特　集

税務データの研究利用可能性とEBPM

近藤 絢子
東京大学 社会科学研究所 教授

1．はじめに

これまで、日本における社会科学の実証分析に用いられるデータは、母集団から抽出された個人・世帯・事業所等を対象に実施される調査に基づくサーベイデータが中心だった。筆者の専門分野である労働経済学でよく利用される、労働力調査（総務省統計局）、就業構造基本調査（総務省統計局）、賃金構造基本統計調査（厚生労働省）などの政府統計もすべてサーベイデータである。サーベイデータにも、知りたいことを直接質問できるなどのメリットはあるものの、情報の正確さやサンプルの漏れの少なさなど、行政記録を活用した業務データのほうが優れている点も多い。

行政業務データの研究利用に関しては、日本は欧米のみならず、台湾や韓国など他の東アジア諸国に比べても大きく後れを取っている。行政記録情報を本来の業務目的以外に用いることに対する抵抗感が大きく、個人のプライバシーを保護するための技術的な解決法がなかなか共有・実践されてこなかったためだろう。近年、ようやく行政業務データの政策立案の基礎資料としての有用性が認知され、大学や民間研究機関との連携による研究利用が少しずつ広がってきた。

そうしたプロジェクトの一つである「EBPM推進のための自治体税務データ活用プロジェクト」（以下「自治体データプロジェクト」と省略）では、東京大学政策評価研究教育センター（CREPE）と協力自治体が連携して、匿名化処理を施した市町村の税務データを研究者に提供するしくみを構築した。CREPEが協力自治体に対して匿名化のためのプログラムなど必要な情報やツールを提供し、自治体がデータの抽出・匿名化を行って、匿名化されたデータをCREPEに提供する。提供されたデータを用いて、CREPEが翌年度の税収の予測や、住民や転出入者の年齢や所得の分布などについての記述的な分析を行い、それぞれの自治体に対してフィードバックを行うとともに、プロジェクトに参加している研究者が提供されたデータを用いて学術研究を行うのである[1]。

自治体データプロジェクトでは今のところ税務データのみの提供だが、税務データと他の行政業務データを接合することで、さらに幅が広がる。例えば、筆者が客員主任研究官として参加している内閣府経済社会総合研究所のプロ

[1] 自治体データプロジェクトは2021年に立ち上げられ、2022年に本格的に始まった。本稿執筆時点（2023年12月）では2023年度のデータのクリーニングや自治体へのフィードバックのための分析を行っている最中であり、学術研究に利用できるのは、2022年度に提供された25市町村のデータとなっている。プロジェクトの全体像については川口・正木（2022, 2023）を参照。

ジェクトでは、首都圏のある自治体より、未就学児のいる世帯について税務データと認可保育所の利用や申し込みに関する記録を接合したものの提供を受けて、認可保育所の利用の可否が母親の就業に及ぼす影響を分析している。

本稿では、まず自治体データプロジェクトで提供している、住民税の課税記録と住民登録情報のみのデータについて、その特徴や利用可能性、限界点などを説明し、具体例として「年収の壁」について分析した近藤・深井（2023）を紹介する。続いて、他の行政業務データとの接合によってさらにどういったことが可能になるかを概観したのち、具体例として先述の認可保育所のデータを用いたFukai and Kondo（2024）を紹介する。

2．税務データでできること、できないこと

自治体データプロジェクトで協力自治体に提供を依頼するデータは、各年初に住民登録がある全住民の生年月・性別・世帯主との続柄・ハッシュ化した世帯IDと、個人住民税の賦課額決定に用いた収入・所得・控除等の情報及びそれに基づき決定された個人住民税賦課額をハッシュ化した個人ID（宛名番号）で接合したものである。最低5年分の提供を求め、同一個人を追跡できるよう個人IDは年度間で共通にしている[2]。

これらの情報から、正確な給与収入・各種所得と世帯構成の情報を含む個人レベルのパネルデータを構築することができる。正確な収入や所得の情報がとれることが、税務データの一番の強みだ。サーベイデータでは、100万円単位でのラウンディングが起こりやすく、そもそも「100万～200万円」などの階級でしか聞いていない場合も多い。年収を訊くと回答率が下がりやすいこともあって、正確な年収を得ることは難しい。住民基本台帳と住民税の課税記録を接合した行政業務データなら、課税額の決定に用いられた正確な所得が全住民についてわかる。とくに給与収入については、非課税となる低所得者も含めて源泉徴収票に基づく正確な額面給与が分かる。

また、他市町村へ転出しない限りデータから脱落しないという利点もある。例えば、サーベイに基づくパネルデータにおいては、家族構成の変化をきっかけとして調査から脱落する確率が上がってしまう問題が知られている。結婚や離婚によって世帯が変わってしまうと、追跡できずに脱落しやすくなるのだ。これは、結婚・離婚や出産などのライフイベントの変化の前後の分析には特に重要な問題となる。市町村の行政業務データについても、家族構成の変化をきっかけに他市町村へ転居することもあるので留意は必要だが、同一自治体内であれば転居しても追跡は可能だ。世帯構成から出産などのイベントを識別できるので、ライフイベント前後の変化の推計には特に適したデータであるといえる。

ただし、自治体の行政業務データには、自治体が持っている情報しか含まれないという欠点もある。たとえば、教育年数は賃金などに大きく影響する重要な変数であり、ほとんどのサーベイデータには含まれているが、自治体が住民の学歴を把握していることは稀である。また、勤め先の業種や企業規模、本人の職種や雇用形態といった情報もないことが多い。さらに言えば、1年間の収入はわかるが労働時間の情報が

[2] 同一個人を追跡できるか否かは研究利用の可能性に大きく影響する。個人の特定リスクが上がるというデメリットがあるのも事実だが、対応は可能である。個人の特定リスクへの具体的な対応策（匿名化処理等）について詳しくは正木（2022）を参照。

ないので、時間当たり賃金もわからない。行政業務データからは、サーベイデータでは得られなかった情報が得られる反面、サーベイデータであれば普通は含まれている情報が抜けていることもある点には注意したい。

このように意外と制約は多いのだが、上手に使えば既存のサーベイデータではできなかった分析が可能になる。その一例として、次節では既婚女性の就労調整行動についての記述的分析を行った近藤・深井（2023）を紹介する。このほか、出産前後の女性の収入の変化（いわゆるchild penalty）の分析にも適したデータであり、欧米を中心とした国際比較研究であるKleven et al（2023）でも多くの国について行政業務データが利用されている。また、給与収入と公的年金収入の両方を正確に把握できるので、在職老齢年金制度や年金支給開始年齢の変化と高齢者の就業についての分析もできるだろう。

3．税務データ＋住民票情報のみでできる研究の具体例：年収の壁の分析

近藤･深井（2023）は､自治体データプロジェクトの参加自治体のうち、分析に必要な条件を満たす16自治体のデータを用いて、いわゆる「年収の壁」の実態について記述的な分析を行った。「年収の壁」とは、社会保険料や税の負担を避けるために一定以下に年収を抑える行動のことで、日本の制度の下では、配偶者の扶養に入っている既婚女性に特に顕著にみられやすい。2023年初の首相の施政方針演説で言及されたことに始まり、秋には年収の壁・支援強化パッケージが打ち出されるなど、改めて注目を浴びている。

政策的に非常に重要なトピックである一方で、「年収の壁」に対する就労調整の実態を正確に把握することは、従来のサーベイデータでは難しかった。住民税所得割がかかり始める100万円、税制上の扶養から外れる103万円、年金の第三号被保険者でなくなる130万円など、100万円台前半にいくつか制度上の閾値が並んでいるが、1万円単位で年収を正確に把握できるデータはほぼないからだ。労働力調査（総務省統計局）などの政府統計の多くは年収を「50万円以下」「50-99万円」「100-50万円」といったカテゴリでしかとらえられないし、実額を記入させる形式であってもキリのいい数字に丸めて回答する人が多い。とくに100万円へのラウンディングが起きると就労調整の実態把握には致命的な問題となる。

この点､市町村の個人住民税の課税記録なら、非課税となるような低収入の個人を含む全住民の正確な給与収入が手に入る。これを使えば、十分な解像度で年収分布をとらえることができ、どの「壁」がどのくらい影響しているのかが一目でわかるようなヒストグラムを描くことができる。世帯構成から配偶者の有無も識別できるし、配偶者の給与収入や合計所得金額などもわかるので、制度上より影響が強いと予想される層を抜き出してみることもできる。自治体データプロジェクトで提供されるデータは、自治体ごとに最低5年間のパネルデータとなっているため、結婚や出産の前後の変化を見ることも可能である。

分析例として、結婚前後の女性の年収分布を比較したヒストグラムを示す（図1）。結婚前は、配偶者の扶養に入ることはできないので、年収の壁は関係ないはずである。結婚前年のデータで作成した、図1の左側のグラフを見ると、税制上の扶養家族の所得上限である103万円にはわずかなスパイクがあるものの、社会保険制度上の扶養配偶者の所得上限である130万円には不連続はない。結婚翌年のデータで同じようにヒストグラムを描いたのが右側だが、103万円と130万円に大きな不連続ができている。つまり、結婚を機に、103万円や130万円に年収を

おさえる女性が一定数存在することがわかる。

近藤・深井（2023）では、第1子の出産を機にさらに多くの女性が103万円や130万円に調整するようになることや、夫の所得が高い方が扶養に入りやすいことなど、様々な分析を行っている。興味のある読者は直接論文を参照されたい。

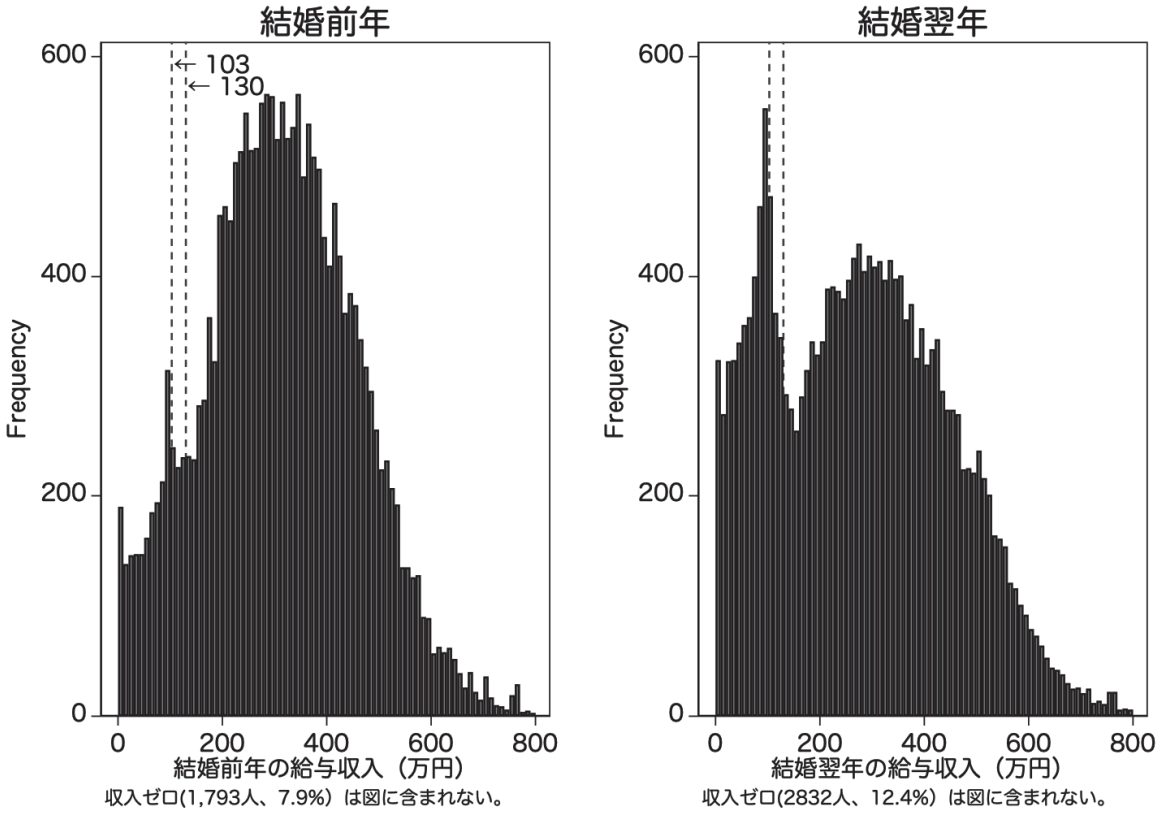

図1　結婚前後の女性の給与収入分布の比較

出所：近藤・深井（2023）図10

4．他の行政記録との接合で応用可能性をひろげる

年収の壁や出産前後の所得の変化などは、税務データと住民票の情報だけもできる分析の例だが、税務データと自治体が行っている様々な行政サービスの利用記録を接合することによって、応用可能性は大きくひろがる。

一つの方向性として、教育関連のデータとの接合が考えられる。欧米での先行研究から、家庭の経済状況や家族構成の変化が子供の発達、学力、進路決定などにさまざまな影響を及ぼすことが知られている。日本でも最近は全国学力テストなど学力テストの個票データを用いた研究ができるようになってきたが、家庭の状況についての情報は、学力テストと同時に行ったアンケート調査に基づくものがほとんどで、自己回答による誤差が問題となる。

もし、学校の成績や特定の学力テストの点数などと、税務データを接合できれば、その生徒の住む世帯の世帯構成や家族の収入についての正確な情報が得られる。さらに学校や学級単位の情報（クラスの人数など）も加えれば、様々な分析が可能になる。

世帯収入を説明変数として使うのではなく、行政サービスの利用状況が個人の収入に与える影響を見ることもできる。自治体が提供する行政サービスは多岐にわたるが、次節では認可保育所の入所の可否が母親の就業や年収に与えた影響を見たFukai and Kondo（2024）を例として挙げる。

具体例にうつる前に、実際に分析用のデータを構築する際の注意点を述べたい。まず、行政サービスの利用の影響を分析する際には、比較対象としてそのサービスを利用していない人も含めたデータセットを作ることが肝要となる。サービス利用者を母集団とするのではなく、全住民を母集団としてサービスの利用についての情報を接合するのである。また､接合に際して､異なるデータソース間で共通の個人IDが振られていない場合は、名前・生年月日などを用いた名寄せ作業が必要となる。この作業には非常に手間がかかるので、これからデータを蓄積する段階で個人を識別する工夫をしておくことが将来の利用可能性をひろげることにつながる。

5．具体例：認可保育所入所選考とその後の母親の就労

出産後も仕事を続けたい女性にとって、子供を保育所に入れられるか否かは、非常に重要な問題だ。だが、欧米における先行研究では、保育所の拡充が必ずしも母親の就業率や収入の上昇につながっていない。日本と違って、保育所の入所条件に就労を課していないことも一因だが、保育所に入れなくても祖父母に頼ったりベビーシッターを雇ったりして就業継続していた世帯が保育所に乗り換える「クラウディングアウト」が起きている場合も多い。

保育所を利用するか否かは自発的な選択なので、単純に保育所を利用している世帯としていない世帯を比べるだけでは、保育所が利用できることの効果は測定できない。先行研究では、保育所の拡充タイミングの地域差を利用した差の差推計や、保育所拡充政策の対象となる学年とならない学年の境目を比べた回帰不連続モデルなど、保育所のキャパシティの変化を保育所利用の操作変数として使うものが多い。だが、こうした手法を使えるのは保育所のキャパシティが外生的な理由で一気に大きく拡大した場合だけだ。

しかし、保育所の入所申し込みに関わる行政業務データがあれば、保育所に申し込んだ世帯だけに限定して、入れた世帯と入れなかった世帯を比較することができる。これだけでも保育所利用の内生性の問題はかなり改善される。入所希望者の中でも必要度に応じて入所できる確率に差がでる仕組みになっている場合は、申し込んだ中での単純比較ではなく保育所の入所選考ルールも加味した推計を行うことで、可能な限り似た世帯で、入れた場合と入れなかった場合を比較することができる。

Fukai and Kondo（2024）は、内閣府経済社会総合研究所のプロジェクトの一環として、首都圏のある市より、未就学児のいる全世帯について、課税記録と認可保育所の入所申し込みと実際の入所状況のデータの提供を受けた。この市では、新年度の始まる4月に認可保育所に入るには、12月までに申込書を提出する必要がある。申込書に記載された情報（父母それぞれの労働時間、兄姉の有無と学年、祖父母との同居など）に基づいて、市がそれぞれの世帯の調整指数を算出し、指数の高い申込者から順に保育所の枠が割り当てられ、同じ指数の世帯の間では前年度の住民税課税額（つまり所得）が低いほうが優先される。入所を希望する保育所はいくつ書いても良く、希望する保育所が全て埋まってしまうと入所不承諾となる。

つまり、保育所に入所できる確率は調整指数と、どの保育所を希望したか、そして前年度の

住民税課税額に依存して決まる。これらを制御すれば、１次選考で承諾されるか否かは、ランダムとみなしてよいため、１次選考の結果を保育所利用の操作変数として用いた。認可保育所に子供を入れられるかどうかのボーダーラインにいるのは、ほとんどの場合共働きで母親が育児休業から復帰するタイミングでの入所を希望する第１子である。操作変数による推計結果は、こうしたボーダーライン上の世帯にとっての認可保育所入所の効果と解釈できる。

得られた推計結果の一部を抜粋したのが表１である[3]。子供を認可保育所に入れることができると、母親の就業率は０歳児で 40%、１歳児で 19%上昇する。すなわち、もともとフルタイム共働きだった世帯の第１子が認可保育所の選考に落ちると、０歳では４人に１人、１歳では５人に１人の母親が仕事への復帰を断念する。年収への効果は０歳児で 141 万円、１歳児で 91 万円だ。出産前の母親の平均年収が約 400 万円なので、年収への影響はほぼ就業確率の変化の影響とみてよい。０歳と１歳の差は、０歳の場合は育休延長ができることがある一方、１歳のほうが認可外保育施設には入りやすいという、認可保育所に入れなかった場合に取られがちな代替手段の差に起因すると推測される。

本プロジェクトで提供されたデータは４年間という短い期間であるため、母親のキャリアに対する長期的な影響を見ることはできないが、もっと長い期間のデータがあれば、０歳や１歳で保育所に入れたか否かが、子が幼稚園や小学校に入れる年齢になった時の母親の年収へ与える影響をみることもできる。さらに、もし市町村が持っているほかの業務データとの接合が可能であれば、保健所の持っている乳幼児健診のデータから子供の発達と保育所利用の関係や小学校での成績との関係を検証することも可能になる。

表１　認可保育所入所が母親の就業や年収に与える効果

	０歳児	１歳児
母親の就業（給与収入が正）	0.403	0.188
母親の給与収入（万円）	141.1	91.1

出所：Fukai and Kondo（2024）Table 6, 7 より抜粋

６．おわりに

本稿では、税務データを中心に、市町村の持つ行政業務データを活用することで可能になる実証分析を、具体例を２つ挙げて説明した。いずれも、サーベイデータでは難しかった分析が行政業務データの利用によって可能になった事例であり、行政業務データの学術研究利用の有用性を示すものといえよう。

データの利用可能性がひろがるとともに、データの扱い方や目的に合わせた分析方法についての知識がより重要になってくる。客観的なデータに基づく政策判断が重要であることはいうまでもないが、具体的な数値やグラフが出てくるだけで、なんとなく信頼できそうな印象を与えてしまい、不適切な処理で間違った数値やグラフでも看過されやすいのも事実である。計量経済の専門的な知識はいらないが、相関関係

[3] ただし、現時点ではまだ未定稿であり、推計値も暫定的なものである点には留意していただきたい。

は必ずしも因果関係を意味しないことや、多くの場合行政サービスの利用は内生的であることなど、基本的な概念の理解は必要だ。中室・津川（2017）、伊藤（2017）など読みやすい入門書がいくつもあるので、データ分析をする当事者だけでなく、そのアウトプットを利用して政策形成する立場にある人は、ぜひ時間を作って読んでみてほしい。

EBPMへの関心の高まりから、行政業務データの利用可能性も広がってきており、喜ばしいことだと思う。しかし、得られた分析結果が政策に反映されなければエビデンスに基づく政策形成とは言えない。私たち研究者の側が積極的に行政に対して研究成果を還元するとともに、それを実際の政策形成にいかに活かしていくかが引き続き課題となるだろう。とりわけ、当初予期していたのとは異なるエビデンスが出た際に、それを受け止めて政策のほうを軌道修正していく姿勢が非常に重要になってくる。PBEM（Policy Based Evidence Making）と揶揄する言葉もあるように、EBPMと言いながら、実際には規定の政策路線に都合の良い結果だけを並べて、都合の悪い結果を無視することは、残念ながら国内外を問わずしばしば起こる。そのようなことを極力なくし、すぐに軌道修正をできるよう、政策に携わる人々のデータ分析に関するリテラシーを高めていくことが求められる。

【引用文献】

伊藤公一朗，2017.『データ分析の力　因果関係に迫る思考法』光文社新書

川口大司・正木祐輔，2022.「CREPEによるプロジェクト設立の背景とねらい」連載「行政データと実証経済学：東京大学CREPE自治体税務データ活用プロジェクトの実践」第1回，『経済セミナー』，日本評論社，2022年6・7月号.

川口大司・正木祐輔，2023.「自治体税務データ活用の課題と可能性」連載「行政データと実証経済学：東京大学CREPE自治体税務データ活用プロジェクトの実践」第8回，『経済セミナー』，日本評論社，2023年10・11月号.

近藤絢子・深井太洋，2023.「市町村税務データを用いた既婚女性の就労調整の分析」RIETI Discussion Paper 23-J-049.

中室牧子・津川友介，2017.『「原因と結果」の経済学』ダイヤモンド社

正木祐輔，2022.「プロジェクト実施における課題と解決のための工夫」連載「行政データと実証経済学：東京大学CREPE自治体税務データ活用プロジェクトの実践」第2回，『経済セミナー』，日本評論社，2022年8・9月号.

Fukai, Taiyo and Ayako Kondo, 2024. Access to Formal Childcare for Toddlers and Parental Employment and Earnings. Mimeo

Kleven, Henrik, Camille Landais and Gabriel Leite-Mariante, 2023. The Child Penalty Atlas. NBER Working Paper #31649.