

## 第5章 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25年度は、統計分類の格付支援システムの研究、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 格付支援システムの研究

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

格付支援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、格付ルールによる格付支援を可能とする技術の研究を行っている。

平成25年度は、現行OCR機を用いて、平成27年国勢調査第1次及び第2次試験調査結果のうち、文字枠が設定された「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欄を対象として文字認識及び知識処理の研究を行った。

具体的には、平成24年度までに蓄積した文字認識に係る知識処理の技法を、新たな認識環境の下でテストを行い、認識結果を検証することにより技法を高めた。また、平成27年国勢調査に向け、誤認識に起因する誤格付を避けるため、格付支援システムによる格付後の格付判定処理を検討し、その仕様をまとめた。

### 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効率化手法及び実用化への研究

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22年国勢調査人口等基本集計のデータを用いて検証を行い、次回国勢調査での実用に向けて「データエディティング効率化手法検討プロジェクト」において検討を行いつつ、研究を進めている。

平成25年度は、これまで行ってきたデータチェック審査の効率化に関する研究成果を基に、エラーとなったデータの特性をロジスティック回帰分析による統計的手法を用いてデータチェック要領改善の指標を求めるシステムを考案した。今後は、事務担当者を中心としたワーキンググループにおいて、平成27年国勢調査への適用に向けた検討を開始する予定である。

#### 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について研究を行っている。

平成25年度は、家計調査における収入金額及び支出金額のレンジチェック（外れ値の検出）への適用について検証を行ったほか、多変量外れ値検出手法の一つであるMSD(modified

Stahel-Donoho)法<sup>19</sup>のコンピューターによる計算負荷軽減のための研究を行い、並列化処理が有効であるとの成果を得た。その研究成果を I E E E（米国電気電子技術者学会）が主催する国際会議(2013 International Conference on Cloud Computing and Big Data)において報告した。

また、家計調査における収入金額及び支出金額のレンジチェック(外れ値の検出)の新たな手法の研究に着手し、26年度中に新たな手法の提案ができるよう、多変量外れ値検出手法の適用などについての研究を行った。

### **第3 事業所・企業調査におけるデータエディティング方法の研究**

事業所・企業を対象とする調査における経理項目に対するデータエディティングの方法について研究を行っている。平成24年度には、欠測値補定における多重代入法の有用性を検証した。

平成25年度は、多重代入法を用いた補定方法について24年度に引き続き実務に適用可能なアルゴリズムに関して調査データなどを用いて調査・研究を行った。その研究成果について統計関連学会連合大会、経済統計学会、科学研究費シンポジウムで報告したほか、「統計研究彙報」に論文を投稿した。

また、諸外国の情報を収集し、「製表技術参考資料23」として「諸外国における最新のデータエディティング事情～混淆正規分布モデルによる多変量外れ値検出法の検証～」を刊行するなど、製表実務での実用化に向けた研究を引き続き行っている。

## **第3節 統計データの提供に関する研究**

### **第1 統計情報の提供方法の強化に関する研究**

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次世代統計利用システム開発計画構想プロジェクトを平成23年4月に立ち上げ研究開発を進めている。

平成25年度は、API機能、GIS機能について一般利用者への試行提供をそれぞれ開始した(第4章第3節第2参照)。セルデータ集計機能は引き続き実証実験を継続して行った。

### **第2 擬似マイクロデータ作成及び利活用手法に関する研究**

匿名データの提供等統計の二次的利用を推進するため、大学などで統計演習等に利用可能な「擬似マイクロデータ<sup>20</sup>」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、平成16年全国消費実態調査データを用いた量的変数の擬似マイクロデータを平成23年度より試行提供を開始した。平成25年度は38件(前年度29件)の利用申出があり、利用者の意見・要望を踏まえた作成方法及び利活用手法の検討を継続中である。

平成24年度より継続している平成14年及び19年就業構造基本調査データを用いた質的変数の擬似マイクロデータの作成方法の研究については、検証結果を統計技術研究会において報告した。

また、中国マカオで開催された「国際統計教育協会(IASE)/国際公的統計協会(IAOS)合同サテライト会合」において、擬似マイクロデータの開発に係る研究成果の報告を行った。

<sup>19</sup> MSD法：平均値ベクトルと分散共分散行列で決まる楕円分布データを対称とする多変量外れ値検出法

<sup>20</sup> 擬似マイクロデータ：統計調査を特別集計した統計表から乱数を発生させて生成した擬似的なデータ。

### **第3 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を行った。

また、統計局との共同研究により、平成25年度は社会生活基本調査B票の匿名データの作成方法を検討した。平成24年度に格付を行った国勢調査の匿名データについては、統計委員会における諮問・答申を経て、平成25年12月から提供が開始された。

### **第4 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、平成25年度は、諸外国における先行研究の調査を行うとともに、平成17年国勢調査のデータを用いて、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、平成24年度に引き続き実施した。その研究成果については、「統計的機密保護に関するUNECEとEurostatの合同ワークショップ」、「統計関連学会連合大会」及び「経済統計学会全国研究大会」等の国内外の研究集会において報告した。

## **第4節 外部機関との連携及び研究成果の普及等**

### **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、匿名データの有用性と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、札幌市立大学（北海道札幌市）で開催された「日本人口学会第65回大会」に参加し、国勢調査の匿名データを使用した分析等の発表を聴取し匿名データの利用について情報収集を行った。さらに、オタワ（カナダ）で開催された「統計データの機密保護に関するUNECEとEurostatの合同ワークショップ」に参加した際に、カナダ統計局を訪問し、担当者から匿名化手法等に関する情報収集を行った。

### **第3 統計技術及び研究成果の普及等**

#### **1 学会等における研究発表**

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4学会に団体加入し、学会の情報入手するとともに、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。25年度は、以下の発表を行った。

平成25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
25. 5. 24 ～5. 25	応用統計学会2013年度年会	・小地域人口統計データの活用とその未来（株式会社NTTドコモとの共同研究）	福島県 福島市	パルセいいざか
25. 7. 6	経済統計学会関東支部 2013年7月例会	・公的統計のデータエディティング：混淆正規分布モデル及び多重代入法の適用可能性	東京都 豊島区	立教大学 池袋キャンパス
25. 8. 22 ～8. 24	国際統計教育協会 (IASE)/国際公的統計協会 (IAOS) 合同サテライト会合	・Statistics education initiatives by Japanese Official Statistics （邦題：日本の官庁統計機関による統計教育イニシアティブ） ・Development of Synthetic Microdata for Educational Use in Japan （邦題：教育用擬似マイクロデータの開発について）	中国 マカオ	
25. 8. 25 ～8. 30	第59回国際統計協会 (ISI) 世界統計大会	・A Post-Aggregation Error Record Extraction Based on Naive Bayes for Statistical Survey Enumeration （邦題：統計調査集計表のためのナイーブベイズに基づいた集計後のエラーレコードの抽出） ・Multiple Imputation of Missing Values in Economic Surveys: Comparison of Competing Algorithms （邦題：経済調査における欠測値の多重代入－様々なアルゴリズムの比較－） ・The Best Stratification to Impute Missing Values of Turnover in Economic Surveys （邦題：経済調査に係る売上高の欠測値補定にふさわしい層分け）	中国 香港	
25. 9. 8 ～9. 11	統計関連学会連合大会	・様々な多重代入法アルゴリズムの比較 ・定性データに基づく企業構造の境界値分析 ・経理項目に基づく企業の構造化分析 －構造の境界値検定－ ・統計データの二次的利用に関する統計センターの取組状況 ・マイクロデータにおけるスワッピングの適用可能性の検証 ・統計におけるオープンデータの高度化	大阪府 豊中市	大阪大学豊中キャンパス
25. 9. 13 ～9. 14	経済統計学会2013年度全国研究大会	・マイクロデータにおける匿名化の誤差の検証 －国勢調査の個票データを用いて－ ・経済調査における売上高の欠測値補定 －様々な多重代入法アルゴリズムの比較－	静岡県 静岡市	静岡市産学交流センター
25. 10. 7 ～10. 9	国際連合・メキシコ国家統計地理情報局 国際地理学・環境会議 (CIGMA 2013) メキシコ国家統計. 地理. 情報局	・Small area statistics on de jure and de facto populations - population census and operational data of mobile phone （邦題：常住地人口と現在地人口の小地域統計 - 国勢調査と携帯電話運用データを基に）	メキシコ メキシコシティ	
25. 10. 28 ～10. 30	統計データの機密保護に関するUNECE & Eurostat 合同ワークショップ	・Assessing the Effectiveness of Disclosure Limitation Methods for Census Microdata in Japan	カナダ オタワ	カナダ統計局

年月日	会議等の名称	発表内容	開催地	開催場所
25. 12. 16 ～12. 19	2013 International Conference on Cloud Computing and Big Data (CloudCom-Asia)	・Parallel computing of the modified Stahel-Donoho estimators for multivariate outlier detection (邦題: 多変量外れ値検出のためのMSD推定量の並列処理)	中国 福州	

## 2 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、外部有識者から統計を取り巻く動向について話を伺うと共に、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25年度 統計技術研究会開催実績

回	開催年月日	議 題
第1回	26. 1. 14	《講演会》 「データに基づく意思決定」 統計数理研究所副所長 丸山 宏 教授 株式会社Preferred Infrastructure 渡部 創史 氏 海野 裕也 氏
第2回	26. 3. 26	・就業構造基本調査を用いた擬似マイクロデータの作成について ・MSD法の並列化

## 3 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25年度 統計センター実務検討会開催実績

回	開催年月日	発表内容
第51回	25. 6. 12	・OCR機による文字認識結果を利用した格付支援システムの研究
第52回	25. 7. 10	・大規模な災害からの情報システムの復旧について
第53回	25. 7. 31	・労働力調査製表システムへの疑義票システムの導入 ・格付支援システムによる産業・職業分類符号格付業務の効率化 －平成22年国勢調査について－ ・審査対象特定システムによる外れ値の検出 ー分析的審査における均一審査の導入ー
第54回	25. 9. 11	・平成24年就業構造基本調査 結果表審査を終えて ・小地域人口統計データの活用とその未来
第55回	25. 10. 24	・平成25年 住宅・土地統計調査の製表について ー住まいから 描く日本の 未来地図ー ・製表グループ進行管理担当の業務について
第56回	25. 11. 13	・OCR機による漢字等認識について ・カンボジア政府統計の作成支援について ー統計センターが取り組む国際協力の一例ー

回	開催年月日	発表内容
第57回	25. 12. 11	・ペーパーレス化への取組 –タブレット端末の試行的導入– ・バーチャルフォーラムいつ使うの? –これからでしょ!–
第58回	26. 1. 22	・統計センターにおけるISMSの現状について
第59回	26. 2. 19	・統計技術研究課における研究業務の紹介 –データエディティングに関する研究を例に– ・初めてのR
第60回	26. 3. 12	・統計センター情報システム基盤について

注) 回は平成20年度からの一連番号

#### 4 製表技術参考資料等の刊行及び学術誌等への投稿

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外における製表技術の研究に関連する出版物の翻訳などの資料を刊行したほか、学術誌等へ投稿した。

##### 平成25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
25. 5	製表技術参考資料22	・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
25. 8	製表技術参考資料23	・諸外国における最新のデータエディティング事情 –混淆正規分布モデルによる多変量外れ値検出法の検証–
25. 10	2012年度統計技術研究会報告	・統計調査における経理項目の欠測値補定方法について –EMBアルゴリズムによる多重代入法– ・UNECE統計データエディティングに関するワークセッションの概要(出張報告) ・Privacy in Statistical Databases 2012 (統計データベースにおけるプライバシーに関する会議) について ・匿名化技法としてのスワッピングの可能性 –国勢調査マイクロデータを用いた有用性と秘匿性の実証研究について–
26. 1	製表技術参考資料24	・匿名化技法としてのスワッピングの可能性について –国勢調査マイクロデータを用いた有用性と秘匿性の実証研究–

##### 平成25年度 学術誌等への投稿実績

投稿年月	学術誌等名	内 容
25. 12	統計研究彙報第71号	・様々な多重代入法アルゴリズムの比較 –大規模経済系データを用いた分析– ・マイクロデータにおける匿名化技法の適用可能性の検証 –全国消費実態調査と家計調査を用いて–

### 第5節 統計活動に関する国際協力

#### 第1 海外への技術協力

平成25年度は、17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、2回にわたり延べ4名の集計・プログラミング専門職員についてカンボジア王国への派遣を行うと共に、我が国で開催されたカンボジア統計局職員に対する研修の際に、統計集計技術に関する研修を統計センターが実施した。

カンボジアでは、日本政府及び国際協力機構の支援の下、2014年3月に、全国の事業所の所在地、従業員数、売り上げ等をサンプル調査する中間年経済統計が実施された。

## **第2 LIS<sup>21</sup>のデータベース利用に関する支援**

LISが整備しているデータベースの利用について、平成21年10月に同機関と協定を締結している。平成25年の11月に同機関と拠出金支払いに係る合意書の締結（平成26年～30年）を行い、政府機関の職員その他国内の大学や非営利機関の研究者が利用するための支援を行っている。

---

<sup>21</sup> LIS (GROSS-NATIONAL DATA CENTER in Luxembourg) : 各国の政府機関等の協力・支援を得て、家計所得に関係する各国の調査データを収集し、国際比較研究に利用可能なデータベースを整備しているプロジェクト。現在、40か国から家計所得に関するデータ提供を受けている。LISのデータは、所得分布に関する国際的なデータベースとして有名なものであり、各国の経済学者や社会学者に幅広く使われている。