

第4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20年度は、統計分類のオートコーディング処理、統計データの欠測値の補定処理、匿名データの作成方法などの研究を行った。

第1節 オートコーディングシステムの研究

第1 経済センサスー基礎調査に係る研究

1 過去の研究成果の実務への適用

平成21年経済センサスー基礎調査における産業分類のオートコーディングシステムは、以前研究・開発した機械学習型¹システムを用いることとし、これを実務へ適用するため、①平成18年事業所・企業統計調査の約50万データを基にしたオートコーディングルール作成用データの整備、②2種類の格付結果を比較し、精度の高い方法を採用する合成方式の導入、③キーワード修正・除外機能の追加を行い、格付率及び正解率を向上させた。

2 格付率及び正解率

事業所分類の格付率は65.4%、正解率は97.7%、企業分類の格付率は74.0%、正解率は96.2%まで向上した。

第2 平成21年全国消費実態調査及び平成22年国勢調査に係る実用化に向けた研究

1 平成21年全国消費実態調査に係る実用化に向けた研究について

オートコーディングシステムの構築に向け、開発方針を決定し、平成19年度に行った民間委託による研究成果の一部を取り入れてシステム開発を進めた。

当初、収支項目分類の格付率は26.1%、正解率は98.4%であったが、研究・開発の結果、平成19年度家計調査データで格付率55.2%、正解率98.7%、平成16年全国消費実態調査データで格付率58.1%、正解率99.3%まで向上した。

2 平成22年国勢調査に係る実用化に向けた研究について

オートコーディングシステムの構築に向け、開発方針を決定するとともに、産業分類及び職業分類の格付テストを行い、格付結果について検証した。

第2節 データエディティングに関する研究

第1 チェック・補定の精度評価の研究

セレクトティブエディティング²を中心とした諸外国の情報を収集するとともに、製表におけるデ

¹機械学習型: 人手による格付結果を学習用データとして、形態素解析など機械処理を行い、確率的に高い順に格付を行うルールを自動的に作成する。

²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定処理)する方法。

ータ処理方法等を調査し、データエディティング方法について精度評価の手法の研究を進めた。

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³及びその改良手法について、統計ソフトウェア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た。

第3節 匿名データの作成方法の研究

第1 秘匿方法及び秘匿処理ソフトに関する研究

オランダ統計局が開発した秘匿処理ソフトウェアである匿名データ作成用の μ -ARGUS、結果表の二次秘匿処理用の τ -ARGUSについて、最新バージョンを使用した機能等の検証作業をサンプルデータ等を用いて進めていたが、平成21年4月から提供する統計調査データの匿名化に際して、 μ -ARGUSの機能では実現できない手法があったことから、プログラムを独自開発することとした。また、結果表の二次秘匿処理に関して、 τ -ARGUSの最新バージョンでも処理機能に制約があることから、その適用を見送った。

今後も、秘匿処理用のソフトウェアについては、開発の動向等の情報収集を進め、必要に応じ適用の可能性について検討する。

第2 ミクロアグリゲーションの有効性に関する研究

平成19年度に行った匿名化技法としてのミクロアグリゲーションの有効性に関する研究結果を踏まえ、平成16年全国消費実態調査の個別データを用いて、国内外の先行研究を基に評価を試みた。また、全国消費実態調査の個別データを用いた研究に引き続き、他調査（平成15年住宅・土地統計調査及び平成18年事業所・企業統計調査）の個別データを用いた研究を行っている。

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への大学教授等外部研究者の活用

外部の研究者を非常勤研究員として採用し、調査票情報の秘匿技法の一種であるミクロアグリゲーションに関する研究を行った。また、大学教授等外部から研究者を招へいした「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、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、オーストリアのウィーンで開催された「Work Session on Statistical Data Editing（統計データエディティングに関するワークショップ）」等、3会議に参加した。

³MSD法：共分散行列と平均値ベクトルを用いて外れ値を検出する、Patak(1990)が提案した方法。

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、特に研究の盛んな国内外の研究動向に関する情報収集が重要であることから、タイのバンコクで開催された「Workshop on Managing Disclosure Risk and Anonymization Tools for Microdata Dissemination (マイクロデータ普及のための開示リスク管理と匿名化ツール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20年度 統計技術研究会開催実績

回数	開催年月日	議 題
第1回	H20. 12. 19	・多変量外れ値の検出について ・複数の公表された集計表からの擬似個別データ作成の試み ・ヨーロッパにおけるデータエディティングと補定に関する取組み動向 ・「統計データ・エディティング (vol. 3) データ品質への影響」の内容紹介
第2回	H21. 3. 16	・住宅・土地統計調査によるマイクロアグリゲーションの試み ・マイクロアグリゲーションにおける開示リスクの評価方法について ・データエディティング及び補定に関する品質評価指標

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、製表業務のマネジメントを含めた製表技術関連文献の翻訳などの資料を4冊(年度目標3冊以上)刊行した。

4 学会等における研究発表

平成20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
H20. 9. 7 ～ 9. 10	2008年度統計関連学会連合大会 (日本統計学会第76回大会)	全国消費実態調査を用いたマイクロアグリゲーションの試み	神奈川県 横浜市港北区	慶応義塾大学 矢上キャンパス
H20. 10. 23 ～ 10. 24	研究集会「官庁統計データの公開における諸問題の研究」	・マイクロアグリゲーションに関する研究動向	東京都 港区	統計数理研究所
H20. 11. 8	経済統計学会関東支部定例研究会	・全国消費実態調査を用いたマイクロアグリゲーションの有効性の検証	東京都 豊島区	立教大学 太刀川記念館

平成20年度 学術誌等への投稿実績

投稿年月	学術誌等名	内 容
H21. 3	統計研究彙報第66号	・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～