約することで、測定誤差のレベルを推定することができる。なお、報告者は欠席であり、座長が要旨のみを紹介した。

トピック（ii）：新たな手法

WP.11 Implementation and Evaluation of Automatic Editing（Jeroen Pannekoek, Mark van der Loo, and Bart van den Broek, オランダ）

企業統計において、自動エディティングは、統計作成プロセスの重要な位置を占めるものである。通常、目的に応じてサブタスクやエディティング関数の定義など、多くの設定を行わなければならない。再利用可能な形で標準化した手法とツールを用いることができれば、自動エディティングシステムの費用対効果、設計、実装、メンテナンスを大幅に改善できる。自動エディティングに関して、オランダ統計局では、標準的な手法を文書化し、Rベースのツールに実装することで、汎用的な標準データエディティング関数を開発している。これを実現するために、データエディティングプロセスをプラグ・アンド・プレイ

で接続できる再利用可能な標準的プロセスステップに分解することで、モジュラー手法を開発している。本稿では、このモジュラーシステムの実装について報告し、各々のプロセスステップの影響を測る指標についても議論している。異なるプロセスステップに応じて、プロセスの進展をモニターするためには、グラフ表示を利用することができる。このようにモニタリングすることで、データ品質やデータエディティングシステムの手法とパラメータに関して、継続的にフィードバックを得ることができ、結果として、標準的な統計作成プロセスを最適化し統合することが可能となる。

WP.12 Presentation and Development of Outlier Treament in HCSO（Gergely Horváth, ハンガリー）

ハンガリーは、外れ値の処置方法についての経験談を報告した。様々な外れ値の対処法を紹介し、現在、ハンガリー中央統計局において実施している方法について紹介を行った。観測ユニットの値を精査するという意味では、ハンガリー中央統計局における現行の外れ値対処法は、通常の統計的エディティング手法と似ている。一方、外れ値を検出した直後に訂正を行うわけではないため、典型的なエラー検出法ではない。少なくとも、方法論者によって値の訂正は行われない。外れ値の変更による影響はユニットの重みを減らすこととなるので、外れ値の変更は推定段階で行う。つまり、外れ値処理の目的は、推定の改善である。現在、最もよい手法は、ウェイトの現象とウィンザー化[24]である。

WP.13 Simulating Multiple Imputation of Water Consumption in the German Agricultural Census 2010（Lydia Spies, Sven Schmiedel, and Katrin Schmidt, ドイツ）

ドイツ連邦統計局では、推定値を公表するかいなかを変動係数の大きさによって決めている。よって、欠測値を補定することによって変動係数にどれだけの影響が出るかを確認できる多重代入法の研究を行っている。実際には、真値が分からないため、異なる手法間でどの手法がよいのか分からないという問題がある。よって、統計環境Rを用いて、繰り返しシミュレーションを実行できるシステムを開発した。このシステムを用いて、マルコフ連鎖モンテカルロ法（MCMC）に基づく多重代入法において、2つのモデル（ホットデックと予測平均マッチング）を比較した研究の報告を行った。2010年の農業センサスの水道消費データを人工的に欠測させた上で、2つのモデルの比較検証を行い、今後の改良案の提示を行った。

WP.14 Data Editing and Scanner Data（Isabelle Léonard, Gaëtan Varlet and Patrick Sillard, フランス）

フランスは、スキャナー・データ・プロジェクトについて報告を行った。研究段階のものであるが、予備的な実験結果について報告があった。2009年以来、フランスでは、消費者

[24] Winsorization: データ内の最大値と最小値の影響を抑える補定手法

物価指数のデータ収集を変更するスキャナー・データ・プロジェクトを実施している。このプロジェクトは、調査員によって収集されていた収集プロセスの一部を、小売業者自身によって記録されたデータに置き換えることを目的としている。このプロジェクトでカバーする範囲は、工業用食品、衛生関連製品、クリーニング関連製品などである。4つの大企業が、データ提供に同意している。大容量のデータを利用することで、新たな統計データプロセスの構築ができ、現行のプロセスの改善を行うこともできる。このプロジェクトの目標の1つは、個別指標の精度向上である。同時に、個別に収集された情報を自動プロセスによって置き換えなければならない。

WP.15 Multiple Imputation Methods for Imputing Earnings in the Survey of Income and Program Participation（María García, Chandra Erdman, and Ben Klemens, 米国）

Survey of Income and Program Participation (SIPP)は、パネルによってデータを収集する縦断調査である。パネルは、2から4年の頻度で面接調査する14,000から65,000世帯から構成されている。2006年に、米国センサス局では、費用を削減し、データ品質を改善することを目的として、SIPPの大幅な再設計を開始した。この再設計では、データの収集方法だけではなく、データ処理についても改善策を模索している。現行のSIPPでは、欠測データの補定にホットデック手法を用いている。本稿では、月次収入データの補定を2つの手法で行った。1つはモデルベースの順次回帰多重代入法(SRMI: Sequential Regression Multiple Imputation)であり、今1つは確率的ホットデックである。シミュレーションに基づき、これら2つの手法を比較した。SRMIは、従来のホットデック手法の代替案として使用でき、推定値を改善できる可能性のあることが分かった。なお、SRMIは、FCSと本質的には同じ手法である。

WP.16 Imputation with Multi-Source Data: the Case of Italian Structural Business Statistics（Marco Di Zio, Ugo Guarnera, and Roberta Varriale, イタリア）

近年、イタリアでは、構造的企業統計を作成する際に、主要な情報源として、行政データを活用している。しかし、行政データを用いてミクロデータを構築するには、異なる集計レベルの推計値に一貫性を持たせなければならず、困難が伴う。すべての変数がすべてのデータ情報源において利用可能ではなく、また情報源は対象となる母集団の一部分のみをカバーしているので、ミクロデータファイルは、補定プロセスを経る必要がある。補定手法は、主要変数間の統計的な関係、バランスエディット、ゼロ過剰な変数の存在といった制約条件のもとで整合性を持つように、異なる手法の組み合わせから構成されている。そのような複雑さがあるので、手法を評価することは、簡単なことではない。本稿では、中小企業の標本調査に基づいた公表推定値との比較を実施し、差異を標本誤差と測定誤差に分解した。このように、異なるエラーのソースごとに影響度を分析することは、結果の妥当性検証として有用であり、文脈を考慮した統計作成プロセスの改善に寄与するこ

とができる。

WP.17 A Generalised Fellegi-Holt Paradigm for Automatic Editing（Sander Scholtus，オランダ）

オランダは、自動エディティングのパラダイムを汎用化し、複雑なエディティング操作を一つに統合する方法を提案した。現在の公的統計において使用されているほとんどの自動エディティング手法は、Fellegi and Holt (1976)のパラダイムに基づいている。このパラダイムによると、エラー特定の問題は、各々のレコードに対して、補定する変数の最小のサブセットを見つけ出すことによって解決される。その結果、レコードは、エディット規則と一貫性を持つことになる。しかし、Fellegi-Holtの自動エディティングでは、個別の値を変更することが重要な要素となる。一方、人手訂正では、複数の値を同時に変更するなど、複雑なエディティング操作が行われるのが一般的である。本稿では、これまでとは異なるアプローチを提案している。エラー特定問題に関して新たな定義をし、エラーによって1つ以上の変数が同時に影響を受ける可能性を扱えるようにした。この新たな手法によるエラー特定は、伝統的なFellegi-Holtパラダイムをスペシャルケースとして内包する汎用的なものである。新たなパラダイムのもとでは、いわゆるエディット・オペレーションの数を最小化することによって、エラーは特定される。エディット・オペレーションとは、一度に1つの変数に対して新たな値を1つ補定するといったものが例として挙げられる。しかし、より汎用的なエディット・オペレーションによって、複数の変数における変更を同時に扱うこともできる。このような汎用化により、自動エディティングがより適切なものとなり、データエディティングプロセスの効率性が改善する。

WP.18 Assessing the Impact of a New Imputation Methodology for the Agricultural Resource Management Survey（Wendy Barboza, Darcy Miller, and Nathan Cruze，米国）

米国農業統計局は、農業資源管理調査において使用した2つの補定手法について紹介した。2年分の調査データを用い、2つの手法による推定値の比較を行った。この調査は、3段階で実施され、複数の情報収集源に基づいている。毎年の農業経営に関する状況を明らかにするものであり、農業経済及び地域経済に関する政策を決定するための唯一の客観的な情報を提供している。

トピック（iii）： データエディティングの実施と関係者の協力

WP.19 Obtaining Wide Support for Statistics Canada's Integrated Business Statistics Program: a Key Task in the Project Plan（Etienne Saint-Pierre and François Couture，カナダ）

カナダ統計局では、Integrated Business Statistics Program (IBSP)という企業調査を実

施するための新たなモデルを実装している。IBSPは、共通の手法を用いて統合されたデータエディティング手法など、共通の処理フレームワークを提供するシステムである。2014年から2017年までに、異なる10のプログラムに属する120以上の経済調査がこの新たな統合的・調和的フレームワークに移行する。このプロジェクトの開始当初から、効率性と品質を最大化するためには、データエディティングの工程に関わる数百人単位の職員からサポートを得ることが重要であると認識されている。本稿では、「自分の担当している調査と回答者は非常に独特なので、いかなる標準モデルにも適合しない」というマインドセットから、どのようにして、「新たな調和的モデルに自分の調査が移行するのが楽しみだ」というマインドセットに移行したか説明している。6つの要素が挙げられている：(1)強力なガバナンスと明確な支持の重要性；(2)新たな手法の開発に調査担当責任者を招くこと；(3)標準化されていない調査の独特な性格という概念の誤りを指摘すること；(4)概念の実証；(5)コミュニケーション・ストラテジー；(6)デザインの選択。

WP.20 Applying Process Indicators to Monitor the Editing Process（Karin Lindgren and Martin Odencrants, スウェーデン）

スウェーデンは、エディティングを監視するためのプロセス指標(process indicator)について報告を行った。実用可能な処理指標を開発し、実験的に調査に応用した。これらの指標を用いて、主にエディティングプロセスの評価を行ったり、データ収集プロセスについての評価を行うことができる。また、測定誤差に関して応用もでき、調査の品質管理にも使用できる。本研究の目標は、SELEKTツールを用いることで、様々な調査に使用できるように一連の指標を定期的に生成することである。

WP.21 Renewal of Editing Practices at Statistics Finland（Marjo Pyy-Martikainen, フィンランド）

フィンランド統計局における統計調査のエディティングについて、調査研究したところ、慣習は多様であり、時としてやや時代遅れなものもあった。さらに、エディティングの慣習について評価を行い、統計品質に与える影響を評価している統計調査は、ほとんどなかった。よって、エディティングの慣習をリニューアルする必要性があった。このリニューアルは、費用や時間の削減につながり、統計作成プロセスの標準化につながると認識されていたので、経営陣からの大きなサポートがあった。エディティングのプロセスモデルを構築し、4つの統計調査を用いて検証を行った。これらの調査に協力した職員は、統計エディティングプロセスを近代化し体系化する必要性を認識するに至り、このリニューアルに賛同した。しかし、エディティング担当職員のすべてが、必ずしもこの新たな試みを歓迎してはいなかった。彼らによると、新たな手法は、理解するのが困難とのことであった。新たな手法とツールに関して、研修を行うことによって、すべての関係者からサポートを得られるのではないかと期待されている。選択的エディティングの実装は、フィンラ

ンド統計局の戦略的目標として設定されており、経営陣もこの目標の達成を強く支持している。この目標を達成するための必要条件は、慣習がリニューアルされようとしている統計調査において、十分なコミットメントと十分なリソースが費やされることである。

WP.22 Implementing a New Editing System – Getting Everyone on Board（Remy Bråthen, Aslaug Hurlen Foss, and Geir Hjemås, ノルウェイ）

ノルウェイは、新たなエディティングシステムの試験的な実装に関して報告を行った。エディティングの枠組みをより標準化し、手順をより効率化し、品質を改善することは有益であることを示し、経営陣に対して新たなエディティングシステムを提示した。現行のシステムから新たなシステムに変更することに関する抵抗感は、文化的なものであって、システムに基づくものではなかった。エディティング手法に変更を施すには、実際にエディティングを担当している職員をプロセスに関与させる必要性が認識される。また、一つの手法だけではなく、補助的な手法を複数用意することも必要である。

WP.23 Questions Raised by the Implementation of the Data-Editing Device for French Structural Business Statistics（Philippe Brion and Johara Khélif, フランス）

2009年以来、フランスでは、ESANEという新たな統計作成プロセスに従って、年次の構造企業統計を公表している。このシステムは、予算削減といった制約の下で品質を低下させないための理論的な解答であった。統計作成プロセス全体を再設計したことで、いろいろな改善につながった。生産拠点は中央化され、選択的エディティングの実装も行われた。こういったベースに従って、フランスでは、2009年に新たな年次統計の実装を開始した。しかし、理論とは裏腹に、エディティング担当職員からの当初のフィードバックは、実務上の困難さを示していた。過去3年間において、エディティングを含む全プロセスの改善を行ってきた。本稿では、今後、どのような改善を実施できるかを議論している：選択的エディティング技術をグローバル・スタンダード化すること；過剰な自動化は、かえって非生産的であること；ユーザーからのフィードバックを利用し、アウトプットエディティングを多用することは有意義であること；新たなプロセスは、すでに、選択的エディティングのもたらす恩恵のよい例であること。

WP.24 Gaining Traction: Management Attitudes Toward Changes in Data EditingPractices（Allyson Seyb, ニュージーランド）

ニュージーランドの代表は、今回、諸事情により欠席となったが、議長であるClaude Poirier氏（カナダ）が代弁した。ニュージーランドの用意した原稿では、主要な企業統計プログラムの一環として、より効率的なデータエディティングプロセスの実施に関する経験談を報告した。エディティングに対する考え方は、経営組織、戦略的プラン、統計作成の手法と技術といった外的要因と同様に、より良いデータエディティングプロセスを成功裏

に採用する国家統計局の能力に影響を与える。従業員の考え方は、パフォーマンス品質に大きな影響を与えるという研究結果もある。2011年に、ニュージーランド統計局では、Statistics 2020というプログラムに着手し始めた。このプログラムは、ニュージーランドにおける公的統計の作成方法を変更するものである。このプログラムの主要な内容は、職員の能力開発、変化に対応できるように職員をサポートすること、強いリーダーシップを持つ職員の育成、パフォーマンス文化の構築である。データ処理プロセスを自動化することによる生産性の向上という組織の目標は、職員、文化、ヴィジョンが三位一体となって初めて現実のものとなる。本稿では、データエディティングに関する組織文化に注目しながら、経験談から得られた教訓について論じている。

トピック（iv）： センサスデータ及び社会データのエディティング

WP.25 Estimation of the Variance due to Imputation for the 2011 UK Census（Damião N. da Silv and Li-Chun Zhang, ブラジル&英国）

英国サウサンプトン大学とブラジルの共同研究では、2011年の英国センサスにおける補定にまつわる分散の推定方法について報告を行った。2011年の英国センサスにおける補定は、CANCEIS（CANadian Census Edit and Imputation System：キャンサイス）においてモジュラー化されており、ドナーベースの手法が実装されている。補定値を提供するドナーは、観測値の集合から一定の確率で抽出される。このように、補定において確率的な性質が備わっており、センサスの推定値に対して補定にまつわる推定不確実性を表す分散が追加されている。よって、この不確実性を測定することは、センサスのユーザーにとって重要な情報となる。本研究で用いた手法は、CANCEISの出力ファイルから得られた情報から補定分散を評価するというシンプルな手法である。つまり、この手法は、CANCEISをプラットフォームとして利用している他の調査にも応用可能である。この手法は、実際に2011年の英国センサスにおいて実装された。

WP.26 Editing the 2011 Census Data with CANCEIS and Options Considered for2016（Lyne Guertin, Marcel Bureau, and Josée Morel, カナダ）

カナダは、CANCEISに施された最新の改善点について報告を行った。2011年センサスにおけるエディティングと補定の処理についての報告し、2016年センサスに向けてどのような改善を行うか、紹介されている。CANCEISは、導出エンジンとドナーエンジンから構成されている。導出エンジンは、確定的補定を行って新たな変数を作成し、ドナーエンジンはドナー補定を実行する。2011年には、CANCEISは、CPUの効率性を高めるために、ネット環境下におけるC#言語で構築しなおされた。インプットデータ辞書ファイルをExcelフォーマットで入力し、アウトプットファイルをHTMLフォーマットで生成できるように、ユーザーフレンドリーに改善された。2011年のセンサスにおいて、CANCEIS

のパフォーマンスは非常に良かった。以下の要件を求める者にとって、CANCEISは、有益なエディティングと補定のツールとなるだろう：確定的補定とドナー補定を実行し、新たな変数を生成するシステム；多数のカテゴリカルな変数、数量変数、英数字の変数を同時に処理する能力；非常に多くのエディット規則を簡単に定義する能力；巨大データファイルを早く効率的に処理する能力；簡潔なユーザーによって定義されたパラメータを通じて、あらゆる側面の処理をコントロールできる柔軟性；カスタムなど複雑なインストール手続きなしで、通常のコンピューティングプラットフォームにおいて即座に使用できるソフトウェア。

WP.27 Automatic Data Editing Experience in 2010 Mexican Census（Isaac Salcedo, メキシコ）

メキシコは、センサスにおける自動エディティングの導入について報告を行った。地方自治体レベルで、エディティングがどのように行われたか紹介された。エディティングを自動化するために理論的ベクトル手法を用いたことにより、センサスの結果をタイムリーに公表することに成功した。

WP.28 Exploring Administrative Records Use for Race and Hispanic Origin Item Non-Response（Sonya Rastogi, Leticia Fernandez, James Noon, Ellen Zapata, and Renuka Bhaskar, 米国）

人種に関するデータは、市民権法などの法律を評価したり、施行したりする際に重要な役割を果たすものである。しかし、センサスにおいて、これらの項目の回答が得られないことがある。こういった項目無回答に関して、伝統的にセンサス局では、最近隣法によるホットデック手法を使用してきた。しかしながら、近年、米国における多様性が増し、近隣コミュニティのあり方が変わってきており、行政データが利用できる場合には、行政データを用いる方がより正確な情報を得られるようになってきた。また、レコード・リンケージ技術によって、2000年センサスの結果と2010年センサスの結果を関連付けて、欠測値に対処することもできるようになってきた。このようにすることで、2010年センサスでは、ホットデックによって補定した人種データを50%減らすことに成功した。本稿では、行政データを活用して、さらなる効率性の追求を行っている。

トピック(v)：国際協力及びソフトウェアとツール

WP.29 Migration of a Large Survey onto a Micro-Economic Platform（Val Cox, ニュージーランド）

ニュージーランドの代表は欠席だったが、議長のClaude Poirier氏（カナダ）が代弁した。2008年より、ニュージーランド統計局では、ミクロ経済プラットフォームというプロジ

ェクトを通じて、経済調査の処理方法を変革している。ミクロ経済プラットフォームは、ユーザーが経済データを読み込み、分析し、公表するための核となるプラットフォームである。ミクロ経済プラットフォームは、処理工程を完全にコントロールできるような柔軟なITツールを提供することで、統計分析者の生産性を最大化できるように設計されている。IT専門家に頼らなくとも、このツールを用いることで、ユーザーは、自分独自の統計出力を設計し、作成することができる。原則として、行政データを可能な限り利用し、調査データは補完的に使用する。このプラットフォームの目的は、全企業に関する核となる情報を縦断的なデータベースとして構築し、経済統計のニーズ変化にすばやく対応できるようにすることである。

WP.30 Towards Generic Analyses of Data Validation Functions（Mark van der Loo and Jeroen Pannekoek, オランダ）

本稿では、データ妥当性検証の分析を汎用化する手法に関して議論している。実装に向けて、データ妥当性検証と様々な国際的標準化についての関連性を指摘している。さらに、3 つの汎用的なパラメータを導き出すために、データ妥当性検証の汎用的モデルを与える。その 3 つとは、データが一定の品質要求を満たしているか否かに関するブール値、妥当性検証の下でのデータに与える影響を測る影響度関数、品質指標値と理想値との差を測る深刻度関数である。後者の 2 つは、妥当性シンタックス(validation syntax: VALS)言語における「相違」と「深刻度」の実現値として解釈できる。主な違いは、妥当性シンタックス言語において、それらの値は自由に指定できるが、オランダのモデルの場合は、妥当性規則の定義から論理的に導出される。これらの測定方法は、手元の特定の規則に関係なく定義されるので、汎用的に実装可能かどうかという疑問がある。そこで、非常によく使用される妥当性規則を調査士、影響度関数と深刻度関数を決定し、線形制約と質的制約の両方の条件の下、これらの測定方法によって汎用的かつ既知のアルゴリズム的処置が可能となることが分かった。つまり、汎用的実装は可能である。

WP.31 New Features of VIM – Visualization and Imputation of Missing Values（Alexander Kowarik, Matthias Templ, and Daniel Schopfhauser, オーストリア）

RパッケージVIMは、3つの目的を持って開発された。1つ目は、データ内の欠測値構造をグラフ手法によって可視化することである。2つ目は、ビルトインの補定手法によって欠測値を補定することである。3つ目は、視覚的なツールによって補定プロセスを検証することである。R初心者のユーザーのために、グラフィカルなユーザー・インターフェースを新たにRパッケージVIMGUIとして導入した。すべての作業は、ポイント&クリックによって簡単に操作できるように工夫されている。本稿では、VIMで利用可能な手法の応用について説明をし、VIMGUIのグラフィカルなユーザー・インターフェースの使用方法についてデモンストレーションを行っている。

WP.32 On Implementing CSPA Specifications for Editing and Imputation Services（Monica Scannapieco, Donato Summa, and Diego Zardetto, イタリア）

CSPA (Common Statistical Production Architecture：シースパ)は、公的統計の作成プロセスを近代化するためのテンプレートである。CSPAには、入力と出力に力点を置いた標準的な方法でインターフェースを定義するスペックが備えられている。本稿では、国際協調によって実証され、オランダ統計局によって開発されたRパッケージeditrulesの機能を活用し、エラー特定におけるCSPAのスペックの実装についての経験談を報告している。この手法は、イタリア国家統計局のCOREプラットフォームにおいて実装されている。エディティング及び補定プロセスを実行するために、どのようにCSPAがエラー訂正と統合されていくのかも示している。

WP.33 Editing and Imputation in the Memobust Handbook on Methodology of Modern Business Statistics（Sander Scholtus and Leon Willenborg, オランダ）

欧州統計システムでは、2011年1月から2014年3月まで、Memobust (Methodology of Modern Business Statistics)プロジェクトを実行してきた。このプロジェクトの主な目的は、最良の慣習を見つけ出し、共通の手法と欧州統計システムのガイドラインを構築することである。このガイドラインによって、回答者負担を軽減し、効率性とプロセス統合を促進して、企業統計の作成をサポートする。この目的を実行するために、企業統計の手法に関するハンドブックを改訂する。このプロジェクトは、欧州統計局の指揮の下、欧州の8つの国家統計機関によって実行されてきた。本稿では、Memobustハンドブックのカバーしているトピック、対象とするグループ、執筆・査読体制について報告している。

WP.34 SAS Enterprise Guide Project for Editing and Imputation（Saara Oinonen, フィンランド）

過去5年に渡り、フィンランド統計局では、エディティング及び補定に関する慣習について特に調査を行ってきた。最初のプロジェクトは2009年1月7日から2011年12月31日まで実施され、エディティング及び補定に関するプロセスモデルを構築した。また、このプロジェクトでは、国際的な動向についても幅広く調査を実施した。その結果、フィンランド統計局では、選択的エディティングこそが、エディティングプロセスの中核となるべきものであるとの結論に至った。今後、選択的エディティングに適している調査すべてにおいて、段階的に、選択的エディティングを実装していく予定としている。2012年には、新たなプロジェクトを立ち上げ、エディティングモデルと手法について、4つの統計調査において試験調査を行っている。ソフトウェアとしては、スウェーデン統計局のSELEKTとカナダ統計局のBANFFを候補としている。いずれも、SASの環境で実行できるものである。この2つのプログラムを用いることで、パラメトリックな汎用的エディティング及び

補定プロセスを実行するための基礎を構築することができる。重要な点は、新たなプログラミングがほとんど必要とされない点である。作成されるデータの構造も、統一的なプロセスに適している。

WP.35 Metadata Driven Application for Data Processing – From Local Toward Global Solution（Rudi Seljak, スロヴェニア）

予算削減の要求が常にある一方で、公的統計には高い品質の結果が求められる。こういった相反する要求の間にある公的統計家は、少ない予算の中で高い品質の統計を作成するという難題にますます直面していくこととなる。スロヴェニア統計局では、6年前から、データ処理を近代化するシステム開発の試みを行ってきた。プロトタイプは、データ妥当性検証、データ訂正、補定、集計、標準誤差の推定、製表といった統計作成プロセスのパーツごとにモジュール化されている。このシステムを徐々に改善して域、2010年農業センサスや2011年人口センサスといった大規模な調査において成功裏に実装してきた。汎用的に開発してきたものの、まだローカルな解決策によって成り立っている部分も多くある。2011年に新たなプロジェクトに着手し、既存の解決策をアップグレードして、1つのグローバルな解決策の構築を目指している。本稿では、新たに開発した汎用ツールの特徴について紹介し、このツールの導入によって、統計作成プロセス全体の設計がどのように変わっていくかを示している。

付録 2：2015 年 UNECE ワークセッション報告論文概要

本付録では、2015年9月のUNECE統計データエディティングに関するワークセッションにて報告された全論文を日本語で簡潔に要約して紹介している。実際の全論文（英語）は、UNECEのウェブサイト[25]にて閲覧及びダウンロードが可能である。以下、WPはワーキングペーパー(Working Paper)の番号を表している。その後に英文タイトルを掲載し、括弧の中に著者名と国名を記し、その下に要旨を掲載している[26]。

WP.1 Provisional Agenda and Tentative Timetable (UNECE)

ワーキングペーパー1番は、報告論文ではなく、ワークセッションのタイムテーブルである。本ワークセッションは、ハンガリーの首都ブダペストにおいて、2015年9月14日（月）に開幕し、9月16日（水）に閉幕した。討議された事項は、以下の6つのトピックであった：(1) 選択的エディティング及びマクロエディティング；(2) エディティング及び補定に関する変更点の運用とサポート；(3) ソフトウェアツールと国際協力；(4) 評価とフィードバック；(5) 革新的手法及びデータ革命；(6) 統計データエディティングの汎用的なプロセスの枠組みを構築する作業部会の報告。報告された論文の数は34（WP.2～WP.35）であった。

トピック（i）：選択的エディティング及びマクロエディティング

WP.17 Selective Editing of Business Investments by Using Administrative Data as Auxiliary Information（Marco Di Zio, Ugo Guarnera, Massimiliano Iommi, and Antonio Regano, イタリア）

企業の投資を推定する国民経済計算で利用可能なミクロデータは、構造的企業統計から流用されている。国民経済計算の推定段階では、他の情報源からのデータも利用しており、データのさらなる検証が可能となっている。行政データでは、2つの変数が投資と高い相関係数を示している。1つは、付加価値税における償却財に関する支出額で、合名会社と株式会社の双方に関して利用可能な情報である。もう1つは、期末における資産と期首における資産との差額から算出した変数で、これは株式会社のみに関して利用可能な情報である。本報告では、このように算出した行政データの変数を補助変数として用いることで、構造的企業統計の投資データに対して選択的エディティング手法を適用している。

[25] http://www.unece.org/index.php?id=37497（2016 年 3 月 2 日アクセス）

[26] 論文の引用には、下記のフォーマットの使用を推奨する。著者名. (2015). “タイトル,” *Work Session on Statistical Data Editing, United Nations Economic Commission for Europe*, Budapest, Hungary, 14-16 September 2015.

WP.18 Output Editing Based on Winsorization in the French SBS Multisource SystemESANE（Thomas Deroyon and Emmanuel Gros, フランス）

ウィンザー化[27]は、ある種の外れ値を検出し、推定量の分散に与える影響を抑えることを目的とする頑健な推定方法である。層化標本抽出では、各々の標本層に関して閾値を設定し、閾値よりも高い値を小さくする。この手法により得られる推定量には偏りがあるが、効率がよく精度が高い。2009年より、ESANE[28]において利用しているが、ESANEにおける統計値には様々な一貫性チェックがあるため、ESANEにおけるウィンザー化は売上高の変数のみに適用している。しかし、1つのコア変数（売上高）のみに基づく手法では、平均的な売上高を持つユニットの検出ができない上に、売上高との相関の低い変数において異常な値を持つユニットの検出もできない。このような欠点に対処するために、本報告では、アウトプットエディティング手法として、売上高以外の変数に対して、Kokic and Bellのウィンザー化手法の適用を行っている。

WP.19 Developing a Theoretical Framework for Selective Editing Based on Modellingand Optimization（Pedro Revilla, スペイン）

選択的エディティングの閾値を設定するためには、スコア関数が用いられるが、どのようなスコア関数を用いるべきかについて、普遍的な答えは見つかっていない。この問題に対処するために、スペインでは、理論的な枠組みの構築を行ってきた。確率的・組み合わせ論的な最適化問題として適切な選択を行う問題を定式化している。この問題を解決する出発点としては、選択されなかったユニットを訂正しなかったことによる推定値のエラーの増加を一定範囲に保ちつつ、人手訂正の数を制限することである。ある種のスコア関数を用いることにより、線形関係の条件化において、この問題を解決することができることを示した。実データを用いた実験においてもよい結果が示されているが、研究はまだ途上である。

WP.20 Changes in Macro-Editing and Score Functions for Dutch STS（Jeffrey Hoogland, オランダ）

オランダ統計局では、経済統計システムを再構築するにあたり、短期統計の統計作成プロセスの見直しも実施し、短期統計の新たなソフトウェアを2015年2月に実装した。2002年に実装された短期統計のソフトウェアIMPECT2には、様々な改善すべき点があった。例えば、ソフトウェアは制御できなくなっており、性能も悪くなる一方であり、方法論上の問題点も指摘されていた。さらに、予算削減のため、より少ない資源でより効率的にエディティングを行うことが求められていた。四半期の短期統計に関しても、新たな統計プロ

[27] Winsorization: データ内の最大値と最小値の影響を抑える補定手法

[28] ESANE: Élaboration des Statistiques ANnuelles d'Entreprises の略で「年次企業統計の精密化」という意味であり、イザーンと読む。

セスを構築した。これらのシステムでは、付加価値税データを利用できる際には、常に付加価値税データを利用している。しかし、付加価値税データは四半期ベースでのみ利用可能なので、月次の短期統計には適用できない。こういった場合には、月次の短期統計のデータを直接利用している。付加価値税データが利用できる際には、月次の短期統計データの人手訂正を行う情報源として活用している。

WP.21 Model-Based Selective Editing Procedures for Agricultural Price Indices（TizianaPichiorri, Daniela Ichim, Maria Liria Ferraro, and Ugo Guarnera, イタリア）

本報告では、イタリア国家統計局において開発してきたモデルベースの選択的エディティングソフトウェアSeleMix(Rパッケージ)を農業指数データに応用している。時系列上の差分を様々に設定したり、データの階層化構造を設定したりすることで、複数の実装戦略を試みた。比較対象として、連続した年について月次の変動は固定という前提に基づく簡易的な手法を用意した。比較の基準は、アルゴリズムの収束、検出した影響のある値の数、公表値への影響である。

WP.22 Selective and Macro-Editing of a Large Business Based Administrative Data Set（David R. H. Hiles, 米国）

雇用賃金四半期センサス(QCEW: Quarterly Census of Employment and Wages)は、月次の雇用データと四半期の賃金データを含んでおり、労働統計局において標本フレームとして利用できるようにエディットを行っている。このデータにおける950万の四半期レコードは、労働統計局のビジネスレジスターの基礎となるものである。これらのレコードは、集計した上で、労働統計局だけではなく、郡や地方レベルにおける雇用や賃金の基準として使用される。このように複数のレベルで使用することは、選択的エディティングとマクロエディティングにより構成されるエディット規則によって可能となっている。雇用賃金四半期センサスのサイズは膨大であり、四半期ごとの公表スケジュールも厳密であるので、横断的かつ縦断的なエディット手法による厳格なエディットが要求される。本報告では、現在使用されているエディットの長所と短所を評価し、今後の改善点を示している。

WP.23 Method for Reviewing Selective Editing Thresholds at ONS, RSI Pilot Study（Sangeetha Gallagher, Ben Graham, and Charlotte Gaughan, 英国）

小売店の売上高調査(RSI: Retail Sales Inquiry)は、英国における小売業における売上高を月次ベースで測る調査であり、エラーの検出には、選択的エディティングを使用している。売上高と雇用という2つの主要な変数に関して、選択的エディティングのスコアは、別々に算出される。小売店の売上高調査における閾値は、2010年に、前年の伝統的な人手によるエディティングで訂正されたデータをもとに算出された。データ及び推定値の

精度を保つためには、エラー検出において使用されている手法を定期的に点検することが必要である。定期的に閾値を点検することにより、選択的エディティングを適用したことによって生じるデータに関する問題も検出できる。選択的エディティングをパスした企業の副標本を抽出し、伝統的な人手によるエディティングを実行することで、閾値は点検できる。

トピック（ii）：エディティング及び補定に関する変更点の運用とサポート

WP.9 Redefining Roles and Responsibilities in a New Harmonized Statistical Production Process: Opportunities and Challenges（Etienne Saint-Pierre, カナダ）

過去5年間に渡って、カナダ統計局では、統計プロセス、組織構造、システムインフラに関して、非常に大きな変更を成功裏に実施してきた。財政的制約のある時代において、こういった変更によって、高品質のデータを作成し続けることができるようになった。最も大きな変更点の1つは、カナダ統計局における経済データの作成に関する手法と手順を完全に見直した点である。新たな経済統計モデルは、統合ビジネス統計プログラム(IBSP: Integrated Business Statistics Program)と呼ばれ、現在、標準化したプロセス・手法・ツールが、70のIBSPプログラムに適用されており、2019年までにさらに80の調査がモデルに組み込まれる予定である。カナダ統計局は、現在、統計調査の統計作成プロセスのすべてが、その調査を担当する部署に任されるという伝統的なモデルから決別しようとしている。こういった変更は、統一企業統計プログラム(UES: Unified Enterprises Statistics Program)によって15年前から始まったが、ここ5年の間に一気に進んだ。統合ビジネス統計プログラムでは、様々な統計作成プロセスにおいて、約10の異なる内部の部署が関わる。同時に、統合ビジネス統計プログラムでは、かつて使用されていた慣例を変更する革新的で調和的な統計プロセスを導入している。担当しているプログラムが統合ビジネス統計プログラムに統合された場合、数百人単位の職員が、自分たちの仕事のやり方を調整する必要がある。本報告では、カナダ統計局における企業調査の新たな実務的な枠組みの紹介を行っている。

WP.10 Managing Changes in the E&I Strategy of the Italian SBS（Orietta Luzi, イタリア）

2013年に、イタリア国家統計局の国民経済計算及び経済統計課では、Frame SBSという新たな情報システムの導入に着手し始めた。この新たなシステムでは、大多数のデータを行政データの統合によって収集し、残りのデータを調査によって収集する。この新たなシステムでは、予想どおり、正確さ・時系列上の統計値の一貫性・費用削減・回答者負担の軽減という点で得るものが大きかった。しかし、初期費用は、経済的な意味だけではなく、組織の文化的な意味も含めて、非常に高いものであった。このシステムの変更は、実際に

完遂するまでに約3年の月日を要したが、イタリア国家統計局の役員からの強いサポートがあり、管理職、統計家、IT専門家、行政データスペシャリストの間で強い協力体制の構築が望まれた。システムに変更を導入することは、新たな手法やツールの実装が必要という意味で、時間や費用がかかるのは言うまでもなく、それ以外にも、調査担当者のトレーニングにも時間を要するし、新たなデータエディティングシステムの原則を受け入れられるようにするためには、非常に時間と労力が必要なのである。こういった移行期の困難さは、統計作成システム全体を標準化し、ITツールを最大限活用して機械化することによって軽減することができた。

WP.11 Implementation of Selective Editing Methods at Statistics Finland Using Innovative and Efficient Team Work Methods（Saara Oinonen, フィンランド）

近年、フィンランド統計局では、エディティングを改善しようと試みており、EG EDITというSASベースのエディティング用ツールを導入している。EG EDITには、スウェーデン統計局において開発されたSASベースの選択的エディティングプログラムSELEKTやカナダ統計局において開発されたSASベースのエディティングモジュールBANFFを組み込んでおり、その他、フィンランド統計局において開発したマクロプログラムも備えている。2014年9月から、8つの統計調査において、選択的エディティング手法を実装するプロジェクトを開始した。この実装では、統計作成に携わる職員、統計技術の専門家、データ収集に携わる職員、IT専門家の協力と努力が欠かせない。効率的なチームワークとwiki情報を活用したトレーニングなどにより、良好な結果が達成されている。これまでのところ、5つの統計調査において、EG EDITによる選択的エディティング手法は成功裏に実装されてきている。2016年中には、8つの統計調査すべての統計作成プロセスにおいて、EG EDITが用いられることとなる予定である。

WP.12 Improvement of the Quality of Statistics by Mezo-Validation（Miklósné Paczári and Katalin Kovács, ハンガリー）

メゾ妥当性検証[29]は、事後的な妥当性検証として定義されているが、その目的は、できるだけ早い段階でエラーを検出することである。したがって、前年のメゾ妥当性検証の経験に基づいて、すでに初期のデータ処理段階において一貫性に注意を払い、その後、メゾ妥当性検証によってチェックを行う。このようにすることで、データチェックや訂正を追加で行うことを避けたり、その量を減らしたりすることができる。将来的には、経済組織、会社グループ、機関など、様々なものに適用できるよう、メゾ妥当性検証を拡張する予定である。究極的な目標は、可能な限り最も包括的な方法で統計データの収集を行うことである。

[29] 原語では mezo-validation と表現されている。従来からの micro-validation と macro-validation のいずれでもない手法

WP.13 Data Collection Optimization – First Attempt (Agnes Andics and Gergely Horváth, ハンガリー)

ハンガリー中央統計局では、調査回答の期限を知らせるリマインダーシステムを実装している。このシステムは、多くの場合、回答者への照会を自動で行うことができるものの、時として、電話や郵送によって照会を行わなければならないこともある。こういったケースでは、データ収集の費用を押し上げることになる。費用対効果という観点から、調査回答期日を過ぎた後になって未回答者から情報を収集することにどれだけの意味があるかを考えなければならない。すなわち、労力に見合うだけデータの品質が改善するのか、もしそうでないとすれば、それだけの労力を別の業務に振り分ける方がよいかもしれない。本研究の目的は、高度な自動化手法を開発し、データ収集の様々な段階における推定量の変化を記録し、ある調査において利用可能なデータの質と量が、どの段階で十分になるかを示すことである。このようなシステムを利用することで、効率性（能率性）の向上につながるものである。

WP.14 Imputation at the National Agricultural Statistics Service（Darcy Miller and Linda J. Young, 米国）

米国農務省の国立農業統計局では、農業センサスと農業関連の標本調査を2つの主なプログラムとして扱っている。農業センサスは、5年ごとに実施され、農業政策の基盤となる情報を提供する。農業センサスで得られた情報は、自治体計画、企業誘致、営業貸付金の額、サービスセンターの充実度、農業プログラムや政策に関する意思決定のために用いられる。農業関連の標本調査では、米国内における農業関係の実質的にすべての側面に関して調査を行い、多くの場合、市場における機微な情報が含まれている。農業センサスと農業関連の標本調査は、相互補完的に、すべての市場関係者に農業部門における需要供給の情報を提供し、そうすることで、競争市場における効率性と公平性を促進するのである。国立農業統計局では、エディティング及び補定の改善策を常に模索しており、改善された手法をより多くの調査に適用するよう心がけている。ISR、IVEware、SignEditといったモダンな統計補定手法の導入は、これまで行ってきた人手によるすべてのレコードの訂正から、推定値の整合性を維持するために分布をモデル化するという文化的なシフトを反映している。現在、複数の調査をSignEditというシステムを用いてエディティングを行っている。国立農業統計局におけるエディティングと補定では、一貫性と処理効率を向上するために人手による介入をできるだけ排除し、エディティングと補定のプロセスにおける変動を説明し、全体的なデータ品質の改善を目標としている。

WP.15 Getting Commitment to a New Editing Strategy（Felibel Zabala，ニュージーランド）

2011年より、ニュージーランド統計局では、公的統計の作成方法を変更する10か年計画(Statistics 2020 Te Kāpehu Whetū)に着手し始めた。このプログラムは、現在進行中であるが、処理プラットフォームの中核をなすのは、徹底して標準化した処理方法の実行である。このシステムにより、エディティングと補定手法の統合化を行うことができ、出力結果の効率性を向上し、品質を改善することができる。社会調査データを処理するための世帯処理プラットフォーム(HHP: Household Processing Platform)を開発した。世帯経済調査(HES: Household Economic Survey)は、現在、世帯処理プラットフォームに移行中である。世帯経済調査を世帯処理プラットフォームに移行することで、エディティングと補定を含む方法論的な改善を実装する機会が得られた。現行の世帯処理プラットフォームには、データをエディットする機能が備えられていない。世帯経済調査は、世帯処理プラットフォームに移行した調査の中でエディティングを要する初めての調査である。必要なエディティング処理プロセスに関して同意を形成するために、世帯経済調査の担当職員とのワークショップを実施したことなど、どのような変更方法を導入したかを報告している。また、新しい手法の理解を促進するために、サポートやトレーニングも提供した。

WP.16 Managing and Supporting Changes Related to Editing and Imputation in the United Kingdom（Jill Tooze and Julie Curzon，英国）

本報告では、英国国家統計局の経済データ部門が、予算削減の要求を満たすために、データ収集と妥当性検証をどのように発展させてきたかについて記述している。エディティングや照会のプロセスを再評価したり、組織構造の改革を行ったり、技術上のプログラムを導入したりといった事例を紹介している。導入された手法の中には、選択的エディティングも含まれている。組織の戦略に変更があった場合、効果と効率を具現化するために、構造・役割・機能もそれに合わせて再編成されなければならない。具体的な教訓として、以下のものを挙げることができる。①機能を分割する際には、細心の注意が必要である；②様々な部署同士の頻繁な意思疎通が必須である；③役割と責任を明確にするために、導入当初の段階で、エリアの境界をはっきりさせておくべきである；④導入は段階的に行い、定期的な評価を行うべきである；⑤器用貧乏にならないように気をつけなければならない；⑥フォーカスグループやワークショップを通じて、職員に変更点を周知するべきである；⑦新しいシステムに信念を持っているリーダーを選ぶべきである。

トピック（iii）：ソフトウェアツールと国際協力

WP.2 Towards a Generic Approach to Validation: the ValiDat Foundation Project（Marco Di Zio, Nadenshda Fursova, Lucas Quensel-von Kalben, and Olav Ten Bosch, イタリア、リトアニア、ドイツ、オランダ）

2009年から欧州統計システム(ESS: European Statistical System)では、国家・領域・プロセスといった境界を越えて、統計作成を継ぎ目なく統合する新たな手法構築に乗り出した。ヴィジョン実行プロジェクト(Vision Implementing Projects)を通じて、欧州統計局は、手法を調和させ共通のインフラを構築するイニシアティブをとった。汎用的統計ビジネス作成モデル(GSBPM: Generic Statistical Business Production Model)では、効率性の向上と質的な改善という意味で、データ妥当性検証に高いポテンシャルが期待されていたため、当初からデータ妥当性検証は主要なエリアの1つとして認識されていた。このポテンシャルが花開くためには、ESS加盟国における非常に多様な統計作成環境を統一的な欧州手法にする必要がある。2014年の後半には、この目的を達成するために、4加盟国（ドイツ、イタリア、オランダ、リトアニア）からの代表を集めて、EU後援による1年間のプロジェクトが開始された。国家統計局におけるデータ妥当性検証に用いられている主な手法に関して、再検討が開始された。成果物の1つとして、共通の方法論的枠組みを構築する「妥当性検証のハンドブック」を挙げることができる。本報告の目的は、このプロジェクトの存在を広く知らしめ、より多くの関係者にアクティブな参加者として関わってもらうことである。

WP.3 The ValiDat Foundation Project: Survey on the Different Approaches to Validation Applied Within the ESS（Sarah Giessing and Katrin Walsdorfer, ドイツ）

欧州統計システム(ESS)加盟国において、典型的な妥当性検証が実際にどのように実装されているかを把握するために、加盟各国の国家統計機関に調査票を郵送した。その目的は、妥当性検証手法の包括的な概略を構築し、応用における実務上の問題点を明らかにすることである。この調査は、ESSにおける妥当性検証手法を体系化する最初の試みである。調査票は、共通部分と領域に特化した部分とに分けられ、技術的な観点と方法論的な観点から妥当性検証について情報収集を行った。

WP.4　Flash Estimates for Short Term Indicators – Data Cleaning with X12 ARIMA（Markus Froehlich, オーストリア）

工業・建設業の短期統計は、絶対データとして利用可能であり、EUレベルで算出する短期統計指標を算出するための主要な統計である。調査期日の最終日から90日以内に絶対データは公表しなければならないが、指標は60日以内に公表することが欧州委員会から要求されている。いくつかの指標については、さらに30日短くなる予定である。現行のデ

ータ処理方法は、このシナリオを実現することが難しく、新たな推定手法を使用しなければならない。EUの要求を満たすことのできる推定方法として考えているのは、単変量及び多変量の時系列による補定モデルである。公表期日までの時間が非常に短いため、データクリーニングとエディティングは、通常の方法で行うことができない。しかし、生データは不正確な値やエラーによって大きく影響を受けているため、データエディティングは根本的に重要である。X12 ARIMAによるデータエディティングの検証を行った。

WP.5 A Formal Typology of Data Validation Functions（Mark van der Loo, オランダ）

データ妥当性検証は、いかなる統計作成においても不可欠なものであり、本報告では、数学的関数として定式化できる形で定義を与える。測定プロセスを丹念に調査することで、データポイント及びデータセットを識別するのに必要な最小の特性を導き出す。妥当性検証プロシージャが関数として適用される領域の種類を判別することで、妥当性検証プロシージャを10個の異なるグループに分類する。妥当性検証関数の領域ごとにデータ特性がどれぐらい変化するかをカウントすることで、自然と妥当性レベルの定義を導き出すことができる。

WP.6 Integrated Data Entry and Validation System in HCSO（Erzsébet Kómár, ハンガリー)

ハンガリー中央統計局における統合的データ入力・妥当性検証システム(ADEL)は、調査データの妥当性検証と訂正の機能を備えている。このシステムは、他のITシステムとも統合されており、通常のコミュニケーションインターフェースを使用している。このメタシステムでは、電子データ収集システム(ELEKTRA)から調査票を入手し、行政データのデータ収集システム(ADAMES)からデータセットを入手し、調査制御情報を読み込んで更新する。そして、ADELシステムは、妥当性検証済みデータを統合データ処理システム(EAR)に提供し、公表用の表計算システムに情報を送る。ADELシステムの標準化は徹底的に行われており、最終的な適用において重要なだけではなく、ヒューマンエラーを最小化するためには途中の段階においても重要である。ADELシステムは、データ入力と妥当性検証の目的で、常に変化し続けている環境の中で、約15年使われてきている。

WP.7 Usage of External Software Tools at SURS – Experiences and Lessons Learned so far（Rudi Seljak, Andreja Smukavec, and Igor Kuzma, スロヴェニア）

スロヴェニア共和国統計局は、明らかに小さな機関であり、既存の開発済みアプリケーションを共有して使用することは、これまでにも頻繁に行われてきた慣行である。カナダ統計局のSASマクロのBanffは、エディティング及び補定のプロシージャを補助する目的で開発されたものであり、スロヴェニア共和国統計局では、2008年から使用している。CALMARは、1990年代初頭にフランス国立統計経済研究所によって開発されたSASマク

ロであり、スロヴェニア共和国統計局では、キャリブレーションを行う必要がある際に使用している。Tau Argusは、統計表を保護するソフトウェアであり、欧州における複数のプロジェクトの成果である。スロヴェニア共和国統計局では10年以上前から使用している。Demetraは、季節性調整を行うソフトウェアであり、初版は1990年代後半に欧州統計局によって開発された。スロヴェニア共和国統計局では、2年前からDemetra+を使用している。反対に、スロヴェニア共和国統計局で開発したSTAGEは、地図作成のウェブアプリであり、地理空間情報の作成に適しており、この製品の共有を行っている。

WP.8 Editing and Imputation in Household Based Surveys – Case of Household Budget Survey in Bosnia and Herzegovina（Edin Šabanović, ボスニア・ヘルツェゴヴィナ）

全数調査であれ標本調査であれ、あらゆる統計調査において非標本誤差は存在する。そのような誤差の例としては、欠測値の問題がある。欠測値には、全項目無回答と一部項目無回答の2種類がある。全項目無回答の問題は、通常、重み付けによって解決できるが、一部項目無回答は補定手法によって解決しなければならない。汎用的統計ビジネス作成モデル(GSBPM)において、エディティングと補定は非常に重要な位置を占めている。本報告では、過去15年のボスニア・ヘルツェゴヴィナにおける世帯予算調査において使用されているエディティングと補定の手法、及びソフトウェアについて紹介している。また、単純な手法から高度な手法まで、異なるエディティングと補定の手法を用いた場合にどれだけの改善がなされるかについても報告している。

トピック（iv）：評価とフィードバック

WP.25 Editing Big Data: An Holistic Approach（Marco Puts and Piet Daas, オランダ）

オランダでは、約6万の道路センサーによって収集された分単位の自動車カウント情報により交通の非常に詳細なイメージを入手することができる。交通マネジメントの観点では、すでに、渋滞予想や移動時間短縮など、いろいろな使用方法が考案されている。オランダ統計局では、このデータを交通統計に使用している。本報告では、オランダの主要幹線道路において2万のセンサーによって収集されたデータを使用している。2010年から2014年までに、トータルで1150億レコードが収集され、80テラバイトのファイルに保存されている。技術的な意味では、このデータはクリアーなデータ構造をしており、非常に構造的だが、データの中身はあまり構造的になってはいない。例えば、道路センサーと中央データベースとの間のシグナルが途切れることにより、値がたびたび欠測したり、センサーは定期的に機能不全に陥ったり、隣接する道路センサー同士の関係は想像しているほど明らかではなかったりという問題がある。それぞれの自動車は異なった速度でセンサーの前を通過しており、1分あたりの標本頻度は「たったの」1台に限られているため、2箇所のセンサーのデータ間に大きな相関を見出すことができない。たとえ、わずか250

メートルしか離れていなかったとしても、こういった問題が起きる。こういったことが原因で、隣接するセンサーの結果を比較することだけでは、データのクリーニングを適切に行うことができないのである。本報告では、このような交通情報に関するビッグデータにおける欠測値の推定方法と隣接するシグナル間の相関を増やす方法について議論している。

WP.26 Editing Process and Its Quality Regarding Design and Production Phases Using Process Metadata and Calculation Modules（Pauli Ollila, フィンランド）

フィンランド統計局におけるエディティング戦略は、デザイン・ITの実用化・検証・生産という4段階で発展している。これらは、汎用的統計ビジネスプロセスモデル(Generic Statistical Business Process Model)において、エディティング戦略を発展させる際に必要とされる4つの段階を反映したものである。エディティングプロセスの構造は、UNECEの作業部会において議論されてきた。パラメータ化と適切なITソリューションを行うのに必要な手法という観点で、統計作成過程、プロセスの流れ、プロセスの段階とつながりにおけるエディティングについても研究している。様々な形でメタデータの果たす役割についても考慮をしている。統計作成環境と検証環境におけるモニタリングと評価についても研究した。プロセス・メタデータ・システムは、モニタリングと評価の性質及び品質を測る基盤を形成している。また、計測のできない類の評価についても、こういった文脈で評価を行っている。改善の期待できるプロセスについても研究を行った。この7段階のカテゴリーは、作業部会におけるプロセスレベルとこれらのレベルにおけるITシステムの実用化に基づいている。

WP.27 Analysis of the Data Preparation Process of the Structural Survey of the Federal Population Census（Daniel Kilchmann and Beat Hulliger, スイス）

連邦人口センサスの構造的調査は、スイスにおけるレジスターと標本調査を統合したセンサスシステムの一部である。2010年以来、毎年、約25万人の標本が選ばれている。一部項目無回答、矛盾、外れ値は、統計データ準備プロセス(SDPP: Statistical Data Preparation Process)において検出され処理される。このプロセスは、2007年に刊行されたEDIMBUSプロジェクト[30]によって推奨されているもので、いくつかの段階に分けて実行されている。スイス連邦統計局では、この統計データ準備プロセスを分析するプロジェクトを開始した。その目的は、統計データ準備プロセスの結果に与える影響をよりよく理解し、また、その影響が統計データ準備プロセスの段階ごとにどのように変化していくのかをよりよく理解することである。このプロジェクトの結果によって、EDIMBUSに基づ

[30] EDIMBUS プロジェクトについては、下記も参考にされたい。小林良行（2009）「ヨーロッパにおけるデータエディティング及び補定に関する調査報告～EDIMBUS プロジェクトを中心に～」, 『統計研究彙報』第 66 号, no.4, pp.101-129. http://www.stat.go.jp/training/2kenkyu/pdf/ihou/66/kobayashi.pdf

く概念的枠組みと統計データ準備プロセスに基づくプロセスデザインが適切なものであるかどうかがはっきりと示される。さらに、この結果によって、統計データ準備プロセスの段階において、どのような指標を算出すればよいかについても示唆が得られる。

WP.28 Editing and Evaluation of Statistics Based on Administrative Microdata – Example by Norway（Aslaug Hurlen Foss and Ane Seierstad, ノルウェイ）

エディティング及び推定のための統合システム(ISEE: Integrated System for Editing and Estimation)は、ノルウェイ統計局において開発中の汎用的ITシステムである。ISEEシステムを汎用的に近代化する新たなプロジェクトを開始した。主要な問題の1つは、行政ミクロデータに適したエディティングシステムをどのようにして構築するかである。評価のレポートモジュールの改善を目指し、人口抑制の汎用化を促進したいと考えている。また、マクロエディティングを効率的に使用することのできるシステムの構築も目指している。行政ミクロデータにとっては、これは集計値レベルのデータを使用することを意味する。ノルウェイ統計局のIT専門家は、統計環境Rによりマクロエディティングのライブラリーを構築できないか検討している。また、ライブラリーとデータベースをつなぐ別のソフトウェアの開発も視野に入れている。統計環境Rによって手法を構築することで、他の統計機関と共有することが容易になるであろう。

WP.29 Evaluation of Census 2011 Survey Estiamtes（Lydia Spies, ドイツ）

2011年に、ドイツでは、初のレジスターベースの人口センサスが行われた。過去のセンサスでは、すべての人口を全数調査していたが、今回のセンサスでは、人口の10%だけを標本調査している（それ以外はレジスターからの情報を利用するという意味）。公表数値の精度は、センサス法において相対標準偏差（変動係数）0.5%以内と規定されている。したがって、正確な分散の推定を行うことは、データ作成プロセスにおいて重要である。すべての調査変数には、補定値が含まれているので、補定による分散も考慮に入れなければならない。この目的を達成するための1つの実用的な方法は、多重代入法(multiple imputation)を用いて複数回の補定を実行することである。そのために、2011年センサスのデータを用いて検証を行っている。補定分散の影響を評価し、前回のセンサスで用いられた単一代入法(single imputation)ではどの程度のエラーの過小評価が起こっていたかを検証する。

WP.30 Using the CURIOS Algorithm to Manage the Prioritization of CAPI Surveys（Antoine Rebecq and Thomas Merly-Alpa, フランス）

フランス国立統計経済研究所(INSEE)では、CAPI[31]調査の優先付けのために、標本の代

[31] CAPI: Computer-Assisted Personal Interviewing（コンピュータ支援による対面調査）

表性を最適化するCURIOSアルゴリズムを用いている。同様の手法として、CATI[32]の優先付けを行うStatCanがあるが、フランスのCAPI調査の構造では、複数の変更をまとめて実行することができないため、StatCanを利用することはできない。我々の手法は、2段階標本デザインに基づいており、第1段階において学習を行い、第2段階において標本デザインの変更を実施する。第2段階の標本は、損失関数の期待値を最小化することで算出される。この損失関数は標本の品質に関わる複数の要因の線形結合に基づいており、期待値はモンテカルロ手法によって算出される。このアルゴリズムの主な目的は、無回答を考慮に入れた上で、ウェイトのばらつきを最小化することである。その結果、特にデータにキャリブレーションが適用される場合には、より頑健な推定量を算出することができる。このアルゴリズムは、すでに、例えば2014年の世帯資産調査など、フランスにおけるいくつかの世帯調査に適用されている。このアルゴリズムは、CAPIを用いるどのような調査にも使用することができる。

トピック（v）：革新的手法及びデータ革命

WP.31 Let the Data Speak: Machine Learning Methods for Data Editing and Imputation（Felibel Zabala, ニュージーランド）

公的統計において、エラーデータや無回答データの処理は避けて通れない問題である。これらの問題は、データ収集の品質に悪影響を与え、結果として、そこから出力される結果の品質にも悪影響を与える。データ収集におけるこのような問題を未解決のままにすると、推定値に偏りが発生し、出力結果の品質低下につながる。エラーがどのようなものであるかを深く理解することで、よい品質のデータを作成できるデータエディティングと補定のプロセスにつながっていく。機械学習を用いることで、様々なデータ収集におけるエラーデータや無回答の特性を理解する一助となる。機械学習とは、明示的なプログラムなしでコンピュータに学ばせる手法である。本報告では、ニュージーランド統計局における世帯経済調査の収入変数に関して、データエディティングと補定を行う際に機械学習を用いる方法について述べている。

WP.32 Estimation and Editing for Data from Differennt Sources. An Approach Based on Latent Class Model（Ugo Guarnera and Roberta Varriale, イタリア）

近年、多くの行政情報源が活用できるようになり、また回答者負担の軽減の目的で、多くの国の公的統計では、「調査票によるデータ収集」から「行政データによるデータ収集」へと移行している。しかし、各々の行政データは、別の目的で構築されているため、定義や変数の尺度など、前処理をしなければならないことが多い。とりわけ、統計アーカイブを構築することが目的である場合には、どの情報源から使い始めるべきかという問題が

[32] CATI: Computer-Assisted Telephone Interviewing（コンピュータ支援による電話調査）

ある。1つの方法として、それぞれの情報源からのデータ品質をあらかじめ分析することで、序列をつけるという方法が考えられるが、情報源同士の序列を明確につけることができない場合が多い。そこで、本報告では、異なる情報源からの情報を同時に利用する手法を提案している。この手法では、真のデータと測定プロセスをモデル化する。この手法を用いることで、情報源ごとの信頼性に応じて、すべての情報に重み付けを行うことができる。

WP.33 An Assessment of the Feasibility of Editing and Imputing Administrative Tax Return Data to Provide a Substitute for Survey Data（Charlotte Gaughan, 英国）

英国国家統計局(ONS)と英国歳入関税局(HMC)は、現在、企業の売上高や収入といった主要な経済変数の収集を重複して行っている。この可能性検証研究の目的は、納税申告データを用いたエディティングと補定によって、調査データの代わりとするのに十分な質のデータが確保されるかどうかを確認することである。英国国家統計局では、以前にも、付加価値税データを用いた手法の検証を行ったことがある。税務データを用いることの限界、税務データと現行の調査データによる推定値を比較するための手法、税務データをエディットし補定するための技術について報告している。この研究の成功は、税務データの最終版と現行の推定値との正確な比較ができるかどうかにかかっている。初期の予備研究は、2012年の売上高変数を用いて行った。今後、他の変数の調査も行う予定である。

WP.34 Multiple Ratio Imputation by the EMB Algorithm（Masayoshi Takahashi, 日本）

米国センサス局、英国国家統計局、オランダ統計局など、公的統計における欠測値は、比率補定(ratio imputation)により処理されることが多い。一方、通常の比率補定は、推定不確実性を評価できず、多重代入法(multiple imputation)の使用が推奨されるが、これまで多重比率補定(multiple ratio imputation)の研究はされていない。本研究では、ブートストラップに期待値最大化法を適用するExpectation-Maximization with Bootstrapping (EMB)アルゴリズムに基づく新たな多重比率補定法を提唱している。本報告では、独自に開発した多重比率補定の*R*関数をシミュレーションデータに適用して検証している。また、実データを用いてその有用性を示している。独自に開発したソフトウェアMrImputationは、*R*関数mrimputeとmranalyzeから構成され、*R*関数mrimputeは多重比率補定を実行し、*R*関数mranalyzeは多重比率補定済みデータを用いた統計解析を実行するものである。

WP.35 New Results on Automatic Editing Using Hard and Soft Edit Rules（Sander Scholtus, オランダ）

収集データ内のエラーを検出するためには、通常、一連のエディット規則に照らし合わせて矛盾の生じる値を探し出すものである。ハードエディット規則はエラー値を確定的に見つけるものであり、ソフトエディット規則はエラーを確率的に見つけるものである。人

手によるエディティングではこの両方の手法を利用しているが、現行の自動エディティング手法は基本的にハードエディットであり、ソフトエディット規則の利用は非常に限られている。自動エディティングの可能性を広げるために、オランダ統計局では、ハードエディット規則とソフトエディット規則の区別ができる新たなエラー特定手法を開発した。現在のところ、この新たな手法は、小規模な合成データによるシミュレーション研究に限られていたが、ハードエディットとソフトエディットを用いたエラー特定に関して、2つの新たな結果を報告する。1つ目は、拡張版のエラー特定問題は、ハードエディット規則のみを用いることで再定式化することができる。これは、既存のソフトウェアを用いることで、この問題を解決できることを示唆している。2つ目は、オランダ構造企業統計の実データを用いて、より現実的な文脈における評価を行ったことである。

トピック（vi）：汎用的なプロセスの枠組みを構築する作業部会の報告

WP.24 Generic Statistical Data Editing Models（Saara Oinonen, Pauli Ollila, Marjo Pyy-Martikainen, Emmanuel Gros, Marco Di Zio, Ugo Guarnera, Orietta Luzi, Li-Chun Zhang, Jeroen Pannekoek, Tetyana Kolomiyets, and Steven Vale, フィンランド、フランス、イタリア、ノルウェイ、オランダ、UNECE 事務局）

前回フランスで開催された 2014 年の UNECE 統計データエディティングに関するワークセッションにおいて、統計データエディティングの汎用的なプロセスフレームワークを構築するべきだと提案された。今回のワークセッションでの報告を目標として、UNECE が音頭を取り、2014 年 8 月に作業部会が設置された。作業部会のメンバーは、フィンランド、フランス、イタリア、ノルウェイ、オランダ、UNECE である。この作業部会の成果物は、一連の汎用的統計データエディティングモデル(GSDEMs: Generic Statistical Data Editing Models)であり、40 ページに及ぶ詳細な文書を作成した。GSDEMs は、標本調査における標準的モデルや手法と同様に、統計データエディティングの標準レファレンスと考えられるべきものである。

製 表 技 術 参 考 資 料 31

平成 28 年 3 月発行

編集・発行　　独立行政法人　統計センター

〒162－8668

東京都新宿区若松町 19－1

電　話　代表　03（ 5273 ）1200

統計情報・技術部統計技術研究課　　TEL ： 03－5273－1368

E-mail ： research@nstac.go.jp