第５章　技術の研究に関する事項

統計センターでは、製表業務の高度化や製表結果の品質の向上、統計ニーズの多様化への対応などに資するため、製表実務に適用可能な研究に重点を置いて研究を進めている。平成22年度は、統計分類のオートコーディングシステムの研究、データエディティングに関する研究、統計ニーズの多様化に対応した製表技術に関する研究を行った。

第１節　オートコーディングシステムの研究

第１　ＯＣＲ機により認識されたデータを用いて直接産業大分類を格付する技術の研究

統計分類の格付業務について、調査票に記入された文字を外部委託により入力した後、オートコーディングを行う場合、文字入力に係る経費及び処理期間の両面において負担が大きく、オートコーディングシステムによる省力化の特性を十分に発揮できているとは言い難い。そこで、オートコーディングシステムによる更なる省力化の可能性を追求するため、ＯＣＲ機により国勢調査の調査票に記入された文字（イメージデータ）を認識し、その結果を用い、格付ルールによるオートコーディングを可能とする技術の研究を行っている。

外部委託により、①フリー記入式の「事業の内容」欄内への文字枠の設定の検討、②イメージデータに対する文字認識技術の検討、③文字認識から格付までのアルゴリズムの検討、の３点を主なテーマとして研究を開始し、その成果を踏まえ、実用化に向けた、更なる段階の研究に着手する予定である。

第２節　データエディティングに関する研究

第１　データエディティングの精度評価の研究

国勢調査等の大規模調査では、データチェックリストの審査に膨大な人員・時間を必要としている。当該審査の効率化を図るため、平成17年国勢調査第１次基本集計のデータの大都市を含む県を用いてテストを行い、その結果を「製表技術参考資料」に取りまとめた。

第２　多変量外れ値の検出方法の研究

調査票の未回答事項を補定する際、外れ値（特異値）は精度に大きな影響を与えるものである。そのため、外れ値を数学的理論により検出する方法を研究している。具体的には、企業財務データについて、繰り返し最小二乗法（IRLS）を適用する研究を行った。その結果については、統計関連学会連合大会及び平成22年度統計技術研究会（第１回）において報告した。

さらに、同じく企業財務データについて、乗率を考慮した産業別試算結果の詳細な分析を進めている。

第３　平成24年経済センサス - 活動調査のデータエディティング方法の研究

平成24年経済センサス - 活動調査では、経理項目が詳細に調査されることとなっている。その詳細さのため未回答が多い場合、結果精度に影響を与えることになる。これを改善するため、経理項目の補定方法（２次試験調査データを用いた回帰分析法等）を検証している。

第３節　統計ニーズの多様化に対応した製表技術に関する研究

第１　各種匿名化手法の研究

諸外国におけるデータ提供の趨勢に対応するため、匿名化手法等に関する諸外国の先行研究の情報収集及び文献の翻訳等を実施し、「Handbook on Statistical Disclosure Control」について製表技術関連資料集の原稿を作成した。

また、平成23年度から提供を予定している労働力調査の匿名データの作成方法について、総務省統計局との共同研究を実施した。

第２　各種匿名化技法による有用性と秘匿性の評価方法に関する研究

匿名化技法の違いが匿名データの有用性と秘匿性に与える影響の評価方法に関し、定量的な分析に基づく相対的評価方法について、諸外国における先行研究の調査を行うとともに、平成16年全国消費実態調査の個票データから、各種匿名化手法及びミクロアグリケーション手法で作成したデータを用いて、確率的リンケージ手法[17]やＲ－Ｕマップ[18]やなどによる実証的研究を行った。

第３　擬似データ作成に関する研究

統計調査の公表済み集計結果表から匿名化データを作成する各種方法論を踏まえ、平成16年全国消費実態調査データの各項目を高次元にクロス集計した集計表をベースとして個別データに近い分布と特性を持つ擬似的なデータを作成し、教育・訓練用データとして提供するための研究を行っている。

第４節　情報収集、技術協力等

第１　外部研究者の採用及び統計センター内研究会での外部研究者の活用

統計学の研究に携わっている博士研究員や大学教育初任段階の若手研究者を非常勤研究員として採用し、データエディティングの精度の評価の研究、匿名データの秘匿性の評価方法などの研究を行った。また、大学教授等の外部研究者で構成する「統計技術研究会」を１回開催するとともに、外部有識者を講師に招いた「統計技術研究会講演会」を開催した。

第２　データエディティング等の研究動向に関する情報収集

データエディティング及びデータ秘匿に関する研究を推進する上で、研究動向に関する情報収集が重要であることから、東京都文京区で開催された「日本人口学会第62回大会」及びギリシャ共和国のケルキラで開催された「Privacy in Statistical Databases 2010（2010年　統計データベースにおけるプライバシーに関する会議）」に参加するとともに、諸外国における人口センサスの匿名データ作成に係る実状把握のため、オーストラリア統計局及びニュージーランド統計局に赴いた。

---

[17] **確率的リンケージ手法**：秘匿処理済データの秘匿性の評価に使われる手法の一つ。秘匿前の原データと秘匿処理済データのペア（組合せ）がリンクされる（一致）ペアかリンクされない（不一致）ペアのどちらに属すると判定するかについて属性値の一致基準及び確率値に従って分類する方法。

[18] **Ｒ－Ｕマップ(Risk-Utility Confidentiality map)**：秘匿処理済データの秘匿性と有用性の関係を表すもの。例えば、縦軸に情報量損失率、横軸に度数１の減少率をプロットすること（散布図）により関係を視覚的に見ることができる。

第５節　研究成果の普及等

第１　統計技術及び研究成果の普及等

１　統計技術研究会

平成22年度 統計技術研究会及び講演会開催実績

回	開催年月日	議　　題
第１回	22.12.22	・企業財務データを用いた売上高のロバスト回帰による補定 ・2010年統計データベースにおけるプライバシーに関する会議（ＰＳＤ2010）出張報告
講演会	23. 1.27	・統計調査における欠測データの補正に関する研究動向

２　統計センター実務検討会

統計センター業務についての研究・開発の成果及び事務改善に関する情報等を共有し、その活用を一体的かつ効果的に推進するとともに、職員の人材育成及び専門性の継承を図るため、統計センター実務検討会を10回開催した。

３　製表技術参考資料等の刊行

研究成果の普及を図るため、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などの資料を３冊刊行した。

４　学会等における研究発表

平成22年度 学会等における研究発表実績

年月日	会議等の名称	発表内容	開催地	開催場所
22. 4.10	経済統計学会関東支部2010年４月定例研究会	・教育用データの作成について	東京都 千代田区	法政大学 市ヶ谷ｷｬﾝﾊﾟｽ
22. 7. 3	「調査データベース公有化における個人データ保護の統計理論」合同研究会	・ミクロデータにおける有用性と秘匿性の評価方法について	東京都 立川市	統計数理研究所
22. 9. 5 ～ 9. 8	2010年度統計関連学会連合大会	・ミクロデータにおける有用性と秘匿性の定量的な評価の試み ・教育用ミクロデータの作成方法について ・多変量外れ値検出法の実データへの適用について －企業売上高のロバスト回帰による補定－	東京都 新宿区	早稲田大学 早稲田ｷｬﾝﾊﾟｽ
22. 9.16 ～ 9.17	経済統計学会　2010年度（第54回）全国研究大会	・教育用ミクロデータ作成の試み －政府統計ミクロデータの利用拡大に向けて－	大分県 大分市	大分大学 旦野原ｷｬﾝﾊﾟｽ
22. 9.22 ～ 9.25	日本行動計量学会第38回大会	・高等教育における公的統計の２次利用の枠組み ～経済・社会科学における実証分析力の育成	埼玉県 さいたま市	埼玉大学
22.10.29 ～10.30	研究集会「官庁統計データの公開における諸問題の研究と他分野への応用」	・教育用ミクロデータの作成 ・ミクロデータにおける有用性と秘匿性の定量的な評価に関する研究	東京都 立川市	統計数理研究所