

税務情報の利活用

— 税務大学校との共同研究と匿名データの提供 —

佐藤 黎

はじめに

国税庁では、確定申告情報などの国税に関する国税庁保有行政記録情報（以下、「税務情報」という。）を保有しており、従前から税務情報を活用した研究に対するニーズがあった。そこで、「オープンデータ基本指針」等の政府方針を踏まえ、学術研究を通じた税・財政政策の改善・充実に資する観点等から、令和4年4月より、税務大学校との共同研究による個票データの提供、令和7年4月より、匿名データの提供を開始した。本稿では、この2つの取組について紹介する。特に、匿名データは提供を開始したばかりであり、本稿を読んで関心を持っていただけたようであれば、国税庁のホームページにアクセスし、詳細を確認した上で利用申請していただければ幸いである。

1 税務大学校との共同研究

(1) 概要

税務大学校との共同研究（以下、「共同研究」という。）は、税務情報の個票データ（以下、「個票データ」という。）を利用して、税務大学校と共同して、我が国の税・財政政策の改善・充実に資する統計的研究等を行っていただく取組である。個票データの利用に当たっては、年に1回程度、公募を実施しており、応募内容について、有識者会議を開催して意見をいただきながら、審査・採択を行っている。

当該取組は税務大学校と共同して研究するこ

とになるが、税務大学校は、税務職員に対する研修を行うだけでなく、税務に関する学術研究なども行っている機関である。共同研究では、税務大学校の職員は、研究者が希望する研究目的に応じて、データのクレンジング処理や、データの管理などを担っている。

本取組の特徴としては、生に近い税務情報を利用して研究を行うことができるところにある。生に近い税務情報を利用する関係上、使用場所は税務大学校の校舎内（埼玉県和光市に所在）に限定している。また、税務大学校の客員教授（国家公務員）に任用して税務情報に関する守秘義務を課すなどの措置を実施している。

(2) 個票データの概要

個票データでは、氏名などの個人情報の削除といった最小限の加工は施しているが、いずれも収入、所得、税額などの定量情報は加工せず提供している。本稿では、直近の公募（令和7年2～3月）の際に提供対象とした税目（所得税・法人税・相続税・消費税の4税目）を例に説明する。

まずは、所得税の個票データについてである。所得税の個票データは、個人の確定申告情報であり、個人の確定申告は、毎年2000万件を超える申告がなされている。研究では、この全ての確定申告情報が利用可能である。個票データとして、実際に提供するデータは、所得税申告書の第一表と第三表の情報である。第一表が所得

税額の計算をする様式で、第三表が（株の譲渡益などの）分離課税を計算する様式になっている。これらの申告様式の中には、納税者の属性に係る定性情報や、収入、所得、所得控除、税額などの定量情報について、合計額だけでなく内訳も含まれる。

次に、法人税の個票データについてである。法人税の個票データは、法人の確定申告情報であり、法人税は、毎年300万件を超える申告がなされている。研究では、この全ての確定申告情報が利用可能である。個票データとして、実際に提供するデータは、法人税額の計算を行う、法人税申告書の別表一の情報である。別表一には、法人の属性に係る情報のほか、売上金額、所得金額、法人税額などの定量情報が含まれている。また、平成28年以降の申告では、申告を行っている法人番号も含まれているので、他の情報との紐付けを行うこともできる。

そして、相続税の個票データについてである。相続税の個票データは、亡くなった方の資産の移転（相続）に係る申告情報であり、相続税は、毎年10万件を超える申告がなされている。研究では、この全ての申告データが利用可能である。個票データとして、実際に提供するデータは、相続税申告書の第一表と第十五表の情報である。第一表は相続税税額を計算する様式で、第十五表は相続財産の種類別の価額の情報を記載する様式になっている。相続税の申告は、相続人の人数などに応じて、提出の基準額が設定されているので、一定以上の資産を有する者を前提とした分析に向いている。

最後に、消費税の個票データについてである。消費税の個票データは、国税に係る消費税の申告情報で、法人消費税、個人消費税の2種類がある。法人消費税は毎年180万件程度、個人消費税は毎年100万件を超える申告がなされてお

り、その全ての申告データを研究に利用可能となっている。個票データとして、実際に提供するデータは、消費税申告書の第一表と第二表である。消費税税額の計算をする様式に記載されている情報を提供する。法人税と同様に、法人の消費税申告については、法人番号をもとに、他の情報との紐付けを行うことも可能である。また、法人税の個票データと法人消費税の個票データの両方を提供することも可能であり、この場合は、法人税の個票データに含まれる法人の属性情報等と、法人消費税の個票データの情報を組み合わせて分析することができる。

なお、本稿では参考として所得税申告書の第一表と第三表を掲載する（図1）。

(3) 利用要件

個票データの利用に当たって、利用者、利用目的、利用期間、利用方法などに一定の要件を設けている。本稿では主な利用要件について紹介する。

まず、利用者についてであるが、個票データの利用者は、公的機関・大学に所属する常勤の研究者等に限定している。1つの研究プロジェクトに最大5名まで参加することができるが、いずれの者も利用者の要件に当てはまっていることが必要である。

次に、利用目的についてであるが、個票データの利用に当たっては、①学術研究の発展に資するもの又は公的機関における施策の推進に適切に反映されるもの、②税・財政施策の改善・充実に資する統計的研究又は国税庁が保有する行政記録情報の整備であることの①②の両方に該当することを要件としている。そのため、個人の特定を目的として使用することや、脱税を意図した研究等に活用することは認めていない。

続いて、利用期間は、原則3年間を上限とし

ている。当初は原則2年間を上限として、やむを得ない場合に限り、最長1年間の延長を認めることとしていたが、個票データのデータ量が大きく、データ分析等に時間を要した研究が多くあり、利用期間延長が相次いだことから、利用期間を見直した。

最後に、利用方法についてであるが、個票データは、前述のとおり税務大学の校舎内でのみ利用できる。個票データそのものを持ち出すことはできないが、

研究生成物の施設外持出しは、チェック基準をクリアすれば可能である。例えば、度数表を施設外に持ち出す場合のチェック内容は表1のとおりである。

図1 所得税申告書の第一表・第三表

[illegible]

(4) 利用環境

令和7年6月現在、税務大学校では、共同研究用の端末を4台用意している。また、当該端末に分析用のソフトウェアを導入している。端末の性能・導入しているソフトウェアは表2のとおりである。

(5) 利用手續

年に1回程度、期間を定めて公募(例年3～5月頃)を行っており、利用を希望する場合には公募期間中に自身の研究内容等を記載した申出書等の必要書類を国税庁に提出する。国税庁において当該書類の記載内容を確認した上で、有識者会議で助言を受け、国税庁が利用の可否

[illegible]

表1 度数表の標準的なチェック内容

チェック内容	利用者が提示する情報
①【度数】各セルが1以上 10未満の調査客体から算出した値でないこと	①【度数】各セルの度数
②【数量】各セルにおいて、70%を超えて寄与する調査客体がないこと	②【数量】各セルにおいて最も大きく寄与する調査客体の値が当該セルの値に占める割合
③【数量】各セルにおいて、85%を超えて寄与する2つの調査客体の合計値がないこと	③【数量】各セルにおいて1番目及び2番目に大きく寄与する調査客体の値及びその合計値が当該セルの値に占める割合
④【度数】行計又は列計の90%超を占めるセルがないこと	④【度数】各セルの行計及び列計に占める割合

表2 税務大学校における共同研究用の端末

	項目	仕様等		
ハードウェア	CPU	インテルXeon Gold 5416S プロセッサー コア数16個×CPU数2個	インテルXeon Gold 6226R プロセッサー コア数16個×CPU数2個	インテル Core プロセッサー i5-13400 コア数10個×CPU数1個
	メモリ	512GB (8×64GB)	1TB (16×64GB)	64GB (2×32GB)
	GPU	NVIDIA T400 4GB		
	ストレージ	10TB (4TB SSD+6TB HDD)	2TB HDD	2TB (1TB SSD+1TB HDD)
	台数	1台	1台	2台
ソフトウェア	機能	名称等		
	統計解析機能	StataMP18(32-core)		StataMP18(6-core)
		R 4.4.1		
		R Studio 2024.04.2		
	データベース管理	Python環境		
		Microsoft SQL Server 2016		
	データ可視化	Quarto 1.5.45		
	エディタ	Sakura Editor		

を決定する。

利用が承諾されれば、誓約書等の必要書類を提出し、国家公務員に任用の上で共同研究を開始できる。

(6) 研究実績等

これまで、5回の公募を実施し、合計で10件の研究が開始されている。研究のテーマとしては、所得税5件、法人税3件、相続税・消費税がともに1件ずつである。その成果として、12件のディスカッションペーパーを国税庁のホームページで公表（令和7年6月現在）している。

2 匿名データの提供

(1) 概要

ここでいう匿名データは、税務情報を個人の識別ができないように匿名加工したデータを指す。我が国の税・財政政策の改善・充実に資する統計的研究を行っていただくよう、匿名データを提供する取組であり、令和7年4月に提供受付を開始した。

本取組の特徴としては、税務情報に個人情報等が識別できないよう匿名加工を施すところにある。匿名加工を施していることから、利用者自身の研究室などで利用可能としており、税務

大学校との共同研究に当たって提供している個票データに比べて利便性が高くなっている。また、申請についても年中を通して可能としており、研究を行いたいタイミングで匿名データを利用することができる。さらに、共同研究と比べて、博士研究員や大学院生（博士課程後期相当に限る）も利用可能（詳細は(3)利用要件参照）としており、より多くの研究者に利用されることを想定している。もちろん、匿名データの利用を踏まえた上で、より詳細な研究を希望される場合は、税務大学校との共同研究に応募することも可能である。

(2) 提供データの概要等

匿名データとして現在提供を行っているのは、所得税の申告書第一表と第三表の情報を匿名加工したものである。提供の元となっている税務情報は、個票データと同様である。このうち、個人の特定リスクが高い情報は匿名加工されることになる。

匿名加工は、大きく分けると3つの加工処理を行っている。1つ目が、標本抽出処理である。具体的には、申告情報のうち1%を標本抽出する処理を行っている。前述のとおり所得税は、毎年2000万件を超える申告があるため、年間20万件超の申告を抽出して提供することとなる。

2つ目が、納税者情報の匿名化処理である。申告情報のうち、納税者の属性などに係る定性情報については、名前、マイナンバーといった個人情報情報は削除している。また、性別については、最新の申告書では記入欄を設けていないため、一律削除を行う。そのほか、個人の特定リスクが高まる情報については、グルーピングをして提供することとしている。具体的には、住所は三大都市圏かそれ以外かの2区分にグルーピング、年齢は5歳階級でグルーピング、業種

は12区分にグルーピングをして提供する。

3つ目が、収入、所得、税額などの定量情報への匿名化処理である。まず、収入、所得、税額といった項目ごとに、定量情報の上位0.5%及び下位0.5%を上限値又は下限値として設定し、その数値を超える場合には、上限値又は下限値に置き換える処理（トップコーディング・ボトムコーディング処理）を行っている。また、この処理に関連して、申告情報の内訳と合計額の関係から、置き換える前の数値が類推できる場合があることから、類推できないよう、一部の項目を削除する処理を行うこととした。これらのほか、桁数に応じたラウンディング処理を行っている。匿名加工の具体的なイメージは図2のとおりである。

まず、各サンプルにおいて、納税者情報の年齢について、5歳階級でグルーピングを行っている。また、住所は三大都市圏かそれ以外か、業種は12区分でグルーピングをしている。これらの情報については数字に符号化している。定量情報には、0が並んでいるが、これは桁数に応じてラウンディング処理を行ったためである。

次に、サンプルNo4.の定量情報のうち、黒枠で囲んでいる箇所が、前述したコーディング処理を行っている箇所である。この例では、上位0.5%を上回る数値が入っていたものになる。事業収入では、上位0.5%のラインが7200万円だったので、仮に1億円だったとしても、7200万円の数値が入るということになる。

また、出現頻度の低い項目に値がある場合、またはコーディング前の値を推測可能な場合は、個人の特定リスクが高まるため、サンプルNo5の定量情報の右側にあるような、総所得金額、所得控除合計額といった一部の項目のみを提供することとしている。

なお、コーディングされた値については、コ

図2 匿名加工のイメージ

…コーディングした値

サンプル NO	納税者情報等			収入金額等			所得金額				総所得金額	所得控除 の合計額	課税所得 金額等	申告 納税額
	年齢	住所	業種	事業収入	給与収入	総合課税	事業所得	利子所得	給与所得	総合課税				
1	3	0	0	1,0980,000	0	3,500,000	3,840,000	0	0	1,750,000	5,590,000	1,300,000	4,280,000	430,000
2	0	1	9	0	3,880,000	3,500,000	0	0	2,660,000	1,750,000	4,410,000	900,000	3,510,000	-130,000
3	2	0	9	0	3,660,000	0	0	0	2,480,000	0	2,480,000	700,000	1,780,000	80,000
4	1	1	6	72,000,000	1,740,000	0	10,000,000	0	1,140,000	0	30,000,000	4,000,000	30,000,000	4,000,000
5	9	0	0								30,000,000	4,000,000	30,000,000	4,000,000

出現頻度の低い項目（0.5%未満）に値がある、またはトップコーディング前の値を推測可能（※）なデータは、内訳（点線内）を空欄化している。
※例えば、所得金額の内訳を合計することにより総所得金額の値が推測可能な場合を指す。

ーディング前の値の推定が困難であることから、分析の補助となるよう、コーディング前（桁数に応じたラウンディング処理後）の平均値・標準偏差もあわせて提供している。

(3) 利用要件

匿名データの利用に当たって、利用者、利用目的、利用期間、利用方法、研究成果の公表に一定の要件を設けている。本稿では主な利用要件について紹介する。

利用者は、①公的機関・大学などに所属する常勤の研究者、②当該研究者が直接指導・監督することを前提に、研究者と同じ機関に所属する博士後期課程相当の大学院生などに限定している（表3）。

利用に当たっては、その目的として、①学術研究の発展に資するものであること、②税・財政施策の改善・充実に関する統計的研究であることの両方に該当することを要件としている。のため、個人の特定を目的として使用すること

表3 利用者の範囲

- ①匿名データの利用者の範囲は、次のいずれかの機関に所属する常勤の研究者
- ・ 公的機関

・ 独立行政法人 ・ 地方独立行政法人

・ 大学 ・ 大学共同利用機関
- ②上記①の利用者が所属する機関と同一の機関に在籍する、次のいずれかの身分を有する研究等の補助者
- ・ 博士研究員

・ 大学院生（博士課程後期相当に限る）

や、脱税を意図した研究等に活用することは認めない。

また、データの利用は、国内の施錠可能な場所、外部ネットワークから遮断された環境に限定して可能としている。データの提供に当たってはCD-R等の磁気媒体を貸し出して行っている。利用している間には、利用者がCD-Rをどのように管理しているかを把握するために、国税庁が様式を定めている管理簿を提出することとしている。仮に、漏えいなどの不適切利用が発覚した場合には、事例に応じて、データの提供禁止や、氏名公表などの措置を講じる可能性がある。

なお、利用期間は、共同研究と比較して提供するデータ量が小さいことから、原則2年間を上限としている。ただし、やむを得ない場合に限り、最長1年間を上限として、延長を認める場合がある。

(4) 利用手続

匿名データの利用申請は、国税庁ホームページに掲載している様式により年中を通して行うことができる。プロセスとしては、まず、正式な申請前に仮の申出書等の必要書類を提出し、

仮の申出書の内容について、国税庁が形式上の確認を行い、問題がなければ正式に利用申請を行う。国税庁から承認が得られれば、利用を開始することができる。なお、仮の申出書の形式上の確認は事業者が行う場合がある。

匿名データの利用による研究成果物については、公表前に国税庁で個人情報などが含まれていないかの確認を行う。問題がなければ、公表することが可能である。なお、当該研究成果物については、国税庁ホームページでもあわせて公表する。

利用後には、公表物以外の生成物は削除して、匿名データ（CD-R）を返却することになる。

おわりに

税務情報が広く活用され、税・財政施策等に資する研究が発展していくことは、国益に繋がるものと考えている。本稿を読んで、税務情報の利用に関心を持っていただけたのであれば、国税庁のホームページに更に詳細な情報を掲載しているため、ホームページを確認の上で利用申請をいただければ幸いである。もし、不明な点があれば、国税庁ホームページの問合せ先に連絡いただきたい。

(さとう れい)

国税庁長官官房企画課データ活用推進室課長補佐)