第４章　技術の研究に関する事項

第１節　技術研究を専任で行う組織の充実

第１　外部研究者の採用に係る検討及び実施並びに統計センター内研究会等への外部研究者の参加の推進

外部の研究機関、大学等との人材交流を推進し、統計センター職員の研究能力の向上及び製表技術の高度化・改善を図るため、外部研究者の採用について検討を進めた。平成16年７月に非常勤職員１人を採用し、集計表の秘匿処理法に関する研究を行った。

また、統計センターの製表業務において中核となる技術の一つであるデータエディティングについて研究を進めるため、平成15年度に引き続き外部有識者をメンバーとした「データエディティング研究会」を開催し、欠測値の補定法の改善、多変量外れ値検出法の適用可能性、調査においてプレプリント方式を採用することによるデータエディティング業務の軽減効果などについて検討を行った。

第２節　研究計画

第１　データエディティングに関する研究

１　研究の概要

製表業務の中核の一つであるデータエディティング（欠測又は矛盾したデータ項目を訂正することを目的とした手法）に関する技術の向上及び業務の効率化に資するため、諸外国における研究動向を把握するとともに、実証研究を進めている。

平成16年度においては、外部有識者をメンバーとした「データエディティング研究会」を開催し、欠測値の補定法の改善等について検討を行った。

研究会の構成員は、研究担当理事を主査とし、外部研究者３人、統計センター４人の計８人となっており、ほかに総務省統計局を含めた関係部門の担当者が参加した。

研究会の開催実績は、表のとおりである。

表　平成16年度データエディティング研究会開催実績

回数	開催年月日	議題
第１回	16.10．8	・年収及びその内訳の欠測値の補定法の研究 ・多変量外れ値検出法の比較 ・密度推定法によるインライアー検出の試み
第２回	17．3.10	・年収及びその内訳の欠測値の補定法の実験結果 ・密度推定法を利用した二変量外れ値検出法の実験結果 ・プレプリント方式のデータ品質改善効果に関する欧米諸国の研究動向

２　データエディティングに関する情報収集及び資料の整備

データエディティングに関する研究を推進する上で、特に研究の盛んな欧米諸国の研究動向に関する情報収集が重要であることから、アメリカ・カナダ統計学会などが共同開催した「合同統計会議」（８月７日～14日、カナダ／トロント）に参加したほか、多変量外れ値検出法の研究動向の調査・分析（製表技術研究レポート１に所収）などデータエディティングに関する文献の収集・分析、国連欧州経済委員会刊行のデータエディティングに関する用語集（製表技術関連資料集１に所収）やカナダ統計局における多変量外れ値検出法の適用例（製表技術研究レポート１に所収）など主要な文献の翻訳を行い、関係部門の参考に供した。

参　考

研究の内容

1　欠測値の補定法に関する研究

平成15年度においては、平成11年全国消費実態調査結果を用い、年収欠測値の補定法の改善について検討を行った。その結果、回帰補定法と最近隣補定法を組み合わせたPredictive mean matching法が、原データの分布を保存した補定が行える点で、従来の回帰補定法に比べてより適切な補定法となる可能性があることを示した。

平成16年度においては、Predictive mean matching法の改良、及び所得格差に関する集計を行う場合などに望ましいもののこれまで適切な補定法が見出されていなかった内訳項目の補定法の検討を行った。その結果、確率的な補定値の選定法を採り入れたFractional predictive mean matching法やFractional dissimilarity matrix法を用いることにより、原データの分布の保存と補定に伴う誤差の抑制を両立させた補定が可能になることが示された。また、各標本の抽出率が異なり、等しいウェイトでない場合でも偏りが生じない方法が見出された。

内訳項目の補定については、多変量に一般化したPredictive mean matching法などの適用を試みたが、年間収入（内訳項目の合計）の補定と両立させることが容易ではなく、各内訳項目と合計の両者を同時に考慮した多次元Predictive mean matching法を適用することによって両者とも比較的良好な補定結果が得られたものの、年間収入の補定の精度を維持しながら内訳項目の補定も行うには、更なる技術的改良が必要と判断された。

2　効果的なデータエディティングに関する研究

(1) 箱ひげ図及び散布図など統計的エディティング、グラフィカル・エディティングの研究

箱ひげ図（Boxplot, Resistant fence）は、１変量の外れ値検出法として広く使われている方法であるが、この手法を含めてデータ分布全体の状況に基づいてチェックを行う統計的エディティングや視覚的なエディティングの方法が関係部門において必ずしも十分に認識されていない現状を踏まえ、諸外国における適用状況をまとめ、周知を図った。その結果、平成14年全国物価統計調査の価格分布データのチェックに活用された。

(2) 多変量外れ値検出法に関する研究

多変量外れ値検出法は、官庁統計の分野では、諸外国でもほとんど使用されていない。これは、従来の多変量外れ値検出法が大量データの処理に適さなかったことや、官庁統計で扱われるデー

タが大きく偏った分布になっていることが多いことなどが理由と考えられる。しかし、異常値の検出を属性間の相関を考慮しながら効率的に行うには、散布図などを利用する視覚的な方法だけでなく、多変量の外れ値検出が利用可能になることが望まれる。

近年、Fast-MCD法やBACON法など計算効率の高い方法が開発される一方、分布の形状に関する前提が少ないノンパラメトリックな方法の研究も行われている。そこで、諸外国における研究動向を調査・分析するとともに、シミュレーション及び小規模な実データを用い、多変量外れ値検出法の比較を行った。その結果、Fast-MCD法は、偏りの少ないデータ分布に対しては外れ値の検出力が高いが、偏りが無視できないデータ分布に対しては検出力が大きく低下する。BACON法は、計算効率が非常に高く、偏りが比較的大きいデータ分布であっても検出力の低下は小さいが、外れ値がデータ分布に近接している場合に検出力が極端に低下する可能性があり、それぞれ一長一短があることが示された。ノンパラメトリックな方法については、密度推定法を利用する方法などの研究を行ったが、実用可能な方法を見出すためには、さらに研究が必要と判断された。このため、Fast-MCD法、BACON法などの業務適用可能性を実際の調査データを用いて検証するとともに、新たな方法の探索が必要と判断された。

(3) インライアー検出法に関する研究

正常値データの分布の中に含まれているエラーデータ（インライアー）については、個票データそれぞれに対して論理的なチェックを行う方法が基本となるが、例えば、通常ピークが一つだけの分布に複数のピークが現れた場合を検出するなど、データ分布の形状から判定する有効な方法があれば、集計結果の正確性確保に役立つと考えられる。そこで、平成14年全国物価統計調査の価格分布データなどを用い、密度推定法によってピークの位置及び数を推定する実験を行ったところ、Silverman検定に類した方法を適用することにより正確にピークの位置及び数を推定でき、価格データのインライアー検出に利用可能性があることが示された。インライアー検出法に関する研究や経験が非常に少なく、データ分布のピーク数を推定する方法の一般性を含め研究の方向性が明確にし難い面があるものの、さらなる研究が望まれると判断された。

(4) プレプリント方式のデータ品質改善効果に関する研究

統計センターの製表業務におけるデータエディティング業務を軽減するには、調査段階で改善を行い、誤りの少ない回答が得られるようにすることが最も効果的であるものの、我が国ではこれに関する研究がほとんど行われていない現状を踏まえて、回答データの品質向上につながる調査段階での改善方策について外国文献の調査を行った。その結果、前回調査の回答を同一被調査

者に示すプレプリント方式が、調査項目によっては有効であり、外れ値の発生などを大幅に減らす可能性があることが判明した。この結果を統計研修所開催の研究報告会などの場においても報告することにより、調査企画を行う統計局職員への周知も図った。

第２　統計分類の自動格付に関する研究

１　研究の概要

製表業務の中核の一つである分類格付事務の自動化を図るため、国内外における関連研究の動向を把握するとともに、実証研究を進めている。

平成16年度においては、事業所・企業統計調査産業分類の自動格付に関する前年度の研究成果をまとめるとともに、自動格付法の改善などを行った。

２　産業分類自動格付システムの研究

平成16年度においては、前年度の実験結果（製表技術参考資料２に所収）に基づいて決定された平成16年事業所・企業統計調査の産業分類審査への自動格付法の適用をより有効なものにするため、事業所・企業統計調査産業分類の自動格付法の改善を行った。

前年度と同様に日本標準産業分類が大幅に改訂されたことも考慮し、新分類に再格付された平成13年事業所・企業統計調査データを用いて実証研究を行った。その結果、前年度の実験で用いた地域別の全事業所データを学習用データに用いる代わりに、審査対象が新設事業所であること、審査済みデータを学習用データに用いることができること、十分な学習用データを確保することを考慮して、（前回調査の）全国の新設事業所データを学習用データに用いることにより、分類性能を向上させることができた。さらに事業の種類のみによる自動格付と事業の種類及び取扱商品による自動格付のうち、第１候補の推定確率が高いほうを選択する合成方式の採用や、単語間の文法的関係を考慮したルール生成を行うことにより、分類性能を向上させることができた。

３　統計分類自動格付法に関する情報収集及び資料整備

製表事務の質の向上及び効率化を推進する上で、統計分類の自動格付は、手書き文字の自動読み取りとともに最も実現が望まれている製表技術の一つであるが、完全といえる水準に引き上げるには技術的に大きな課題がある。

そこで、国内外における統計分類の自動格付法あるいはテキストの自動分類に関する研究につ

いて情報収集を行うため、情報処理学会自然言語処理研究会及び言語処理学会に参加し、新しいテキスト自動分類法の研究動向の把握に努めたほか、国内外の関連文献の収集・分析を行い製表技術参考資料２にまとめ、関係部門の参考に供した。

第３　統計ニーズの多様化に対応した製表方法に関する研究

１　研究の概要

製表業務の効率を維持しつつ統計ニーズの多様化に柔軟に対応できる製表方法について検討するため、国内外の研究動向等を把握するとともに、実証研究を行っている。

平成16年度においては、オーダーメード集計に適した既存ソフトウェアのリストアップ及び機能比較、オーダーメード集計や詳細な集計の増大に伴って開発の必要性が高まる秘匿処理法の検討を行った。

２　オーダーメード集計に関する研究

前年度実施した地方自治体におけるニーズ把握の結果、現状ではニーズの量的規模や傾向を特定することが困難であることが示唆された。このため、当面は、市販のものも含めて既存ソフトウェアの利用可能性について検討することとし、アドホックな集計に適する可能性のある集計用ソフトウェアのリストアップ及び機能比較を行った。その結果、市販ソフトウェアの中にかなり豊富な機能を有し、利用可能性があるものが見出されたが、いずれも秘匿処理の機能を欠いており、秘匿処理法について別途検討する必要のあることが明らかとなった。

３　集計表の秘匿処理に関する研究

小地域統計などの詳細な集計や企業統計などではプライバシー保護等の観点から集計表の秘匿処理が必要となる場合が多い。ニーズの多様化に伴い、詳細な集計が増大し、さらにオーダーメード集計が実施されると秘匿処理を効率的に行う必要性が高まるが、現在は効率的な秘匿処理法が開発されていない。

そこで、平成11年サービス業基本調査結果を用い、EUの秘匿処理に関するプロジェクトSDC及びCASCが開発した秘匿処理ソフトウェアτ-ARGUSの利用可能性に関してテストを行うとともに、米国センサス局の方法を採用したプロトタイプ・システムSCSを開発し、τ-ARGUSとの機能比較を行った。その結果、τ-ARGUSは実用面で機能が不十分であるのに対し、SCSは実用性に優れていると判断されたことから、業務への適用を目指してSCSの開発を継続することとした。

４　オーダーメード集計及び集計表の秘匿処理法に関する情報収集及び資料整備

集計表の秘匿処理法に関する研究を推進する上で、特に研究の盛んな欧米諸国の研究動向に関する情報収集が重要であることから、アメリカ・カナダ統計学会などが共同開催した「合同統計会議」（８月７日～14日、カナダ／トロント）に参加したほか、秘匿処理法に関する文献の収集・分析、統計データ機密保護に関する国連欧州経済委員会／EU統計局合同ワークセッション作成の統計データ開示抑制に関する用語集（製表技術関連資料集２に所収）の翻訳を行い、関係部門の参考に供した。

また、米国・カナダの人口センサスなどにおけるオーダーメード集計の実施状況及びそれに対応した秘匿処理法についても情報収集を行い、製表技術等研究報告会で報告するなど関係部門への周知を図った。

第４　情報処理技術に関する研究

１　プログラミング言語に関する研究

現在、主に使用しているプログラミング言語であるVisual Basicは、マイクロソフト社が提供しているＯＳであるWindowsにおいてのみ動作が可能であり、統計センターのＬＡＮシステムにおけるＯＳの選択肢を狭めている。

このため、ＯＳの選択肢を広げるとともに、ＬＡＮシステム切替え時に生じるプログラムの書き換え等の負担を軽減する目的から、機種やＯＳに依存しないプログラミング言語であるJavaについて調査・研究を行っている。

平成16年度は、研究・開発用ＬＡＮシステムを用いて、Javaによるテストプログラムを開発し、LinuxなどのWindows以外のＯＳでの動作検証を行い、動作することを確認した。

今後は、更に詳細な問題点を把握するため、製表システムの開発に試験的に使用することを予定している。

２　プログラミングの標準化等に関する研究

統計センターにおけるシステム開発業務については、開発者の経験と技量などによっては、近年の情報通信技術の進展に伴う技術の習得や開発環境変化への対応などが負担となることもある。このような状況は、システム開発効率の低下やシステム品質のバラツキを生じさせる要因ともなっている。

このため、開発者の経験と技量に依存する部分をできる限り縮小させることを目的に、プログ

ラミングに係る標準化等について調査・研究を行っている。

平成16年度は、ＬＡＮシステム切替えに伴うプログラミング言語のVisual Basic 6.0からVisual Basic .NETへの移行に係るコンピュータシステムガイド「プログラム開発基準編」の改訂を行った。

なお、「プログラム開発基準編」については、さらに改善に向けての検討を行っているところである。

また、総合テスト及び本番運用（再演算時の運用を含む。）に係るドキュメント及び業務手順の見直しを行い、これを踏まえコンピュータシステムガイド「運営管理基準編」の改訂を行っているところである。

このほか、既存システムをモデルとした開発業務の標準化、部品化等の検討を行うため、開発標準策定関連サービス[i]を試行的に導入した。

３　ナレッジマネジメントシステムに関する研究

事務の効率化、情報の共有化等を図ることを目的としたナレッジマネジメントシステムについて、民間企業における活用状況の聴取や新聞、雑誌等からの事例についての情報収集を行うとともに、統計センター内における適用可能な業務の洗い出しを行った。

その結果、各部門における情報共有化システムの運用状況を踏まえて「文書管理」、「職員情報共有化」についてシステム化を決定した。さらに、「業務ノウハウの蓄積・共有」について、市販ツール（グループウェア型）の試行体験を実施したが、使い勝手・効果などの面から導入には至らなかった。

以上のことから、単純な文書共有については、システム構築を進めるとともに、統計センターの業務のノウハウの共有化については、共有の在り方（機能、体制など）を更に検討する必要があるとの結論に達し、平成17年度の基盤整備活動に引き継ぐこととした。

第５　その他の研究等

１　製表技術参考資料等の刊行

製表技術の普及及び研究の促進を図るため、統計センターにおける製表技術の研究成果や、国内外における製表技術の研究動向の調査分析結果、製表業務のマネジメントを含めた主要な製表

---

[i] **開発標準策定関連サービス**：システム開発における開発プロセス、ドキュメント、設計方法などの標準化への支援をＩＴ関連企業が有料で行う技術サービスのことをいう。

技術関連文献の翻訳などの各種資料を刊行した。

平成16年度の刊行実績は、表のとおりである。

表　平成16年度製表技術参考資料等刊行実績

刊行年月	資料等名	内容
16．4	データエディティング研究会報告	前年度のデータエディティング研究会における年収欠測値の補定法の研究成果、エディティングの汎用的手法Fellegi-Holt法、新補定法NIMの欧米諸国における研究動向の調査分析結果など
16．5	製表技術関連資料集第１号	国連欧州経済委員会刊行の「データエディティングに関する用語集」及び「データエディティングの効率性評価：一般的枠組み」の翻訳
16．8	製表技術研究レポート第１号	多変量外れ値検出法に関する国内外の研究動向
16．8	製表技術参考資料第２号	平成15年度に行った事業所・企業統計調査産業分類の自動格付法の研究結果及び国内外における統計分類の自動格付法の研究動向
17．1	製表技術関連資料集第２号	統計データ機密保護に関する国連欧州経済委員会／EU統計局合同ワークセッション刊行の「統計データ開示抑制に関する用語集暫定版」及びEU統計局刊行の「統計品質に関する用語集」の翻訳
17．3	製表技術関連資料集第３号	EUにおける統計品質の定義、品質標準報告書、LEG勧告とその実施状況等をまとめた文献の翻訳
17．3	製表技術関連資料集第４号	職業・産業分類格付業務における継続的な品質管理、物価指数作成におけるISO9000の適用経験、生活行動分類の格付誤りの影響分析などの米英の文献の翻訳

２　学会における研究発表

平成16年度は、日本統計学会第72回大会（9月3～6日、富士大学（花巻市））において次の２件の研究発表を行った。

○Predictive Mean Matching法による年収欠測値の補定

○多変量外れ値検出法の比較