

## 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23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。

平成22年度から外部委託により、「事業の内容」欄を対象とした産業大分類の格付について①フリー記入欄内の文字枠の設定の検討、②OCR機による文字認識技術の検討、③文字認識から知識処理、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行っている。平成23年度は、「勤め先・業主などの名称」、「事業の内容」及び「本人の仕事の内容」の3欄を対象として産業・職業分類自動格付を行い、実践的なレベルのテストの実施による研究に着手しており、24年度にも平成27年国勢調査第1次試験調査結果を対象とした研究を継続する予定である。

### 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。同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検証を行い、次回国勢調査への実用に向けてプロジェクトにおいて検討を進めており、平成22年国勢調査データを用いた検証方法を検討中である。

#### 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。平成23年度は、シミュレーションにより様々なデータ条件下での繰り返し最小二乗法（IRLS）の挙動を整理した。

また、企業財務データを用いて売上高の従業者数による補定の試算とともに補定により得られるIRLSウェイトにより標本調査データの乗率調整を試みた。IRLSは外れ値の影響を緩和または除外することができるので、ごく少数の特異なデータの動きによる変動を抑えた補定値が得られ、乗率の補正にも役立つ可能性を持つ一方、高度な統計ソフトや科学技術計算ライブラリなどがなくても実用化が可能である。さらに、サービス産業動向調査の売上高補定への実用可能性を確認するため、調査データを用いた試算を行っている。

なお、シミュレーションにより様々なデータ条件下での繰り返し最小二乗法（IRLS）の挙動を整理した結果については、応用統計学会年会で報告した。企業財務データを用いて売上高の従業者数による補定の試算とともに補定により得られるIRLSウェイトにより標本調査データの乗率調整を試みた結果については、国際統計協会（ISI）第58回大会及び統計関連学会連合大会において報告した。

### **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、2次試験調査データにより、ホットデック法の一手法として経理項目の補定方法について研究を行い、比率による補定が真値との乖離が少ないとの結果を得た。

## **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、製表技術関連資料集を刊行した。

また、労働力調査の匿名データ作成について、総務省統計局と共同研究を行い、統計委員会の了承を得て、提供を開始した。さらに、平成24年度中に提供予定の国勢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21年家計調査を用いて、かく乱的匿名化手法を適用したデータの作成及びこのデータに対する有用性と秘匿性の定量的評価による検証を実施した。その結果、秘匿化されたマイクロデータの秘匿性と有用性の程度の定量的、客観的な評価の可能性を確認した。

### **第3 擬似データ作成に関する研究**

公的統計のマイクロデータの利用を推進するため、大学などで統計演習に利用可能とする「教育用擬似マイクロデータ」の作成について、統計委員会等の意見・要望があった。これを受けて、統計センターでは、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として、教育用擬似マイクロデータを作成した。

平成23年度は、教育用擬似マイクロデータの実用性の検証・評価等を行うため、8月から、主に大学、研究機関等の教育関係者からの利用を想定した一般への試行的な提供を実施した。実施に当たり、統計センターホームページに提供方法について紹介したほか、各種学会等において紹介した結果、45件の利用申出があり、大学等での授業、統計演習等において、約400名の院生・学生による利用があった。

また、今回の試行提供に当たりアンケートを、利用者に対して実施した。

今後は、アンケート結果を踏まえ、実用性の検証・評価等を行い、実用化に向け検討する。

## 第4節 情報収集、外部機関との連携等

### 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若手研究者を非常勤研究員及び任期付研究員（上級研究員）として採用し、データエディティングの効率化の研究、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 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、京都市で開催された「日本人口学会第63回大会」及びスペイン王国のタラゴナで開催された「Joint Work Session in Statistical Data Confidentiality（統計データの機密保護に関する合同ワークセッション）」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、英国国家統計局においてヒアリングを行った。

### 第3 学会との連携等

統計センターでは、日本統計学会等、統計技術との関連が強い学会に団体加入し、学会の情報を入手するとともに、定期的で開催される会合において発表を行っている。また、匿名データの提供及びオーダーメイド集計などの統計センターの業務を広く知らしめるための広報も実施している。

## 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

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の研究成果を報告し、外部研究者を始めとした有識者から意見等をいただき、それぞれの研究の方向性や手法を検討するため、統計技術研究会を2回開催した。

平成23年度 統計技術研究会開催実績

回	開催年月日	議 題
第1回	23. 12. 26	・経済センサスー活動調査における経理項目の補定方法について ・英国における政府統計マイクロデータの作成状況について
第2回	24. 3. 13	・教育用擬似マイクロデータの作成と試行提供状況 ・統計データの機密保護に関する合同ワークセッション（出張報告）

#### 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。

### 平成23年度 統計センター実務検討会開催実績

開催年月日	テーマ
23. 6. 8	・平成24年経済センサスー活動調査における経理項目の補定方法の比較・検討
23. 7. 13	・平成22年国勢調査 産業・職業分類自動格付について ・収支項目分類における自動格付の課題へのアプローチ
23. 8. 3	・平成22年国勢調査 抽出速報集計を終えて ・震災時の職員の安全確保について
23. 8. 31	・セミオーダーメイド集計のシステム（プロトタイプ）について
23. 9. 21	・平成24年経済センサスー活動調査の製表について ・OCRの課題と今後の対応について
23.10.12	・統計データ二次利用の申出状況について
23.11. 9	・確率による審査対象の選択 ～平成17年国勢調査での検証結果報告～
23.12.14	・平成22年国勢調査 高齢者重複チェックシステムの概要 ・新規採用職員及び経験1年の者の育成 ・共同利用システム監視室の入退室セキュリティ対策
24. 1. 18	・マクログラフとマンダラに基づく可視化審査による分析的審査の効率化 ・H22国勢調査 抽出速報集計 産業・職業分類符号格付事務 Gラーニング ・プログラム開発等業務への派遣職員の活用について（中間報告）
24. 2. 8	・OCR機を利用した国勢調査の産業・職業大分類の自動格付の研究 ・職員の健康管理について ～生活習慣病から身を守るために～

### 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、国内外における製表技術の研究動向の調査分析結果の資料を刊行した。

#### 平成23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
24. 3	製表技術関連資料集10	欧州統計機構ネットワークによる統計的開示抑制 ESSNet SDC

### 4 学会等における研究発表

#### 平成23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
23. 6. 4	応用統計学会2011年度大会	・ロバスト回帰を用いた外れ値に対する乗率の補正	大阪府 吹田市	大阪大学 吹田キャンパス
23. 8. 21 ～ 8. 26	2011年 国際統計協会第58回大会 (ISI 2011)	・ Multivariate Outlier Detection for Regression - Imputation and Aggregation Weight Calibration by ISLS - ・ A Method to Quantitatively Assess Confidentiality and Potential Usage of Official Microdata in Japan	アイルランド ダブリン	ダブリンコンベンションセンター
23. 9. 4 ～ 9. 7	2011年度統計関連学会連合大会	・教育用擬似マイクロデータの試行提供について ・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と集計乗率の補正- ・マイクロデータにおける攪乱的手法の有効性の検証	福岡県 福岡市	九州大学 伊都キャンパス

23. 9. 14 ～ 9. 15	経済統計学会第55回（2011年度）全国研究大会	・教育用マイクロデータの作成方法と 試行提供について	東京都 八王子市	中央大学 多摩キャンパス
23. 10. 21 ～10. 22	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」	・教育用擬似マイクロデータ（全国消費実態調査）の作成方法と試行提供 ・公的統計の二次利用（匿名データ、オーダーメード集計）の現状 ・家計調査マイクロデータを用いた攪乱的手法の有効性に関する研究	東京都 立川市	統計数理研究所
23. 11. 11	平成23年度 公的統計のマイクロデータの利用に関する研究集会	・教育用擬似マイクロデータの作成と試行提供ー公的統計のマイクロデータ利用の推進に向けた取組ー	東京都 立川市	統計数理研究所
24. 3. 2 ～ 3. 3	第8回統計教育の方法論ワークショップ	・教育用擬似マイクロデータの開発とその提供ー統計教育の教材として利用するためにー	東京都 国立市	一橋大学 国立キャンパス

## 第6節 技術協力の実施

### 第1 海外への技術協力

平成23年度は、15年度から始まったカンボジア統計局への技術支援のため、関係機関（日本政府及び国際協力機構）からの要請に応じ、3回にわたって専門職員の派遣を行った。カンボジアでは、日本政府及び国際協力機構の支援の下、2011年3月に、全国の事業所の所在地、従業員数、売り上げ等を全数調査する経済センサスをカンボジア史上はじめて実施し、2012年3月に確報結果が公表された。

### 第2 国の行政機関及び地方公共団体への技術協力

製表業務の技術協力の一環として、総務省統計局が主催する地方別事務打合せ会、実務研修会等に対して、同局と連携しながら職員の派遣を行った。

表 平成23年度の主な派遣実績

業務内容	人数
平成23年度労働力調査 実務研修会	1
平成23年度労働力調査 調査員合同指導会	1
平成23年度家計調査 実務研修会	1
平成24年経済センサス - 活動調査 地方別事務打合せ会	4

## 第7節 次世代統計利用システムの研究開発

統計センターでは、公的統計の利用可能性の拡充及び統計利用者の利便性向上を図るため、公的統計に係る公表情報その他統計センターが蓄積し、又は独自に作成・収集する各種の統計情報を、インターネット環境とGIS<sup>16</sup>等のアプリケーション、API<sup>17</sup>を通じて、受益者負担（有償）

<sup>16</sup> GIS : Geographic Information System。地図に関する属性情報をコンピュータにより解析するシステム。

<sup>17</sup> API : Application Programming Interface。OSなどの基本ソフトウェアが開発ツールに対し提供している機能を開発ツール（アプリケーション・ソフトウェア）から呼び出す際の取り決め、仕様のこと。

により提供することを目的とした次世代統計利用システム開発計画構想（GAUSSプロジェクト）を平成23年4月に策定した。

次世代統計利用システムの構築に当たっては、総務省情報流通行政局と連携して、独立行政法人情報通信研究機構が管理・運用している新世代通信網テストベッド（JGN-X）を活用し、平成23年12月からAPI機能による実証実験を開始した。GIS機能及びセルデータ集計機能の実証実験については、平成24年4月以降開始する予定（5月に開始した）。