

海外の公的統計や行政記録情報に関するデータリンケージの現状について

中央大・経済 伊藤 伸介

異種の統計調査の個票データや行政記録データとの間のデータリンケージに対する関心が高まっている。統計調査の中には、世帯・人口系の統計調査と事業所・企業系の統計調査が存在するが、前者では個人や世帯、後者においては企業や事業所が調査対象であることから、世帯・人口系の個票データと事業所・企業系の個票データのリンケージは容易ではない。また、公的統計データと行政記録データのリンケージや、行政記録データ同士のリンケージの可能性については、リンケージを行うための個体識別子(名前、住所、個体識別番号等)をどういった形で保管するかを検討することが求められる。

北欧諸国、さらにはオランダといった国々では、行政記録データに基づいたレジスターベースの統計の作成が行われている。そのために、各レコードに一意的な個体識別番号が付与されており、行政記録データ間のリンケージが可能になっている。例えば、デンマークでは、中央人口レジスター(Central Population Register)が、個人番号を管理しており、名前、年齢、性別、結婚状態、住所等を含む個人や家族の情報がレジスターに登録されている。デンマーク統計局では、統計作成において必要な属性情報を行政記録から取得することができる。そして、個人番号を用いて、社会給付状況、保健衛生、学校教育活動、雇用状況といった社会人口的な属性に関する情報のリンケージが可能になっている。一方、イギリスの場合、デンマークやオランダのようなレジスターベースの統計を作成している国とは異なり、リンケージを可能にするための共通のIDが個体レコードには存在しない。そこで、名前と住所を用いて仮名化(pseudonymization)を施した上で、確率的なリンケージが行われる。また、イギリス国家統計局は、患者レジスターや学生レコードといった行政記録データを対象に、スコアを用いたマッチングやロジット回帰モデルに基づくマッチング等、様々なマッチングの方法に関する研究を行ってきた(伊藤(2017))。

他方、諸外国においては、学術研究目的のために行政記録データの利用や公的統計マイクロデータとリンケージされた行政記録データの活用も広範に展開されてきた。例えば、イギリスでは、Administrative Data Research Network(ADRN)と呼ばれるプロジェクトが進められており、公的統計のマイクロデータと行政記録データのリンケージや行政記録データ間でリンケージされたデータの提供が行われている。ADRNは、学術研究目的のための非識別データ(deidentified data)を以下のように作成・提供している。①ADRNは、承認された研究プロジェクトを対象に、データの管理者(データを保有している政府機関)にそのプロジェクトに関連するデータの提供に関する交渉を行う。②データの保管者が、データの提供に同意した場合には、各レコードから個人を直接識別することが可能な情報(名前、生年月日等)を切り離し、代わりに一意の参照番号を付与したデータをオンサイト施設が備わっているADRC(=Administrative Data Research Centre)に送付する。③直接識別情報と各レコードに付与される参照番号が、信頼される第三者機関に送られる。なお、第三者機関には、データは送付されない。④信頼される第三者機関は、一意な参照番号と直接識別情報のマッチングを行い、マッチングされた一意な参照番号の組(index key)のみを残した上で(直接識別情報を消去)、このindex keyをADRCに送付する。⑤ADRCは、index keyに基づいて複数のデータ間のリンケージを行った上で、index keyと参照番号を消去する。⑥研究者は、ADRC内部のオンサイト施設においてこのリンケージデータにアクセスし、研究を進める。

本報告では、海外の公的統計や行政記録情報に関するデータリンケージの現状を明らかにすることによって、公的統計の分野におけるデータリンケージの可能性について議論していきたい。

参考文献

伊藤伸介(2017)「公的統計における行政記録データの利活用について—デンマーク、オランダとイギリスの現状—」『経済学論纂(中央大学)』第58巻第1号, 1~17頁