

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21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 平成22年国勢調査に係る実用化に向けた研究

1 格付テスト及び検証結果

平成17年国勢調査の抽出速報集計用の文字入力済データを用いて、ルールベース型のオートコーディングシステムのテスト及び結果の検証を行った。格付テストに当たっては、新分類符号への組替え、ルールの整理を繰り返し、格付率及び正解率はそれぞれ産業分類で40.2%、97%以上、職業分類で44.2%、97%以上となり、両分類とも目標とする格付率40%を達成した。

2 外部委託による機械学習型格付システムの開発

外部委託で行った研究成果を分析・検討した結果、機械学習型格付システムの有効性を確認できたことから、同システムの開発を外部委託で行った。同システムの基本機能として①ルール作成機能、②自動格付機能、③格付結果分析機能を備えたものとし、格付率40%、正解率97%を目標としている。

第2 平成24年経済センサスー活動調査に係る実用化に向けた研究

1 研究・開発方針の決定

現行の事業所・企業に関するオートコーディングシステムは、導入されて10年余が経過し、現在のLAN環境などへの適合性の低下による及び度重なる手直しによるシステム更新業務の負担増となっている。このため、現行のハードウェアに適合し、更なる精度を持つシステムを新たに開発することを「経済センサスー活動調査 製表検討会議」で決定した。

2 外部委託による機械学習型格付システムの研究・開発

新システムは、平成21年度に外部委託で開発した国勢調査用のオートコーディングシステムの格付手法を参考に21年度から22年度に掛けて研究・開発を行うこととし、システムの構築は22年度末までに完了することとした。新システムの格付性能は、格付率55%以上、正解率は97%以上を目標としている。

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

セレクトティブエディティング¹⁴に関する諸外国の情報の収集に加え、平成17年国勢調査の速報

¹⁴ セレクトティブエディティング：エラーデータについて、集計値への影響度やエラーの度合いを所定の算式によって推定（スコア化）し、その大きさが一定値以上のものについて人手による審査を重点化（それ以外はコンピュータにより補正処理）

集計及び第1次基本集計のデータを用いて、データチェックリスト審査の必要度を予測式により算出する、当該エディティングの考え方を導入した手法の有効性等について検討を行っている。

第2 多変量外れ値の検出方法の研究

カナダ統計局において実用化されたMSD法¹⁵及びその改良手法について、統計解析ソフトウェアRによりプログラムを作成し、シミュレーションデータによる性能評価を行っている。

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 諸外国のデータ提供方法の調査・研究

匿名データ及びオーダーメイド集計表の適切かつ円滑な作成・提供を行う方法の研究のため、諸外国におけるリモート集計等のオンライン方式による提供方法の調査を行った。

また、マイクログリゲーションによる匿名化手法の事業所・企業データへの諸外国での適用例の調査・研究を行った。

第2 匿名データ作成における各種匿名化技法の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ついて、定量的な分析に基づく客観的な評価のための指標の作成が可能か、平成16年全国消費実態調査について、マイクログリゲーション手法で作成したデータを用いた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、その派生的な方法として、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作成した。これ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、学生、研究者等に教育・訓練用データとして提供すること及び二次利用におけるサテライト機関でのオーダーメイド集計の際の利用者用サンプル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

外部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2回開催した。

第2 情報収集

1 データエディティングに関する情報収集

データエディティング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

する方法。

¹⁵ MSD法 (Modified Stahel-Donoho 法) : 多変量データを様々な角度から射影することにより、分布の外側に位置するデータを見つけ出し、外れ値として検出する手法。Stahel(1981)及びDonoho(1982)が提案したSD法をPatak(1990)が改良した。

ことから、スウェーデンのストックホルムで開催された「Modernisation of Statistics Production（統計作成の近代化に関する会議）」及び東京都立川市で開催された研究会「データ解析環境Rの整備と利用」に参加した。

2 データ秘匿に関する情報収集

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、スペインのビルバオで開催された「Work Session on Statistical Data Confidentiality（統計データの機密保護に関するワークショップ）」等、3会議に参加した。

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21年度 統計技術研究会開催実績

回	開催年月日	議 題
第1回	21. 11. 13	<ul style="list-style-type: none"> ・擬似データの作成について（中間報告） ・各種匿名化技法における有用性と秘匿性の評価方法について ・多変量外れ値の検出法－MSD法とその改良版について－
第2回	22. 3. 16	<ul style="list-style-type: none"> ・擬似データの作成方法について ・諸外国における統計データの機密保護

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4冊刊行した。

4 学会等における研究発表

平成21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
21. 7. 4	経済統計学会関東支部定例研究会	・マイクロデータにおける秘匿性の評価方法に関する一考察	東京都千代田区	専修大学 神田キャンパス
21. 9. 6 ～ 9. 9	2009年度統計関連学会連合大会（日本統計学会第77回大会）	・マイクロアグリゲーションにおける秘匿性の評価方法について	京都府京都市京田辺市	同志社大学 今出川キャンパス 京田辺キャンパス
21. 9. 18	「公的統計の有効な提供方法に関する研究」第1回研究会	・マイクロデータにおける秘匿性の評価方法に関する一考察	東京都渋谷区	青山学院大学 青山キャンパス
21. 10. 22 ～10. 23	研究会「官庁統計データの公開における諸問題の研究」	・マイクロデータにおける秘匿性の評価方法に関する一考察	東京都立川市	統計数理研究所

平成21年度 学術誌等への投稿実績

投稿年月	学術誌等名	内 容
21. 12	統計研究彙報第67号	・多変量外れ値の検出－MSD法とその改良手法について－