EBPM推進のための自治体税務データ活用プロジェクト

令和4年（2022年）5月

趣旨

デジタル改革関連法の施行を見据え、複数の地方自治体及び東京大学政策評価研究教育センター（CREPE）が連携し、①政策現場におけるEBPM（evidence-based policy making）の推進と②アカデミアにおける実証研究の発展を目指す。

背景

- 地方行政のデジタル化・スマート自治体化が主要な政策課題に
- 国や自治体におけるEBPMの推進
- AI（人工知能）等のビッグデータの分析技術の発展
- 2000年代以降、行政データを用いた実証経済学が発展

目的

① 政策現場におけるEBPMの推進

- 必ずしもリソースに余裕がない自治体も含めて、最新の経済学やAIに基づくデータ分析を利用可能にする。
- 法学者・情報工学者の参画のもと、個人情報保護と両立したデータ利活用を実現

② アカデミアにおける実証研究の発展

- 匿名化されたミクロデータを基に研究を実施

概要

- 自治体が、匿名化された個人レベルの税務情報をCREPEに提供
- CREPEが、計量経済学の知見やAI（人工知能）を用いて税収予測等のデータ分析を行い、参加自治体にフィードバック
- CREPEが、提供されたデータに基づき学術研究を実施

応募要件

募集対象自治体

- 全ての市区町村（※1）

※1　2022年度は、都道府県の新規募集は行っていませんが、柔軟に対応可能ですので、御関心がある都道府県は御連絡ください。

応募要件

- 以下のデータを匿名化（※2）した上で、CREPEに提供できること。

※2　匿名化については、CREPEが匿名化のツールを配布するため、匿名加工技術の知識やプログラミングの知識は不要です。

必要データ（匿名）

少なくとも以下を含む（※3）個人単位のデータ（※4）

・年度
・宛名番号（※5）
・控除前の収入金額：営業等、農業、不動産、利子、配当、給与、給与（専従）、雑所得（公的年金）、雑所得（その他）
・控除の金額：配偶者控除、配偶者特別控除、扶養控除、医療費控除、基礎控除、控除合計
・所得合計
・住民税額：均等割、所得割、住民税額合計（利子割、配当割、株式等譲渡所得割を含んだもの）
・属性に関する項目：生年月、性別、世帯番号、続柄

給与、給与（専従）、雑所得（公的年金）**以外**については、控除後の所得金額があれば、控除前の収入金額がなくても応募可能とします。
その他、そもそもシステムで保持していないデータ項目がある場合は、御相談ください。（2022.6.1追記）

※3　ここに示す以外のデータ項目も、データ分析において有用である可能性があるため、できる限り幅広に御提供いただけると幸いです。
※4　納税の有無にかかわらず、当該自治体の全住民を含むもの。ただし、匿名化の過程において、ランダムサンプリングを行う場合もあります（p. 5参照）。
※5　自治体独自に各個人に一意に割り振った番号のことであり、「個人番号」いわゆる「マイナンバー」とは異なります。提供に当たっては、世帯番号とともに、ハッシュ化（p. 11Q5及び参考資料2「匿名化についての考え方（案）（公表版）」参照）していただきます。

- 提供いただいたデータに基づく研究結果の公表（※6）に同意すること。

※6　自治体の同意がない限り、自治体名を明らかにすることはありません。

参加費

- 無料

参加自治体に提供する予定の分析

- 参加いただいた自治体には、CREPEから以下の分析結果を提供する予定です。

①2023年度の個人住民税の税収予測
・計量経済学の知見及びAIを用いた予測を提供します。
・性別や年齢等の属性に応じた区分や、所得割・均等割等の区分に応じた予測を提供可能です。

②その他の分析結果（以下は一例）
※ これ以外にも、どのような分析が有用かなどのアイディアがあれば、応募書類において御提案ください。

（所得・格差関係）
・国全体あるいは他の自治体と比較した所得分布の特徴
・市区町村における、過去x年の格差トレンド
・マクロ経済変動（コロナ危機、金融危機など）と所得格差の関係
・コロナ禍において最も所得の変動があったのはどの層か
・所得格差はどの程度固定的であるか
・属性ごとの所得リスクを考慮すると、どのような再分配政策が望ましいか（どういった層をターゲットとすべきか。どの程度の規模の政策をどういった手段で行うのが良いか）
・政策実施にかかるコストをどのような手段で負担するのが望ましいか（所得税・消費税・保険料・公債など）

（雇用・家計関係）
・103万の壁、130万の壁など、制度の壁による就労調整の実態把握
・どのような層が就労調整しやすいか（夫の年収、家族構成、子供の年齢等）
・同居家族の存在が女性の労働供給にどう影響するか（介護、子育て等）
・子育て期の女性の労働供給の分析
・女性の出産前後での所得の変化
・出産前後での所得減が子供が成長するにつれて回復するか否か、女性個人の所得や世帯所得で見た格差が広がるか否か
・祖父母との同居によりどの程度緩和されるか

（人口増減・移動関係）
・どのような属性（年齢・性別・所得）の住民がどの程度転出入し、どのようなトレンドにあるか
・どのような世帯において子が生まれているか

できれば、税務担当課様以外にもこの資料を共有し、応募を御検討ください。

追加データによる追加分析

- 本プロジェクトの応募に当たっては、p. 2記載の税務データを御提供いただくことが要件となりますが、その他の福祉データや教育データ（匿名化したもの）も併せて御提供いただくことが可能であれば、以下のような追加分析も可能です。

  ※1 以下はあくまで一例

  ※2 匿名化やデータ間の紐づけについては、CREPEにおいてツールを配布いたします。

追加データの例	追加分析内容の例
保育所の利用の有無、入所選考の点数等	・保育所の拡充によりどの程度女性の就業が増えるか ・それによりどの程度の税収増が見込めるかのシミュレーション ・特に恩恵を受けるのはどういった層か（出産前の本人所得、世帯所得、家族構成）
学力調査結果、問題行動調査結果	・どのような世帯の子どもが学力面のリスクを抱えており、支援が必要か ・教育によって将来の所得にどのような影響があるか ・いじめや長期欠席などの問題行動のリスク要因と有効な予防策は何か ・学力と問題行動の間にはどのような関係があるか

- もし御関心があれば、個別に協議いたしますので、応募書類においてお知らせください。

  ※3 応募時において、追加データの提供を確約いただく必要はなく、個別協議の後、提供の可否を御判断ください。

できれば、税務担当課様以外にもこの資料を共有し、応募を御検討ください。

個人情報保護・匿名化

- 自治体様には、匿名化したデータをCREPEに提供いただくこととしています。本プロジェクトでは、自治体様の多様なニーズに応えるため、個人情報保護及び匿名化の専門家の監修の下、2つの匿名化手法を御用意しています（※1）。

  ※1 詳細は参考資料1「個人情報保護制度との関係」・参考資料2「匿名化についての考え方（案）（公表版）」参照。最終的に、どこまでの匿名化が必要十分かの判断は、各自治体における個人情報保護条例に照らし、各自治体において判断いただくこととなりますが、その際、これらの資料を参考にしてください。

  Ⅰ 簡易な匿名化（❶～❸）

  Ⅱ 高度な匿名化（⓪～⑦。デジタル社会形成整備法（令和3年法律第37号）施行後の匿名加工情報制度において求められる水準の匿名加工）

  ❶⓪ 氏名、個人番号（いわゆるマイナンバー）、住所等を事前に削除
  ① 特異な世帯の世帯番号の秘匿
  ❶② 宛名番号、世帯番号のハッシュ化
  ❷③ 宛名番号をキーにして、複数年のデータの結合
  ④ 所得、賦課額等のトップコーディング
  ❸⑤ 生年月日の月単位への丸め
  ⑥ 生年月、性別等について、「3-匿名性」を判定し、「3-匿名性」を満たさないものは、「3-匿名性」を満たすまで秘匿化
  ⑦ 各世帯番号（のハッシュ値）について、ランダムサンプリング

- Ⅱはより匿名性が高いものとなりますが、Ⅰの方がより質・精度の高い分析結果を御提供できます。本プロジェクトには、個人情報保護法上の匿名加工情報制度がそのまま適用されるわけではありませんので、自治体様におかれては、いずれの匿名化手法を取るか御判断ください。

- 匿名化方法について確認いただいた後、CREPEから自治体様に、R（※2）による匿名化プログラムを配布しますので、CREPEが配布するマニュアル（※3）に従って、自治体様のPCにおいて実行いただきます。

  ※2 学術的な統計分析に広く用いられている無料のプログラミング言語
  ※3 参考資料3「RStudioの利用ガイド」参照。作業量の参考としてください。

- 併せて、CREPEにおいても、個人識別行為の禁止等を定めた「自治体税務データ活用プロジェクトにおけるデータ取扱規則」、「自治体税務データ活用プロジェクトにおける安全管理措置等措置等に関する規程」を定めており（※4）、個人情報の管理に万全を尽くしています。

  ※4 http://www.crepe.e.u-tokyo.ac.jp/research/research20210624.html

想定スケジュール（案）

【新規自治体様】

2022年5月	東京大学	プロジェクトに参加したいと考える自治体を募集する。
6～7月	東京大学	協力自治体を選定する。
	協力自治体	保有する税務データのデータ項目・定義を東京大学に伝える。
	東京大学	協力自治体に、データの概要（データの形式・分布）を確認する。
7～10月	東京大学	匿名加工の具体的手順書を作成し、協力自治体に送付する。
	協力自治体	手順書に沿って税務データを匿名加工し、東京大学に送付する。（分析結果の提供時期について、2023年1月以降を希望する場合は、提供を受けたい2ヶ月前までにデータを東京大学に提供する。） ※ 必要に応じて、個人情報保護制度上の手続（例：審議会、契約）を踏む。
10～12月	東京大学	協力自治体からもらったデータを分析し、他自治体との比較も含め、詳細な分析結果を返す。(協力自治体の要請により、2023年1月以降の提供になる可能性もある)

※　以上のスケジュールは現段階の想定であり、進捗に応じて、今後、変わりうる。

想定スケジュール（案）

【継続自治体様】

2022年5月	東京大学	プロジェクトへの参加を継続したいと考える自治体を募集する。
5～6月	協力自治体	データ項目・定義を東京大学に送付するとともに、匿名加工・データ項目について方針を確認し、変更を希望する場合その旨を東京大学に伝える。（変更を希望しない場合(◆)へ）
	東京大学	協力自治体に、データの概要（データの形式・分布）を確認する。
6～8月	東京大学	匿名加工の具体的手順書を協力自治体に送付する。
	協力自治体	(◆)手順書に沿って税務データを匿名加工し、東京大学に送付する。（分析結果の提供時期について、2022年12月以降を希望する場合は、提供を受けたい2ヶ月前までにデータを東京大学に提供する。） ※ 必要に応じて、個人情報保護制度上の手続（例：審議会、契約）を踏む。
8～11月	東京大学	協力自治体から提供を受けたデータを分析し、他自治体との比較も含め、詳細な分析結果を返す。(協力自治体の要請により、2022年12月以降の提供になる可能性もある)

※　以上のスケジュールは現段階の想定であり、進捗に応じて、今後、変わりうる。
特に、協力自治体が昨年度と異なる匿名加工やデータ項目を希望する場合は、新規参加自治体と同じスケジュールになる可能性がある。

研究体制

総括班

川口 大司
北尾 早霧
近藤 絢子
古川 知志雄
正木 祐輔

所得リスク・格差班（税務データを活用した所得リスクと所得格差の分析）

○	北尾 早霧	東京大学大学院経済学研究科　教授 独立行政法人経済産業研究所　上席研究員（特任）
	鈴木 通雄	東北大学大学院経済学研究科　准教授
	山田 知明	明治大学商学部　教授

雇用・社会保障班（セーフティーネットと雇用・家庭）

○	近藤 絢子	東京大学社会科学研究所　教授
	深井 太洋	筑波大学人文社会系　助教 東京大学政策評価研究教育センター　招聘研究員

税理論・実験班（最適税制論理論モデルと徴税率フィールド実験）

○	古川 知志雄	横浜国立大学大学院国際社会科学研究院・経済学部　准教授 東京大学政策評価研究教育センター　招聘研究員
	別所 俊一郎	東京大学大学院経済学研究科　准教授

学術利用基盤整備班（個人情報保護、匿名化を含めた行政記録情報の学術利用基盤整備）

○	川口 大司	東京大学大学院経済学研究科　教授 東京大学公共政策大学院　副院長・教授 東京大学政策評価研究教育センター　前センター長
	佐藤 一郎	国立情報学研究所　教授
	宍戸 常寿	東京大学大学院法学政治学研究科　教授
	正木 祐輔	東京大学公共政策大学院　准教授 総務省大臣官房秘書課　課長補佐

※ 2022年5月時点

※ 「○」は班長

※ 教育データが得られた場合など、必要に応じて、その他の研究者も参画予定

応募方法

応募自治体から直接、「応募様式」を電子メールで提出してください。

提出先：　jichitai_data[at]e.u-tokyo.ac.jp

＊ “[at]”の部分を“@”に変えて送信してください。

提出期限：2022年6月21日（火）

※　都道府県が取りまとめる必要はありません。

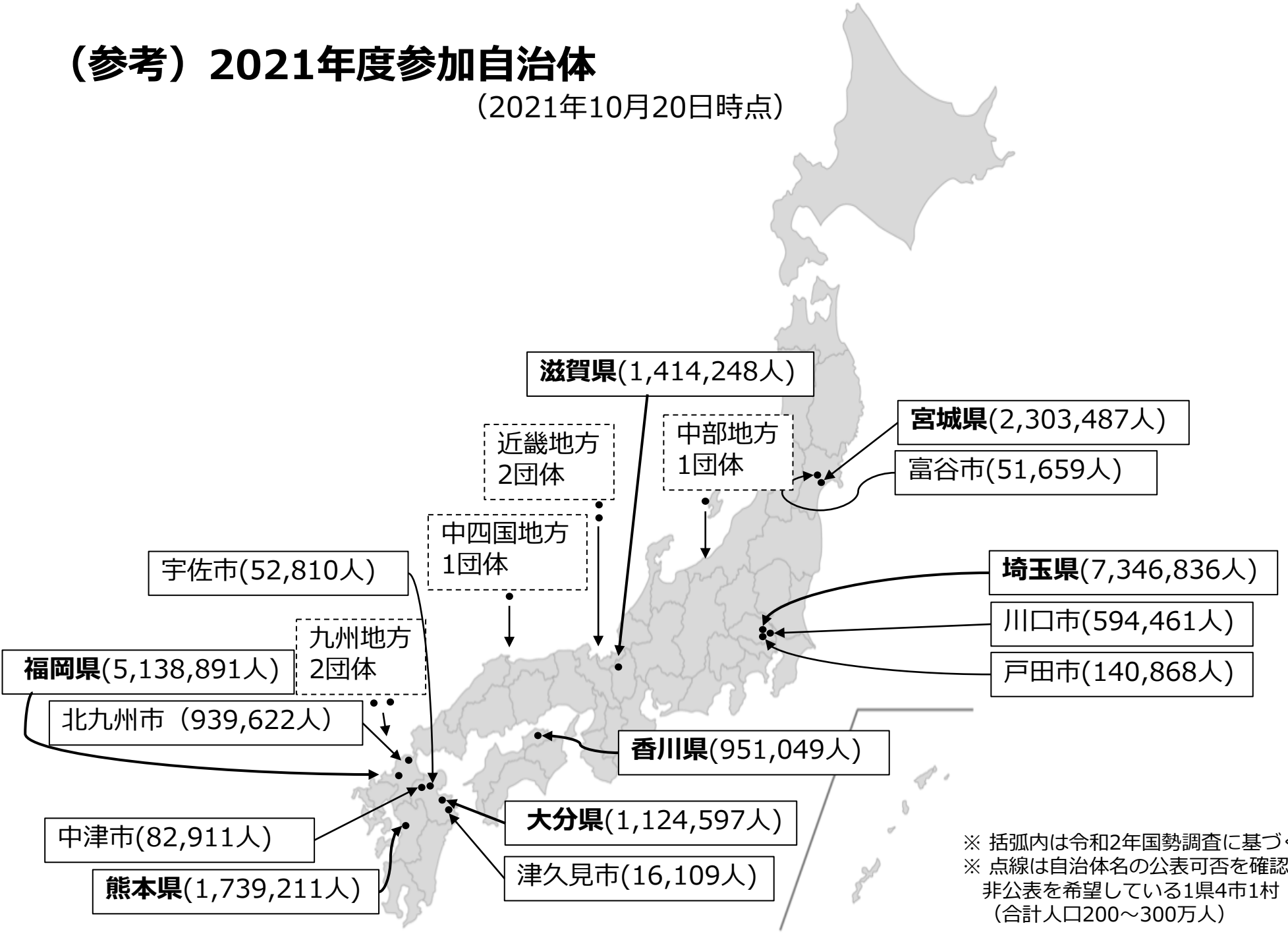
※　応募について御関心がある場合は、未確定の場合も、提出期限より前に早めにお知らせいただけますと助かります。

※　不明点、疑問点等あれば、遠慮なく問い合わせてください。

※　期限に間に合わない場合は、期限までに御相談ください。

（参考）2021年度参加自治体

（2021年10月20日時点）



※ 括弧内は令和2年国勢調査に基づく人口
※ 点線は自治体名の公表可否を確認中又は非公表を希望している1県4市1村（合計人口200～300万人）

Q&A (1)

全般

Q1　匿名化に関する知識は必要ですか。

A1　いいえ、必要ありません。個人情報保護及び匿名化の専門家の監修の下、デジタル社会形成整備法施行後の匿名加工情報制度において求められる水準の匿名加工の案を、CREPEから考え方とともに御提示します。

Q2　匿名化等のため、プログラミングのスキルは必要ですか。

A2　いいえ、必要ありません。匿名加工の案について自治体様の確認をいただいた後、CREPEが匿名化ツールを配布します。自治体様においては、CREPEが配布するマニュアルに従って匿名化ツールを実行いただくこととなります。

Q3　データ分析のスキルは必要ですか。

A3　いいえ、必要ありません。いただいた匿名化データをCREPEにおいて分析し、分析結果をお返しします。

Q4　税務データと世帯データ、普通徴収のデータと特別徴収のデータなど、自治体側でデータ間の紐付けを行うことが必要ですか。

A4　いいえ、必要ありません。紐付けのキーとなるデータ項目（宛名番号など）が入っていれば、CREPEにおいて紐付けします。

匿名化

Q5　紐付け等のために、宛名番号・世帯番号を提供する必要がありますか。

A5　宛名番号・世帯番号そのものを提供いただく必要はありません。自治体様において設定いただくパスワードに基づいてハッシュ化された宛名番号・世帯番号を提供いただくこととなります。

ハッシュ化とは、元のデータから一定の計算手順に従ってハッシュ値と呼ばれる規則性のない値を求め、その値によって元のデータを置き換えることにより、データを不可逆的に別の形に変える方法をいいます。

例としては以下のようなイメージです。

宛名番号	ハッシュ化した宛名番号
12345678	aks;ldfjpawnefdoiewadlksfdajf
23456789	k3298refcsna3489elsjodslerire
34567890	kjlz-9erj8ufdjiawe8re;ofjwloe

ここで、「aks;ldfjpawnefdoiewadlksfdajf」から、元の「12345678」を復元することは不可能です。そのため、宛名番号について言えば、当方が宛名番号を知りえないまま、年度をまたぐ個人の紐付けが可能となります。

ハッシュ化については、CREPEが配布する匿名化ツールに組み込まれているため、自治体様においては、パスワードを設定いただければハッシュ化を行うことができます。

Q&A (2)

自治体側の負担

Q6　参加に当たって、費用負担（CREPEへの委託料など）は必要ですか。

A6　いいえ、必要ありません。ただし、システムからデータを抽出する際にシステムベンダに依頼しなければならない場合などに費用が発生する可能性があります。

Q7　自治体側にどれくらいの作業負担が発生しますか。

A7　2021年度に参加いただいた自治体の作業時間の中央値は12時間ですが、50時間を超える自治体も複数あり、自治体によって大きな差がありました。

2022年度は、2021年度に作業時間が長い自治体からヒアリングを行い、自治体様の作業負担軽減に取り組みます。

Q8　データ抽出作業にはどの程度の期間を要しますか。

A8　自治体によってまちまちですが、ベンダへの依頼を行う場合、CREPEから自治体への匿名化ツール配布後、自治体様がデータを抽出し、CREPEに提供するまでに1か月～3か月を要する自治体がほとんどでした。自治体様の負担軽減に向けて当方も可能な限りの支援を行いますが、早めの御準備をいただけますとスケジュールどおりの御提供ができるかと思います。

税収予測の精度

Q9　税収予測の精度はどれくらいですか。

A9　提供いただけるデータの数（人口）、年数等によりますが、2021年度はほとんどの自治体で平均誤差率は1–2%でした（個人住民税）。

2022年度は、更なる精度向上を目指しています。

データ内容・形式

Q10　現在、システムベンダごとにデータ構造・データ形式が異なっていますが、提供するデータのデータ構造・データ形式の指定はありますか。

A10　ありません。データ構造・データ形式の差異については、CREPE側で対応するので、心配いただく必要はありません。

具体的には、参加決定後、データ構造をお聞きし、それに基づいてCREPEにおいて自治体様のデータ構造に合わせた匿名化ツールを配布します。2021年度も、異なる7社の税務システムの自治体様からデータをいただきました。

Q11　何年分のデータが必要ですか。

A11　少なくとも5年分のデータは御提供いただいております。なお、機械学習を用いた予測を行うため、可能な限り多くの年度分のデータをご提供いただけますと予測精度の向上が期待できます。

Q&A (3)

PCスペック等

Q11　匿名化を行うPCのスペックはどの程度必要ですか。

A11　自治体様の人口規模によって変動がありますが、昨年度ご参加いただいた人口10万程度の自治体様でプロセッサ：Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz、メモリ：16.0 GB のPCをお使いになった場合はデータ処理に大きな問題はなかったようです。

希望があれば、メモリ16GBの小型PC（モニタ・キーボード・マウスは自治体様に御用意いただきます。）を貸し出すことを検討しております。

Q12　匿名化を行うPCはインターネットに接続できる必要がありますか。

A12　インターネットに接続できる場合、匿名化に用いるプログラムであるRのインストールやアップデートがオンラインでできますが、インターネットに接続できなくても、必要なプログラムはCREPEが配布するため、問題ありません。

Q13　匿名化データの提出方法はどうなりますか。

A13　ファイル転送システム、記録媒体の郵送など、自治体様の都合の良い方法で結構です。

Q14　PCに外部のソフトウェアをインストールすることができません。匿名化にRを用いることは必須ですか。

A14　匿名化に要する時間をできるだけ減らすため、Rの使用を強くお勧めしております。

自治体様によって必要とする匿名化の水準は異なるため一概には言えませんが、Rが利用できない場合、「3-匿名化」など一部の高度な匿名化手法が利用できなくなる可能性が高いため、氏名を落とすなどの簡易な匿名化でCREPEにデータを提供することに同意いただく必要があります。

契約・協定

Q15　契約や協定は必要ですか。

A15　必ずしも必要ないと考えていますが、必要があると考える自治体は、個別に御相談ください。