DP/24-5

経済財政分析ディスカッション・ペーパー

銀行口座リアルタイムデータを利用した法人企業経済動向分析の手法

栗山　博雅・岩上　順子・酒巻　哲朗

Economic Research Bureau

CABINET OFFICE

内閣府政策統括官（経済財政分析担当）付

目次

銀行口座リアルタイムデータを利用した法人企業経済動向分析の手法 *

栗山　博雅 †・岩上　順子 ‡・酒巻　哲朗 §

【要旨】

オルタナティブデータとは、クレジットカードデータや携帯電話の位置情報データなど、従来の公的統計等とは異なる情報源や入手経路を通じて、新たに利用可能となったデータであり、その即時性の高さなどから近年活用の機運が高まっている。

今回、内閣府事業において、個社が特定されないように配慮しながら、都市銀行の銀行口座データを用いて法人企業に関する経済動向分析を行った。法人企業に関する代表的な公的統計である法人企業統計調査に対して、銀行口座データは高い速報性が期待でき、また含まれる中小企業のサンプルサイズが大きいなど、既存の公的統計と異なった特徴を持ち、企業の経済動向を把握するために活用できる可能性がある。

本稿では、こうした銀行口座データの活用に向け、まず、今回の銀行口座データに含まれるサンプル企業の特徴を確認した。企業の所在地が東京都に集中し、中小企業の中でも資本金や従業員規模が大きい企業のサンプルが多いが、業種別分布に関しては公的統計との類似が見られるなど、我が国の企業分布と比べた際の特徴を指摘した。

次に、銀行口座データを用いて各種指標を再現するにあたり、仕訳ルールを設定し、銀行口座による財務データの再現度合いを確かめ、銀行口座データを用いて業績を再現できる企業が全体の中では数少ないことを把握した。

こうした状況を踏まえ、再現度合いの向上を目指し、複数の条件を用いた絞り込み、及び機械学習手法を利用した企業サンプルの絞り込みを行った。

その上で、絞り込みを行った銀行口座データによる再現値と法人企業統計を始めとする公的統計の比較を行い、速報性の高い銀行口座データを用いて、公的統計の売上高や人件費の動向を、業種別の数値を含めて一定の精度で把握できることを確認した。

本稿での成果は、今後銀行口座データを用いた経済動向把握や、政策課題分析を進めていくうえで、分析手法や結果の解釈についての基礎的な知見となることが期待される。

---

* 本稿は、「令和５年度『リアルタイムデータを活用した経済動向分析（法人銀行口座データ活用)』」事業（東京大学エコノミックコンサルティング株式会社への委託調査）において得られた成果の一部を報告するものである。本稿において結果を利用した銀行口座データの分析に当たっては、一橋大学の植杉威一郎教授、東北大学の久保田荘准教授、早稲田大学の遠山祐太准教授、宮川大介教授、慶應義塾大学の星野崇宏教授、東京大学の渡辺安虎教授にご指導いただいた。また、本稿の作成においては、内閣府政策統括官（経済財政分析担当）の林伴子氏、内閣府政策統括官（経済財政分析担当）付参事官（企画担当）付調査官の石井一正氏から有益なコメントを頂いた。ここに記して感謝を申し上げる。ただし、本稿に残された誤りは筆者の責に帰すものである。本稿で示された見解は筆者の個人的なものであり、必ずしも内閣府の見解を示すものではない。

† 内閣府政策統括官（経済財政分析担当）付参事官（企画担当）付内閣府事務官

‡ 内閣府政策統括官（経済財政分析担当）付参事官（企画担当）付参事官補佐

§ 内閣府経済動向特別分析官（政策統括官（経済財政分析担当）付）

１．　はじめに

（１）　分析の目的

本稿では、リアルタイムデータを用いた経済動向分析の一環として、法人企業の銀行口座データを用いたリアルタイムデータ分析の手法を紹介し、政府統計である法人企業統計等と比較した際の特徴を考察する。

近年、特に経済・社会状況が刻一刻と変化するコロナ禍の経験以降、既存の公的統計とは異なる情報元を用いた、即時性の高い「オルタナティブデータ」活用の機運が高まっている。クレジットカード情報を利用した消費データや、スマートフォンの位置情報に基づく人流データなどのオルタナティブデータの多くは、統計作成のためのデータ収集・集計といったプロセスを取る既存統計と比較して、データが利用可能になるまでの時間が短く、速報性が高いリアルタイムデータであるといった特徴がある[1]。また、アンケート調査や、標本調査の形をとることの多い既存の統計と比べ、週次や日次単位でのデータや、ピンポイントな位置情報データ、特定のサービスやアプリのユーザー全数のデータを用いた詳細な属性の分析など、細かい粒度のデータを利用できるといった点においても、従来の公的統計では把握できなかった経済動向を観察できる可能性がある。内閣府でも、携帯電話の位置データを利用した「モバイルビッグデータ」[2]、家計の出納や資産管理に用いられる「家計簿アプリデータ」[3]、賃金や賞与といった給与の計算に用いられる「給与計算代行サービスデータ」[4]を用いたリアルタイムデータの活用を行ってきた。

今回の法人企業の銀行口座データを用いた分析は、企業動向に関する公的統計、特に法人企業統計の公表は四半期毎かつ公表まで2か月以上要する[5]といった性質を持つため、企業動向に関する分析の即時性の向上につながると考えられる。

そのため、本稿では公的統計とリアルタイムデータの比較という点に着目し、政策分析を行うに当たっての前提となる、銀行口座データを用いたリアルタイムデータ分析におけるデータ処理方法の解説、法人企業統計の再現性を確認することによる即時的な分析への活用可能性の検証について、これまでの作業で明らかになった点を紹介する。

リアルタイムデータを活用した分析が考えられる一例として、コロナ禍においては、法人企業、特に中小企業に対して、持続化給付金の給付や雇用調整助成金制度の拡大、時短協力金の給付、実質無利子・無担保融資（いわゆるゼロゼロ融資）といった様々な支援策が行われたが、本稿において扱う法人企業の銀行口座データの特徴を踏まえ、これらの政

[1] 大久保他（2022）
[2] 井上他（2019）
[3] 内閣府政策統括官（経済財政分析担当）（2023）、小林・鈴木（2022、2023a）、小林他（2023b）
[4] 都竹他（2024）
[5] GDP 四半期別速報の１次速報（１次 QE）に法人企業統計を用いることを主目的として、附帯調査による法人企業統計の一部早期化の試みが行われていた（財務省財務総合研究所（2020、2021、2022））。

策効果をリアルタイムデータを用いて分析することで[6]、コロナ禍における大規模な支援策の効果を検証することが可能となる。これにより、将来に類似の大規模支援が政策として求められた際に、リアルタイムデータを用い、最新の情報を効果的な政策立案に役立てることが出来るようになると考えられるなど、今後の更なる研究が期待される。

（２）　分析の手順

本節では、具体的な分析のプロセスについて概説する。

銀行口座データは、銀行に口座を保有する企業の日々の入出金の記録である。銀行は取引企業の属性情報や与信先を中心に財務諸表の情報も保有しており、これらも分析に活用する。以下では、銀行口座に付属する企業の属性情報と日々の入出金の記録を「銀行口座データ」、銀行が保有する取引企業の財務諸表の情報を「財務データ」、入出金の記録から推計した財務情報を「再現データ」と呼ぶ。今回の法人企業銀行口座データの分析に当たって、主に①法人企業銀行口座データの基本的な性質の確認、②業績等に関するマクロ的動向の確認を行った。①法人口座データの基本的な性質の確認においては、イ）サンプルに含まれる中小企業の属性が日本の中小企業全体と比べて偏りがないか、企業数や財務データの分布を経済センサスと比較する[7]、ロ）銀行口座データにより企業業績等に関する諸指標（売上高・費用・営業利益・人件費・従業員数等）を再現し、財務データと比較する、ハ）前項で財務データの再現比率（再現データ÷財務データ）が低いサンプルが多かったため、より当てはまりの状況が良い企業にサンプルを絞り込み、財務データの再現度合の向上を図る、といったプロセスを取った。

②業績等に関するマクロ的動向の確認では、イ）サンプル全体や絞り込みによって再現度合が向上したサンプルについて諸指標を集計し、法人企業統計と比較し、またロ）業種や指標を絞って鉱工業生産指数・商業動態統計・総雇用者所得との比較を行った。

なお、コロナ禍における各種支援策の効果等にかかる実証分析は今後の検討対象となるが、①ハ）において行った絞り込みにより、法人企業統計を一定程度再現できていることが②イ）において確認できたため、①ハ）にて絞り込んだサンプルを対象に分析を行うこととしている。

---

[6] コロナ禍における企業を対象とした政策効果の分析の一例として、Kawaguchi et al.(2023) による経営者を対象にしたアンケート調査による分析があり、持続化給付金の受給により「企業の存続確率」に対する経営者の見通しや実際の存続確率が上昇すると分析しており、中小企業庁（2022）でも分析の概要が取り上げられている。但し、Kawaguchi et al.(2023)内でも指摘されているように、経営者の見通しは事後的に確かめられた実際の存続確率に比べて過度に悲観的であり、リアルタイムデータの活用は、こうしたアンケート調査によるバイアスを補完する役割も果たしうると考えられる。

[7] 本稿では基本的に法人企業を対象に分析を行ったが、①イについては、個人事業主も対象とした。

２．先行研究

この章においては、オルタナティブデータを用いた経済動向分析の事例、及び銀行口座データを活用した事例を紹介する。

オルタナティブデータには様々な種類があるが、内閣府においても複数のデータを用いた分析により、その有用性の検証を進めてきている。井上他（2019）では、モバイルビックデータによる位置データを用い、オフィス街及び繁華街・住宅街の夜間の滞在人口を分析し、若年層や男性を中心とした残業時間の減少の可能性や、中高年層を中心とした朝方シフトの動きといった働き方改革の進展について分析した。

「家計簿アプリデータ」の活用については、消費動向に関する一連の分析がある。小林・鈴木（2022）では、家計簿アプリデータの特徴や分析に当たっての留意点について論じ、アンケート調査の併用による家計簿アプリデータの属性の偏りや収入・支出の補足可能性の検討、前月比や前年同月比といった変動を中心にした公的統計との比較を行い、小林・鈴木（2023a）ではログイン履歴を用いて口座連携の十分性の高いサンプルへの絞り込みの手法や、日時データの観察、月内や週内における数値の周期性の検討を行った。こうした検討結果を踏まえ、内閣府政策統括官（経済財政分析担当）（2023）では家計簿アプリデータを用いて、コロナ禍における特別定額給付金が家計消費に与えた影響を分析した。給付金支給の５週間前から 10 週間後までの期間において、累積で給付額の 22%程度が消費増加効果として計測されたこと、特に低所得世帯において比較的大きな効果（等価所得の下位３分の１のグループで 32%程度の消費増加効果）が観測されたことが分析結果として得られた。当該推計結果は、並行して実施した家計調査を用いた消費増加効果の推計結果（17%）とおおむね整合的であり、リアルタイムデータを用いた分析にある程度の頑健性が見られることも明らかにした。また、小林他（2023b）では、家計簿アプリデータを用い、子育て世帯への臨時特別給付の消費増加効果の計測を試みており、一定の有意な累積の消費増加効果が観測された。

また、鈴木・森（2023）では、クレジットカードデータを用いた個人消費動向把握の手法について検討し、クレジットカードデータが公的統計による消費支出全体の動向を相応の精度で捉えていること、財・サービス別に見ると財支出に関しては公的統計との相関係数が低くなることを明らかにした。その上で、財については一部の業態の販売額をＰＯＳデータ等で置き換えることで精度が向上し、消費支出全体としても公的統計の変動との相関係数が高まり、ＲＭＳＥが低減するなど、パフォーマンスの向上が確認された。更に、クレジットカードデータに基づく消費データを一定の仮定の下で実質化することを試み、公的統計の実質値の変動を概ね捉えていることを明らかにした。

都竹他（2024）では、「給与計算代行サービスデータ」の活用に向けたマニュアル的な位置づけとして、データの項目の確認、前処理を行った上でのデータセットの構築、サンプルの代表性についての検討、賃金や総労働時間に関する公的統計との比較を行い、水準

や変動に関して公的統計との一定程度の整合性を確認し、分析を通じて認識されたデータの強みや政策課題への効果的な分析の可能性について指摘した。

こうした検討の成果は、クレジットカードデータの「月例経済報告等に関する関係閣僚会議資料（令和５年４月 25 日等)」への活用、給与計算データの「月例経済報告等に関する関係閣僚会議資料（令和６年５月 27 日）・（令和６年６月 27 日)」への活用等、政府による経済動向分析の場面で活用されている。

銀行口座データを我が国における経済分析 [8]に活用した例としては、Kubota et al.（2021）のコロナ禍における特別給付金の分析が挙げられる。当該研究においては、みずほ銀行の銀行口座データから、コロナ禍における個人を対象とした特別給付金の入金のタイミングと消費の増加について分析し、入金から１週間以内において支出の増加がみられたこと、また入金から１か月以上に渡って緩やかな支出水準の高止まりが見られたことを明らかにした。更に、受給者の経済的な状況も分析対象とし、総資産額よりも流動資産額が消費性向に影響を与える可能性を示唆している。

３．データ

（１）　使用したデータの概要

本節では、実際に分析対象としたデータセットの概要及び特徴について記載する。

当分析では、みずほ銀行が保有している銀行口座データを用い、所在地、業種、従業員数、資本金額、設立年月日、上場区分、口座残高といった属性情報を保有している法人企業約 62 万社が含有されているデータセットを利用した。このうち、中小企業基本法上の定義 [9]に基づき、中小企業約 49 万社を抽出して財務データを結合した。主に与信先である約 13 万社 [10]に関しては、銀行が保有する最新財務データが利用可能なことに加え、信用調査会社（帝国データバンク）より基本情報データ約１万２千件、財務情報データ約８千件を購入して財務データを補完した。

---

[8] 海外において個人を対象とした銀行口座データを活用した事例としては、コロナ禍において、アメリカにおける当初の幅広い所得階層における急速な消費の落ち込みと、その後の低所得層における速やかな消費の回復を明らかにし、給付金と失業保険の拡充を中心とした”American Rescue Plan”が労働市場の落ち込みの中で消費を下支えした可能性を示唆した Cox et al.(2020)、取引データから American Rescue Plan 内の給付金に対する消費性向が 46%であり、加えて 10%が負債の返済に回ったと推計した上で、貯蓄がほとんどなく当期の給与に頼った生活している層の消費性向は 60%に上ったと推計した Karger & Rajan (2020)、スペインにおいて休業要請は消費への影響が大きかったが、人数制限の影響は比較的小さかったこと、高所得者が多い地域における支出額が急減したこと、低所得層が多い地域において出勤等のための人の移動が多かったことを明らかにした Carvalho et al. (2021)などがある。

[9] 業種分類において「製造業その他」においては資本金３億円以下または従業員数 300 人以下、卸売業においては資本金１億円以下または従業員数 100 人以下、小売業においては資本金５千万円以下または従業員数 50 人以下、サービス業に関しては資本金５千万円以下または従業員数 100 人以下（中小企業庁（2024))。「または」が条件となっているので、後述するように資本金１億円以上の中小企業も一定数存在する。

[10] 財務データが使用可能な約 14 万社のうち、今回分析対象とする中小企業として分類された企業が約 13 万社となった。

また、法人企業に加え、個人事業主に関しても個人事業主約 102 万人の銀行口座データから、最新財務データを結合できる約 1 万８千社を抽出した[11]。分析においては個社が特定されないように配慮している。

（２）　銀行口座データの性質

銀行口座データは、文字通り企業が保有する銀行口座のデータであり、企業の取引の摘要、取引日時、取引金額等を入手することができる。ある企業の取引が１つの銀行における口座で完結しており、取引を財務諸表の項目に適切に仕訳できる場合、理論的には銀行口座データから、企業の売上・費用及びそこから算出できる利益、人件費や従業員数といったデータを把握できることになる。

銀行口座データは、他のオルタナティブデータと同様、速報性が高い、サンプルサイズが大きい、記録が自動的で記入漏れ等のリスクが低い[12]といった点で、公的統計や企業によって公開される財務情報等に対し優位な点を持っていると考えられる。

他方、他のリアルタイムデータと同様に、サンプルの分布が偏っている可能性がある。今回利用する銀行口座データが大手都市銀行であるみずほ銀行のものであることを考えると、サンプルが比較的規模の大きい、大都市圏の法人に偏る可能性もあり、（３）でサンプルの偏りを検証する。

銀行口座データから財務諸表の情報を再現する際には、個別の取引を財務諸表のどの項目に分類するかといった仕訳ルールが必要になる。具体的には財・サービスの販売や購入、人件費など売上や費用に該当する項目と、借入や返済などの金融取引、税の支払いや補助金の受取などを区別する必要がある。その際、同一企業間の入出金の扱い、摘要欄が空欄の取引をどのように扱うか、といった点も検討すべき課題となる。こうした財務諸表の再現にあたっての仕訳ルールについては４．（１）にて検討する。

更に、企業の多くが複数口座を使い分けている場合、銀行口座データで再現できる売上や費用、人件費といった項目が過小となる恐れがある。例えば、ある企業が給与等の人件費の支払をみずほ銀行以外の銀行口座で行っている場合、この企業の人件費は０と推定されてしまうが、これは明らかに実態に即していない。こうしたみずほ銀行で捕捉出来る企業の活動に限りがあるという点については、口座の利用状況からみずほ銀行を主要な取引銀行としていると判断される企業にサンプルを絞ることなどで対応することが考えられる。

---

[11] 後述するように、個人事業主に関しては企業と比較しても更に所在地や業種の偏りが激しく、今回は分析対象としては用いなかった。

[12] ただし、会計処理の様式がある程度共通しており、監査等の対象となりうる企業の財務情報は、家計の情報に比べて一般的により正確性が高いと考えられており、この点ではリアルタイムデータの相対的な優位性は低くなると考えられる。例えば、家計を対象にした総務省「家計調査」について、宇南山（2011）では「『繰り返しクロスセクション統計』としては国際的にも最高水準の統計である」と評価しつつも、「記入の必要性に関して誤解の余地があり記入漏れが生じる可能性」があると述べており、リアルタイムデータの活用によって、家計調査の記入漏れによる誤差を補正できる可能性がある。一方、企業の財務情報に関しては、こうした家計の情報に比べると記入漏れ等は元々少なく、記入漏れの補正という形に関しては、リアルタイムデータが活用できる余地は相対的には少ないと考えられる。

具体的な方法は４．（３）にて検討する。

（３）　銀行口座データに含まれる企業サンプルの特徴

本節では、銀行口座データに含まれる企業サンプルの基本的な統計量の特徴を、経済センサスと比較する。以下の節で示すように、経済センサスと比較して銀行口座データ内の企業サンプルには、①業種別の分布は似通っており、②所在地別の分布は東京都が多く、③資本金規模別で見ると規模の大きい企業が多く、④従業員規模別で見た場合も規模が大きい企業が多く、⑤個人企業の分布は法人企業と比較しても偏りが顕著、といった特徴があった。

比較の手法としては、属性（業種/都道府県/資本金級数/従業規模）別の分布を経済センサスと銀行口座データに含まれる企業のそれぞれについて作成して比較するとともに、参考として、属性別の分布に対してカルバック・ライブラー・ダイバージェンス（KL 距離）を用いて、両分布の近さを定量化した[13]。

留意点として、銀行口座データの企業サンプルは中小企業基本法の定義に基づき、資本金額及び従業員数の双方を中小企業の判定に用いて抽出しているが、経済センサスや法人企業統計では、同様の基準で中小企業を抽出することができず、比較する範囲を厳密に合わせることはできない。そのため、業種別の企業数分布及び売上高分布の比較においては両者のサンプルを資本金１億円未満（サービス業、複合サービス業では資本金 5000 万円未満 [14]）の企業に絞り、資本金額において中小企業の要件に近いサンプル同士の比較となるようにした。

また、経済センサスのデータは「令和３年経済センサス－活動調査」に基づき 2020 年のデータとなっているが、財務データは最新の事業年度を用いており、異なる事業年度の集計値であることに留意が必要である。企業毎に入手できる最新の事業年度が異なり、2023、2022 年といった近年のデータが多いが、2015 年以前のデータも一定数存在する [15]。

---

[13] カルバック・ライブラー・情報量は、P(i)と Q(i)を各分布におけるビン i の割合とした場合、

$$KL(P||Q) = \Sigma_i P(i)\log\frac{P(i)}{Q(i)}$$

として定義される、「真の」分布 P と予測値 Q の近さを定量化した指標であり、本節においては経済センサスのデータを真の分布 P、銀行口座データを Q として算出し、対数の底は e とした。KL 距離は常に０以上となり、P と Q の分布が全てにおいて等しい時（業種別分布の場合、センサスと銀行口座データにおいて各業種の企業の割合が全く同じ場合）、KL 距離は０となる。相関係数等と異なり、値の絶対量によって分布が「近い」「遠い」と判断することはできないが、後述のように、従業員規模別分布を全国と東京都で算出した際に、後者の KL 距離が前者の KL 距離よりも小さくなったというように、相対的な分布の近さを判断するに当たり用いることができる。留意点として、P と Q の KL 距離 KL(P||Q)と Q と P の KL 距離 KL(Q||P)は異なる値を取る（Raschka and Mirjalili (2017) を参考としたが、同著では機械学習に当たっての情報利得量（Information Gain）としての解説が主であり、KL 距離の概念は情報利得量の算出に用いる控除項の一つとなっている）。

[14] 前述のように中小企業法上では、サービス業に加え小売業における中小企業の資本金条件も 5000 万円以下だが、経済センサスにおいては「卸売業、小売業」として分類されるため、サービス業・複合サービス業以外では資本金１億円以下の企業に絞り比較した。

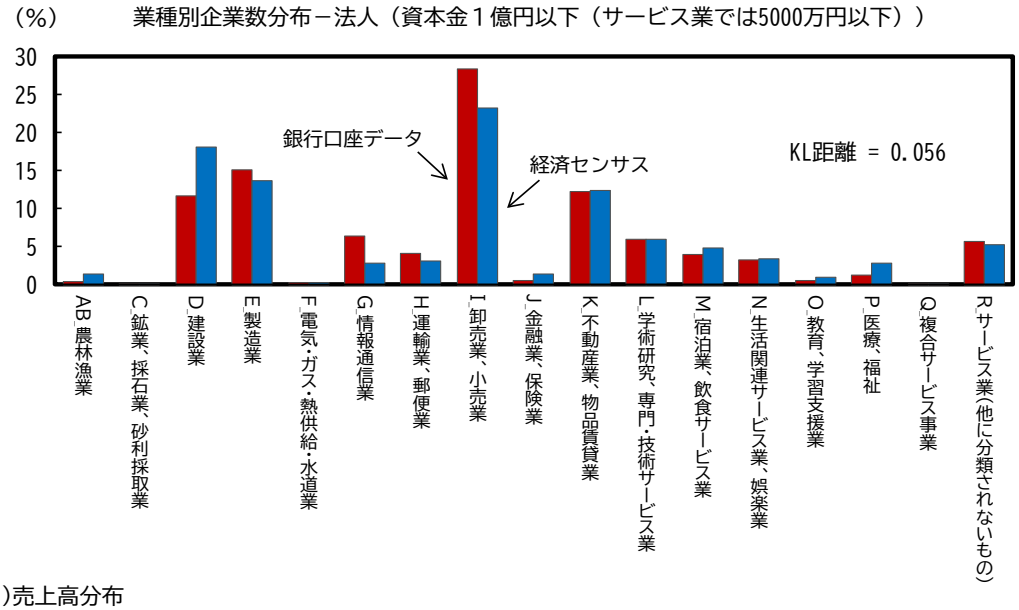
[15] 財務データを入手可能な中小企業約 13 万社のうち、最新の決算年として入手可能な決算年の分布は最

①　業種別企業数・売上高分布

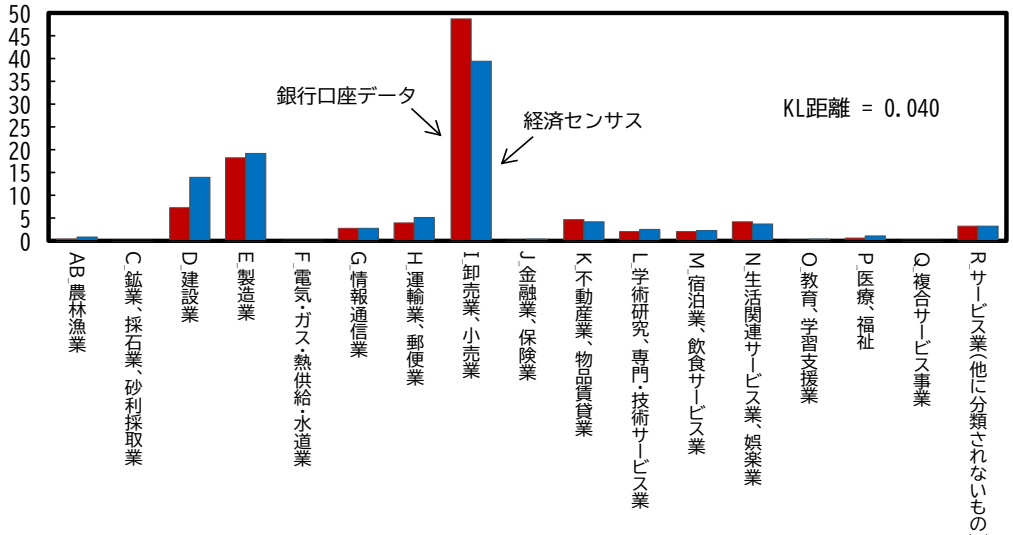
資本金１億円未満（サービス業、複合サービス業では 5000 万円未満）の法人に関して、銀行口座データに含まれる企業数の業種別分布と、経済センサスの業種別の企業数分布・売上高分布を比較する（図表３－３－１）。

（図表３－３－１　業種別企業数分布）

(1)企業数分布



(2)売上高分布



（備考）みずほ銀行が保有する財務データ及び総務省・経済産業省「令和３年経済センサス－活動調査」から作成。産業分類は「令和３年経済センサス―活動調査」の大分類に準拠。

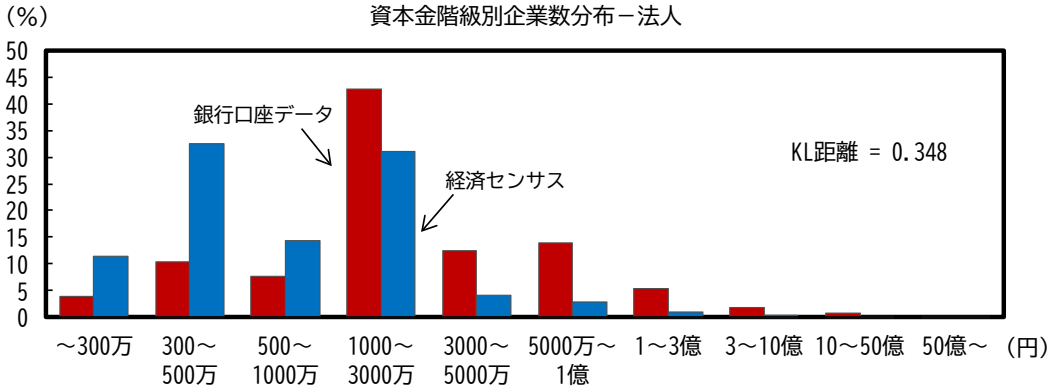
頻値が 2022 年で約４万２千社、次いで 2023 年の約１万３千社だが、最新の財務データが 2015 年以前となる企業も約２万４千社存在する。決算年数を数字と見立てて単純平均を取ると、2019 程度となる。

業種別の企業数割合の分布を見ると、経済センサスで 10%以上の割合を占める製造業や不動産業・物品賃貸業の割合がほぼ一致するなど、業種別の企業数分布は概ね似通っていると考えられる。主だった個々の業種を見ると、銀行口座データに含まれる企業の割合は、建設業でセンサスよりやや小さく、卸売業・小売業及び情報通信業においてやや大きいといった特徴がある。また、業種別売上高分布をみても、企業数分布と同様、センサスと分布の特徴が概ね似通っており、特に卸売業、小売業は売上高で見た場合、全体に占める割合が高くなる、といった特徴を捉えられていると評価できる。

②　資本金規模別の企業数分布

資本金に関しても、銀行口座データと、経済センサスを比較した（図表３－３－２）。

（図表３－３－２　資本金階級別企業数分布）

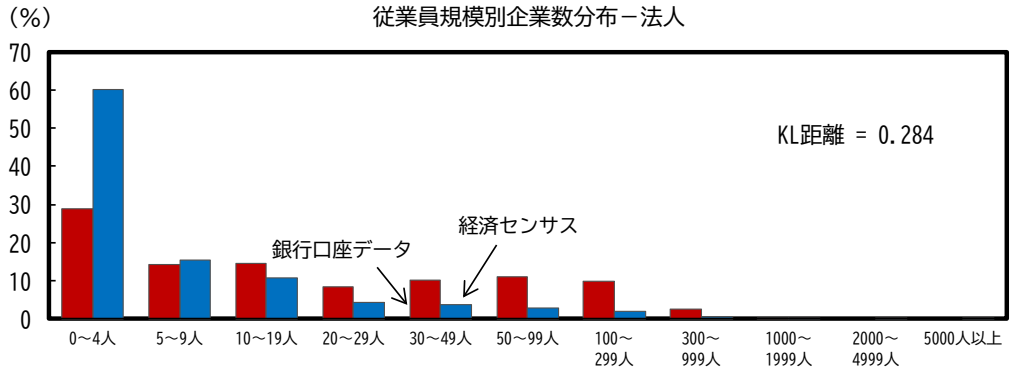


大手都市銀行のデータを用いていることもあり、経済センサスでは全体の 50%以上を占める資本金 1000 万円未満のサンプルが少なく、資本金 1000 万円～１億円未満程度の中小企業としては比較的資本金の額が大きいサンプルが多い。後述の従業員数等も考慮すると、法人企業銀行データにおいては比較的規模の大きい中小企業がサンプルとなっていると考えられる。

③　従業員規模別の企業数分布

従業員規模別の企業数の比較は下図（図表３－３－３）の通りとなった。

（図表３－３－３　従業員規模別企業数分布）

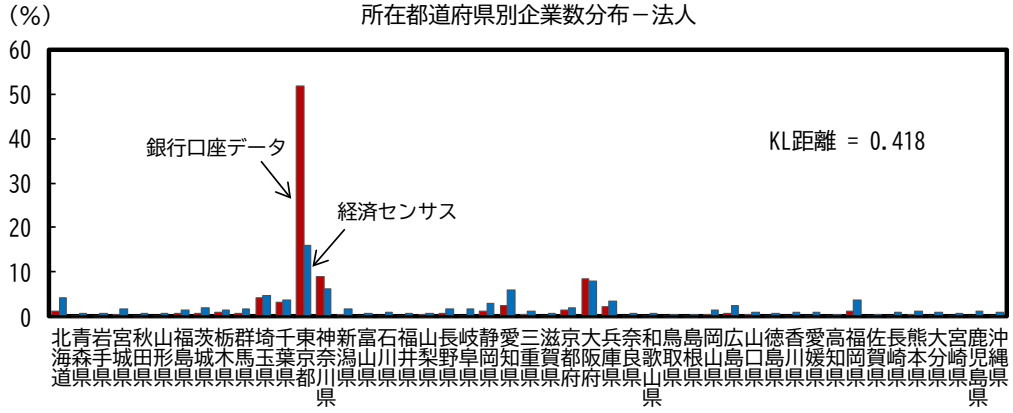


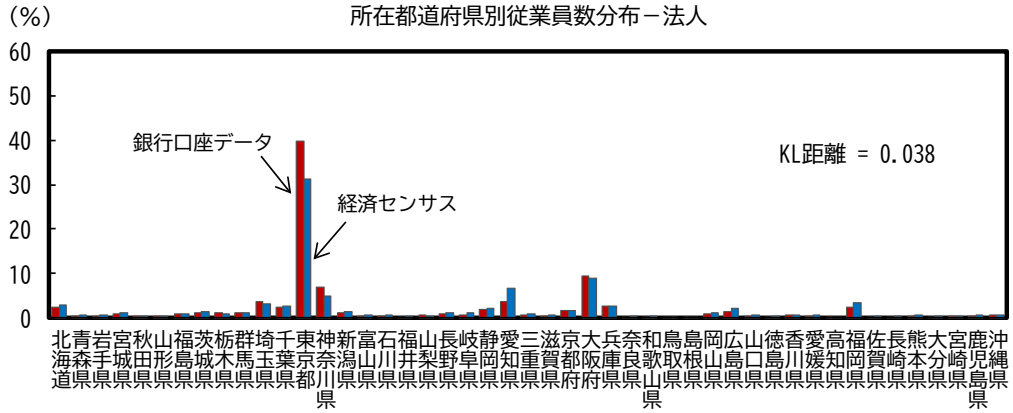
銀行口座データでは、経済センサスでは 60%程度を占める従業員０～４人の規模の企業の割合が相対的に小さく、従業員 30 人以上の比較的規模が大きい中小企業の割合が大きい。

④　所在地別分布

都道府県別の企業数分布、及び都道府県別の従業員数分布を経済センサスと比較すると下図（図表３－３－４）の通りとなった。

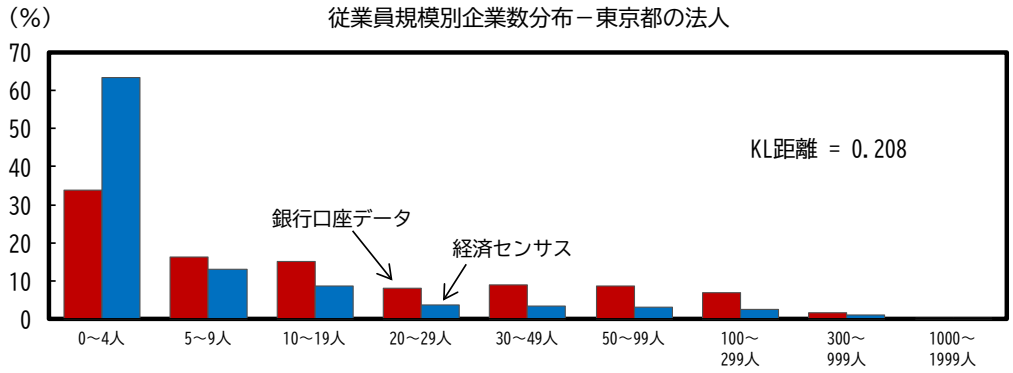
（図表３－３－４　所在都道府県別企業数及び従業員数の分布）





企業数でみると、東京都に所在する企業が多く、企業サンプルの所在地が東京都に偏っていることが分かる。一方、従業員数でみると、企業数ほどは東京都への集中が見られず、KL 距離を見ても、企業数分布の KL 距離が 0.418 であるのに対し、従業員数分布の KL 距離は 0.038 と、従業員数ではより経済センサスに近い分布となっていると考えられる。経済センサスによると、一企業当たりの従業員平均人数は全国平均が 25 人であるのに対し、東京都では 48 人のため、一般論として東京都に特段小規模企業が多いといった事象があるとは考えにくい。そこで、東京都に所在する企業に絞り、銀行口座データ及び経済センサスの従業員数規模別の企業数分布を比較すると下図（図表３－３－５）の通りとなった。

（図表３－３－５　東京都所在企業における従業員数規模別企業数分布）

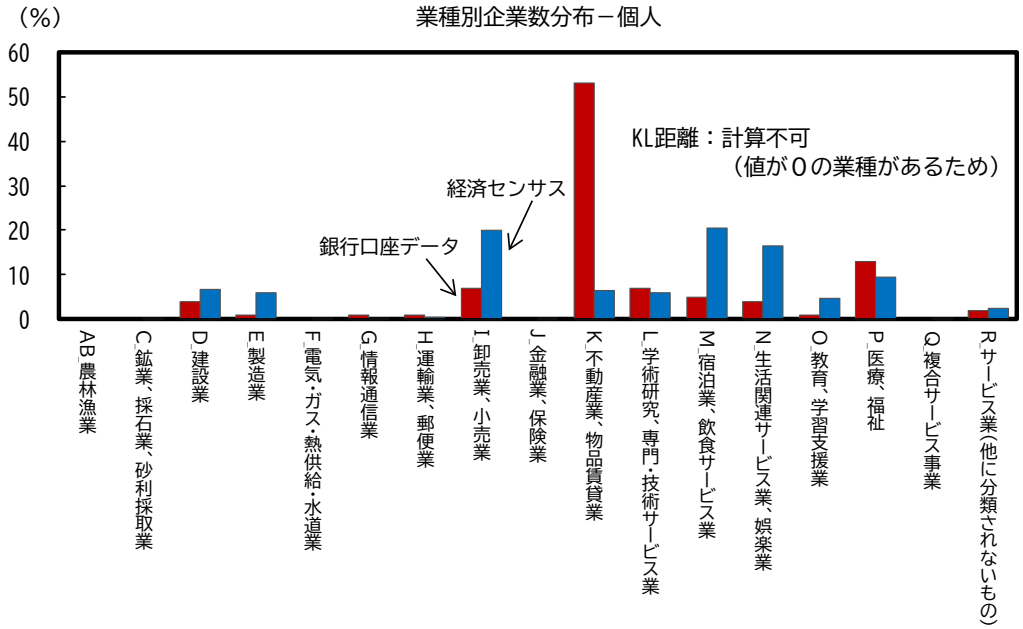


従業員数０～４人の企業の割合は、センサスと比較して銀行口座データでは依然として小さいが、全国でみると銀行口座データの０～４人の割合が 29%であったのに対し、東京都では 34%と割合が高い。KL 距離で見ても、全国の場合 0.284 であったのに対し、東京都の場合は 0.208 と、より近い分布となっている。東京都以外においては、銀行口座データのサンプル企業は比較的規模の大きな企業が多いことに対し、東京都では従業員数５人未満といった、規模の小さい企業が比較的多く銀行口座データに含まれている。

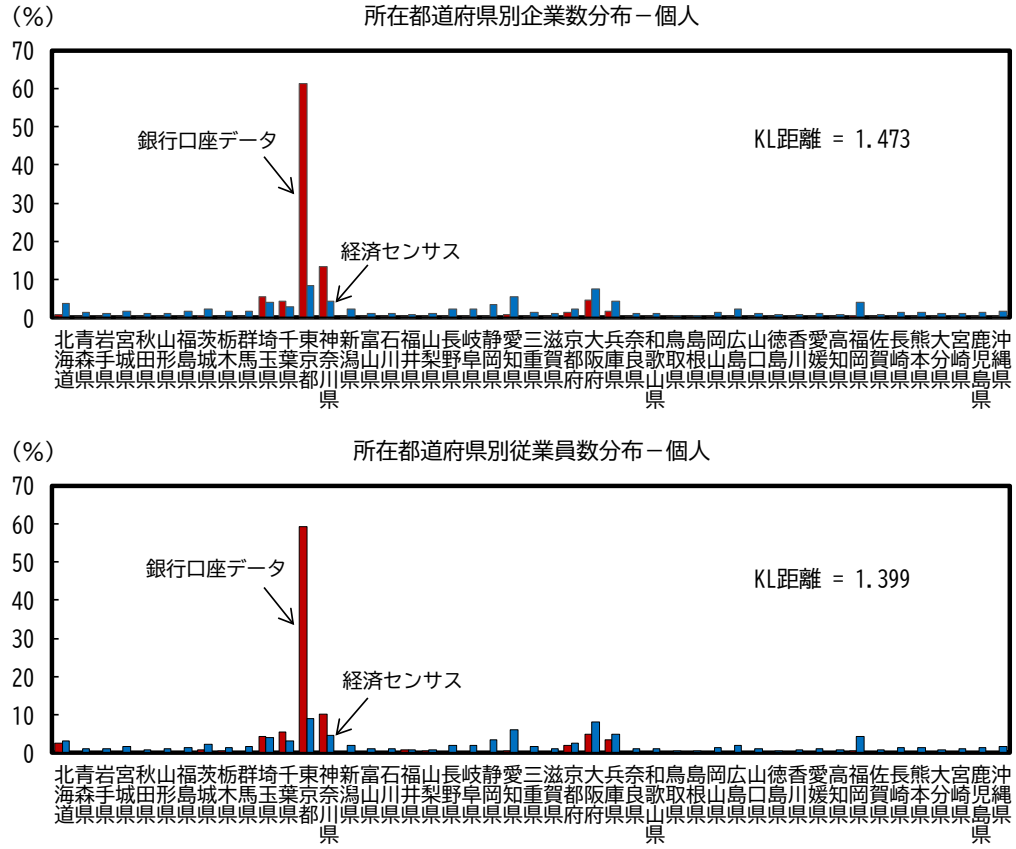
⑤　個人企業の分布について

ここまで法人企業に関する分布を確認してきたが、個人企業についても分布を確認すると以下の通りとなった。個人企業に関しては、不動産業・物品賃貸業への企業数の偏り（図表３－３－６）や、法人企業以上にセンサスと比べた際の東京都への集中（図表３－３－７）がみられる。財務データと組み合わせて比較することが可能なサンプルが限られる中で業種の偏りが見られることもあり、今回は公的統計との比較を行わなかったが、こうした個人企業の銀行口座データの偏りに留意した上でどう分析していくかは、今後の課題である。

（図表３－３－６　個人企業の業種別企業数分布）



（図表３－３－７　個人企業の都道府県別企業数分布・従業員数分布）



４．銀行口座データによる再現値と財務データの比較

前章では、銀行口座データに含まれる企業が、どの程度日本の企業全体の分布と似通っているかを見てきたが、本章では、オルタナティブデータである銀行口座データそのものが、どの程度口座を保有する企業の財務データを再現できているかという観点から、銀行口座データから年間の売上高・費用・営業利益・人件費・従業員数の再現を行い、財務データと比較する。まず、銀行口座データを利用し売上等を再現するに当たり、個別の取引に用いる仕訳ルール（例えば、同一企業間のやり取りを除く、摘要欄が「利息」「融資」「還付金」等となっている取引を除外する）について概説する。次に、銀行口座データを用いて全てのサンプル企業に対して売上高等を再現した再現データを財務データと比較した結果を示す。その上で、全てのサンプルでは再現度合いが著しく小さいものが多数になってしまうため、従業員数や人件費、売上といった項目の再現度合いが高い企業のみにサンプルの絞り込みを行い、再現度の高い 540 社に関する比較結果を示す。更に、前述の複数条件を用いた 540 社への絞り込みに追加して、機械学習の手法である「Isolation Forest」を用いた絞り込み手法を示す。

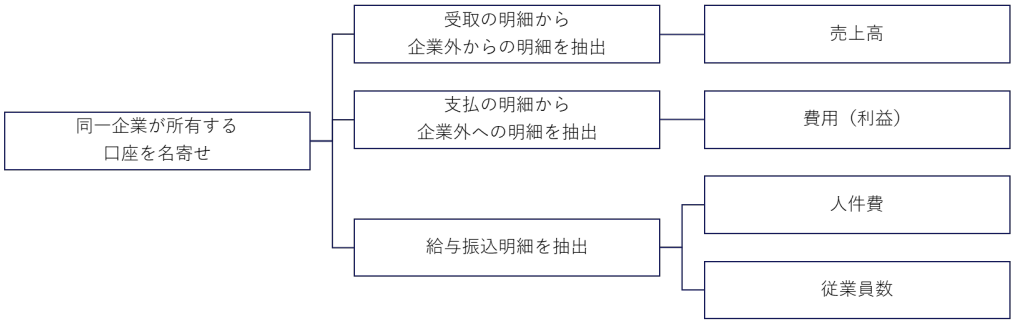
（１）　　仕訳ルールの設定

本分析に関しては、有識者の意見も踏まえ、下記及び図表４－１－１の仕訳ルールを設定し、財務情報の再現データを作成した。

①銀行口座データに紐づけられている「企業 ID」ごとに口座のやり取りを集約
②ルールベースで除外する取引を取り除く
- 税金等納付金（地方公金）を除外
- 更に、摘要欄で「利息」「融資」「還付金」「株配」「積立」「返済」「税金」等を含むものを除外
- 同一企業間のやりとりを除外[16]

③入金を合計　→　「再現売上高」とする
④出金を合計　→　「再現費用」とする
⑤再現売上高－再現費用　→　「再現営業利益」とする
⑥出金（＝再現費用）のうち、摘要欄が「給与代り金」「賞与代り金」「社会保険料」となっているものを合計　→　「再現人件費」とする
⑦出金（＝再現費用）のうち、「給与代り金」を利用している振込先件数の合計（月内の重複除く）　→　「再現従業員数」とする

（図表４－１－１　仕訳のフロー）



①のプロセスを通じて、同一企業が所有する口座を「名寄せ」することによって、同一銀行内にある複数の口座、例えば「人件費の支払用」や「売上の入金用」といった用途別の口座、取引先によって使い分けている複数の口座を網羅的に銀行口座データとして用いることができる。②の仕訳ルールでは、記帳コードや摘要欄の情報により金融取引、税・補助金の受払を除き、入出金先の情報により同一企業内の取引を除くことで、集計対象の取引を売上に近い入金、費用に近い出金に絞ることを試みている。なお、摘要欄が空欄の

[16] 自らが保有する他行の口座からの振込・他行の口座への振込を売上や費用として分類してしまうことを防ぐことが目的。

取引は分析結果に大きな影響は与えていないことを確認している[17]。③④⑤に関して、再現売上高や再現費用及び再現営業利益を算出しているが、これらは銀行口座データによって算出しているため、現金の動きに着目した、いわゆる営業キャッシュフローに概念としては近いと考えられる。今回の分析では財務諸表上の営業利益を比較対象としたが、こうしたキャッシュフローを見るに当たっても、銀行口座データには利用価値があると考えられる。

（２）　仕訳ルールに基づく売上高等の再現

本節では、前節で構築した仕訳ルールに基づき、銀行口座データによる売上高・費用・営業利益・人件費・従業員数の再現データを、財務データと比較する。銀行口座データの入手可能期間が 2018 年１月以降のため、2018 年度以降の財務諸表データがある企業[18]について比較を行った（63,149 企業、延べ 225,349 会計年度）。利用可能な年数[19]及び事業年度毎にデータが入手可能な企業数は図表４－２－１の通りとなった。

（図表４－２－１　利用可能な年数及び事業年度毎の企業数）

利用可能な年数	企業数
1年間	2,098
2年間	2,229
3年間	20,433
4年間	34,451
5年間	3,938

事業年度	企業数
2018	4,473
2019	55,030
2020	60,166
2021	60,705
2022	44,975

この銀行口座データ全体には、口座を開設したが普段の取引は別の金融機関で行っている企業（売上及び費用がほぼ０となる）や、人件費の支払は別の金融機関で行っている企業（人件費及び従業員数がほぼ０となる）なども含まれており、後述するように、銀行口座全体で見ると売上高・費用・人件費・従業員数ともに０近辺に分布の山ができる結果となった。以下の各項では、売上高・費用・利益・人件費・従業員数に関する再現データについて、再現データ÷財務データの値、すなわち銀行口座データから再現した財務情報が、銀行が保有している当該企業の財務諸表のデータをどの程度再現できているかを「再現比

[17] （２）の検証作業を摘要欄が空欄のサンプルを除いて実施したところ同様の結果が得られたことから、空欄のサンプルによる大きな影響はないと考えられる。

[18] 各企業の再現売上高等を、財務諸表データの事業年度に合わせて（例えば、2022 年１月～2022 年 12 月を対象とした財務諸表データが入手可能な場合、再現売上高も 2022 年１月から 12 月に合わせる）比較した。銀行口座データが 2018 年１月以降のため、2018 年決算のデータがある企業は 2018 年時点で 12 月決算の企業に限られている。例えば、８月が決算月で 2017 年９月～2018 年８月が事業年度の企業の場合、2017 年９月～12 月の銀行口座データが入手不可のため、2018 年度のデータは取得できない。

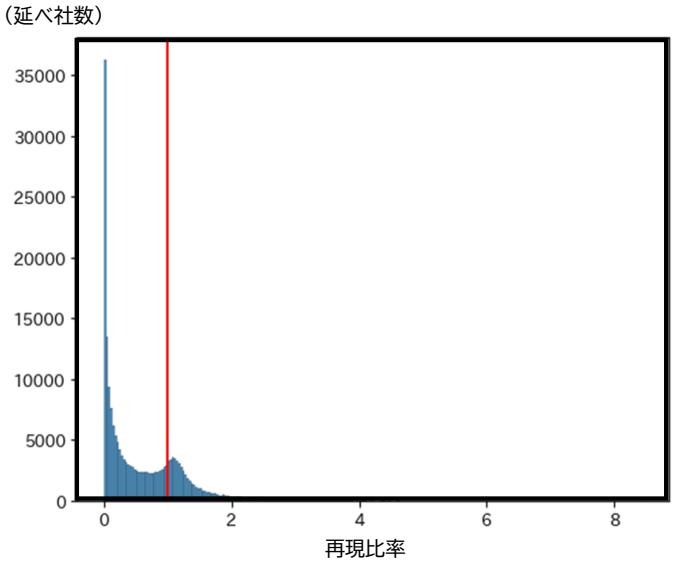
[19] 例えば、2022 年度決算のデータのみ利用できる企業は「１年間」、2020 年度・2021 年度・2022 年度決算のデータを利用できる企業は「３年間」利用可能とする。

率」[20]として、各項目別にみていくことにする。なお、各項目によって、財務指標が０または負のサンプルを除外し、またヒストグラムからは１パーセンタイル以下、99 パーセンタイル以上の値を削除している。

①　売上高

売上高に関する再現比率は図表４－２－２の通りとなった。

（図表４－２－２　売上高の再現比率）



再現比率(x)の範囲	延べ社数
0.0 ≦ x < 0.1	71,338
0.1 ≦ x < 0.2	19,219
0.2 ≦ x < 0.3	12,766
0.3 ≦ x < 0.4	9,525
0.4 ≦ x < 0.5	8,185
0.5 ≦ x < 0.6	7,164
0.6 ≦ x < 0.7	7,045
0.7 ≦ x < 0.8	6,837
0.8 ≦ x < 0.9	7,245
0.9 ≦ x < 1.0	8,292
1.0 ≦ x < 1.1	10,302
1.1 ≦ x < 1.2	9,901
1.2 ≦ x	35,444

（備考）赤線は 1.0 を示す。ヒストグラムでは１パーセンタイル以下、及び 99 パーセンタイル以上のデータを削除している。再現データまたは財務データが０及び負のサンプルは除外した。再現データ及び財務データがいずれも正の値を取る有効なサンプルサイズは延べ 213,268 社、企業数は 59,158 社。

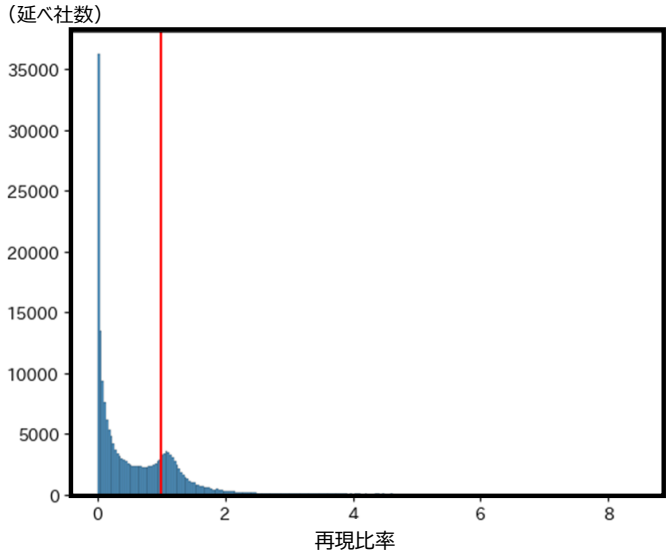
図表から分かるように、０近辺に企業が集中しており、多くの企業で再現データが財務データを全く再現できていないことが分かる。他方、1.0 近辺においても小さいながらも山が生じており、一定数の企業においては再現データによって財務諸表を概ね正確に再現できていることが分かる。なお、1.0 を大きく超える企業に関しては、例えば本来売上に含まれない資金の貸借や一時金などを「売上」として認定してしまっている、といった可能性が考えられる。

[20] 財務諸表より、ある会計年度の売上高が１億円であった企業において、再現データの売上高が１億円となった場合再現比率 1.0（完全に再現できている）、２億円となった場合 2.0（何らかの要因で売上を過剰に計上している）、1000 万円となった場合 0.1（売上の１割しか再現できていない）となる。

② 費用

費用に関する再現比率は図表４－２－３の通りとなった。

（図表４－２－３　費用の再現比率）



再現比率(x)の範囲	延べ社数
0.0 ≦ x < 0.1	71,338
0.1 ≦ x < 0.2	19,219
0.2 ≦ x < 0.3	12,766
0.3 ≦ x < 0.4	9,525
0.4 ≦ x < 0.5	8,185
0.5 ≦ x < 0.6	7,164
0.6 ≦ x < 0.7	7,045
0.7 ≦ x < 0.8	6,837
0.8 ≦ x < 0.9	7,245
0.9 ≦ x < 1.0	8,292
1.0 ≦ x < 1.1	10,302
1.1 ≦ x < 1.2	9,901
1.2 ≦ x	35,444

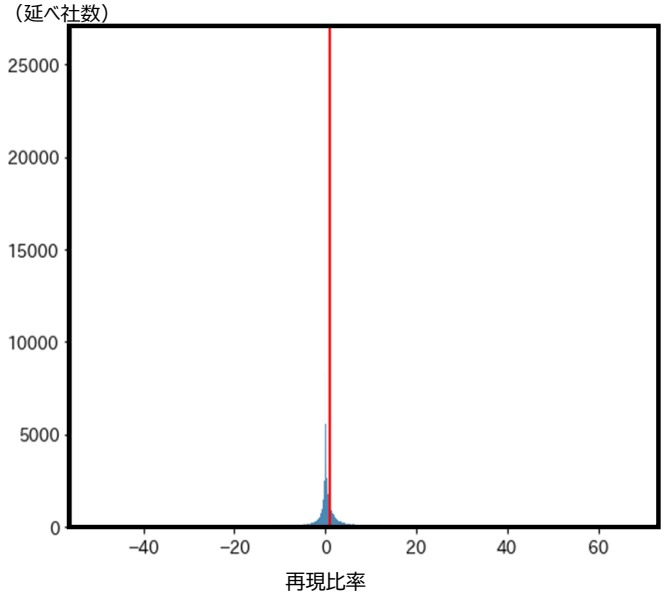
（備考）赤線は 1.0 を示す。ヒストグラムでは１パーセンタイル以下、及び 99 パーセンタイル以上のデータを削除している。再現データまたは財務データが０及び負のサンプルは除外した。再現データ及び財務データがいずれも正の値を取る有効なサンプルサイズは延べ 196,883 社、企業数は 53,027 社。

こちらも売上高と同様、０近辺に企業が集中しているが、１近辺に小さな山が出来ているという結果となっており、大多数の企業に関してはほとんど財務諸表を再現できていないが、一部の企業に関しては比較的正確に再現できているという結果となっている。なお、費用に関して留意すべき点として、同一企業間のやり取りの全てを除外できていない、会計上の投資が費用としてカウントされている、財務データでは会計基準に基づき減価償却費等を費用に含むが、銀行口座データではあくまで金銭の移動を伴う取引のみ記録されるため含まないなど、再現データと会計上の費用で対象が異なる場合がある。

③　利益

利益の再現比率については図表４－２－４の通りとなった。

（図表４－２－４　利益の再現比率）



再現比率(x)の範囲	延べ社数
x ＜ 0.0	80,743
0.0 ≦ x ＜ 0.1	22,238
0.1 ≦ x ＜ 0.2	8,506
0.2 ≦ x ＜ 0.3	6,418
0.3 ≦ x ＜ 0.4	5,918
0.4 ≦ x ＜ 0.5	4,472
0.5 ≦ x ＜ 0.6	3,907
0.6 ≦ x ＜ 0.7	3,419
0.7 ≦ x ＜ 0.8	3,077
0.8 ≦ x ＜ 0.9	2,791
0.9 ≦ x ＜ 1.0	2,521
1.0 ≦ x ＜ 1.1	2,322
1.1 ≦ x ＜ 1.2	2,183
1.2 ≦ x	46,483

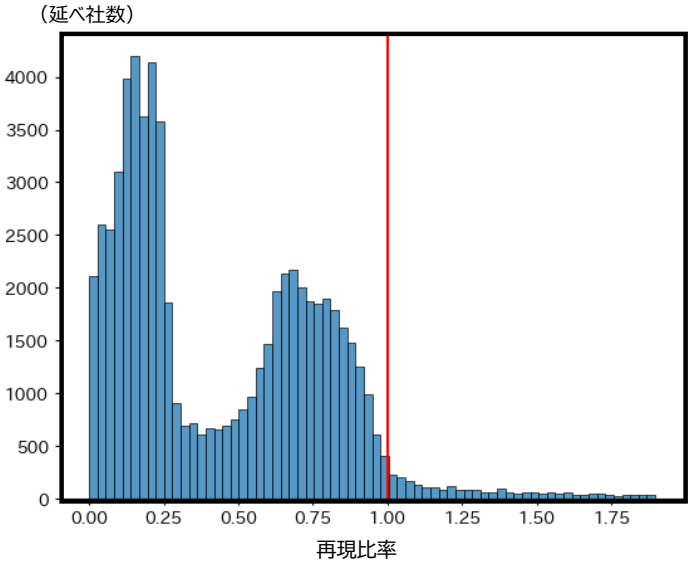
（備考）赤線は 1.0 を示す。ヒストグラムでは１パーセンタイル以下、及び 99 パーセンタイル以上のデータを削除している。有効なサンプルサイズは延べ 196,878 社、企業数は 53,027 社。

０近辺に企業が集中するのみで、１近辺に山は見られず、利益の再現データの作成については特に課題が残る結果となった。「再現売上高－再現費用」で再現利益を作成しているため、再現データによる売上高または費用が正確でない場合、利益が極端に上振れないし下振れする事象がみられ、利益の再現比率が－5.0 を下回る、あるいは＋5.0 を上回る企業も一定程度存在する。なお、費用の項でもふれたように、銀行口座データの性質上、金銭の出納を伴う取引を把握しているため、仮に銀行口座データを用いて各取引を漏らさず捕捉して再現データを作成できた場合、営業利益ではなく営業キャッシュフローがより当てはまりのよい比較対象となると考えられる。

④　人件費

人件費の再現比率については図表４－２－５の通りとなった。

（図表４－２－５　人件費の再現比率）



再現比率(x)の範囲	延べ社数
0.0 ≦ x < 0.1	132,799
0.1 ≦ x < 0.2	13,899
0.2 ≦ x < 0.3	9,677
0.3 ≦ x < 0.4	2,423
0.4 ≦ x < 0.5	2,441
0.5 ≦ x < 0.6	3,802
0.6 ≦ x < 0.7	7,187
0.7 ≦ x < 0.8	6,774
0.8 ≦ x < 0.9	5,853
0.9 ≦ x < 1.0	2,805
1.0 ≦ x < 1.1	704
1.1 ≦ x < 1.2	353
1.2 ≦ x	3,202

（備考）赤線は 1.0 を示す。ヒストグラムでは１パーセンタイル以下、及び 99 パーセンタイル以上のデータを削除している。再現データまたは財務データが０及び負のサンプルは除外した。再現データ及び財務データがいずれも正の値を取る有効なサンプルサイズは延べ 191,928 社、企業数は 51,984 社。

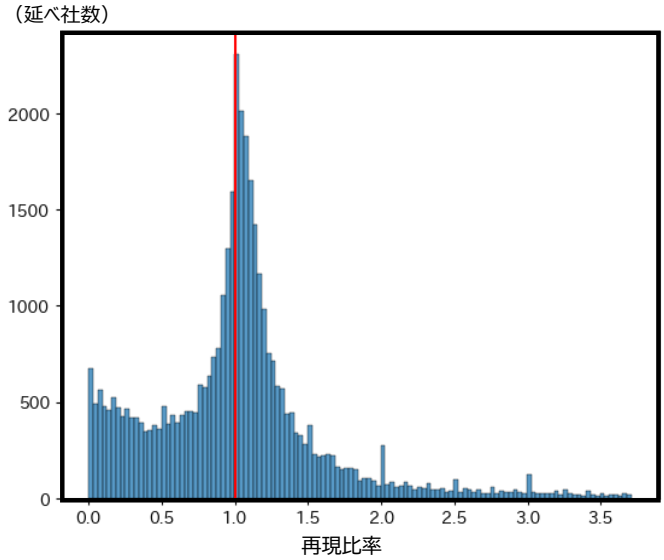
売上高等と同様、０付近が大半だが、1.0 より少し小さい 0.7～0.8 のあたりに山がある分布となっている。人件費に関しては、売上や費用と異なり複数の銀行口座を取引先に応じて使い分けている事例が少ないため、他行が人件費の支払口座の場合、人件費をほとんど再現できず、みずほ銀行が人件費の支払口座の場合、人件費を高い割合で再現できるといった、再現比率の二極化が起こっているのではないかと推測される。なお、1.0 より低い部分に山が生じている理由として、財務データ上で人件費として含まれる派遣労働者に関する人件費や委託事業の人件費、社会保険料以外の福利厚生費 [21]が、給与・賞与及び社会保険料のみをカバーしている再現データでは拾い切れていないことが理由として考えられる。

21 具体的には、企業が直接医療機関に支払う健康診断費用や企業による借上げの社宅の貸与に係る費用など

⑤　従業員数

従業員数の再現比率については図表４－２－６の通りとなった。

（図表４－２－６　従業員数の再現比率）



再現比率(x)範囲	延べ社数
0.0 ≦ x < 0.1	157,373
0.1 ≦ x < 0.2	1,539
0.2 ≦ x < 0.3	1,458
0.3 ≦ x < 0.4	1,245
0.4 ≦ x < 0.5	1,156
0.5 ≦ x < 0.6	1,395
0.6 ≦ x < 0.7	1,368
0.7 ≦ x < 0.8	1,648
0.8 ≦ x < 0.9	2,229
0.9 ≦ x < 1.0	4,151
1.0 ≦ x < 1.1	6,617
1.1 ≦ x < 1.2	4,291
1.2 ≦ x	12,130

（備考）赤線は 1.0 を示す。ヒストグラムでは１パーセンタイル以下、及び 99 パーセンタイル以上のデータを削除している。再現データまたは財務データが０及び負のサンプルは除外した。再現データ及び財務データがいずれも正の値を取る有効なサンプルサイズは延べ 196,510 社、企業数は 52,917 社。

従業員数に関しても、売上高等と同様に０付近が大半だが[22]、1.0付近の割合が比較的目立つ分布となっている。なお、2.0 や 3.0 近辺に局所的な企業の集中が見られるが、これは１か月における複数回に分けた給与の支給や、各種手当や賞与の支給を通常の給与と重複して計上していることが一因と考えられる。分析の過程においても、一か月における従業員数の推計を給与等の振込先の延べ件数からユニークカウント数に変更したところ、2.0 や 3.0 の局所的な山は低くなったが、なお影響が取り切れず局所的な山が残存している結果となっており、同一人物に対して月内に複数回給与等を振り込んでいる事象をどう的確に判別するかは、従業員数の再現データの作成に当たっての課題となる。

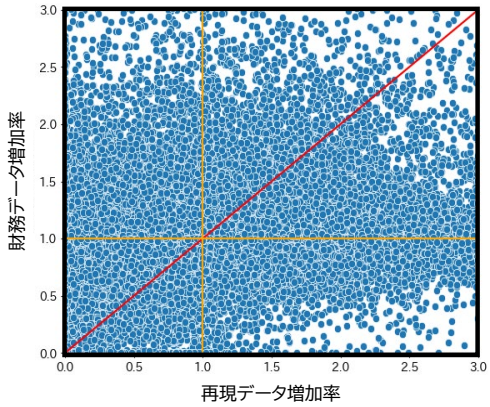
[22] １パーセンタイル以下のデータを図表に掲載した場合、社数の軸が拡大しその他の分布の動きが見えにくくなることから削除しており、結果として図表上では０付近の企業が目立たない形となっている。

⑥　増加率の比較

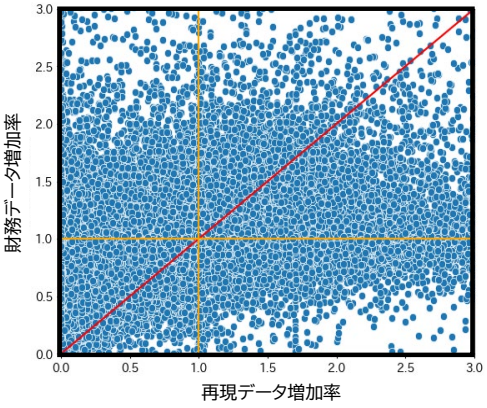
ここまで、再現データが財務データの「水準」をどの程度再現できているかを分析し、売上高等の再現比率が著しく低い企業が多いこと、また比較的正確に再現できる企業も一定数存在することが明らかになったが、再現の「水準」が低くとも、増加率を的確に把握できていれば、データとして活用できる可能性もある。一例として、ある事業年度の財務諸表上の売上１億円を再現データでは 1000 万円しか捕捉できていなかったとしても、翌事業年度の売上 1.2 億円を再現データで 1200 万円分捕捉できている、すなわち売上高の20%増加が再現データ上でも見られれば、再現データによって売上高等の増加率を捕捉できていると評価することができる。そこで、財務データにおける売上高等の増加率と、再現データの売上高等の増加率をプロットし、何らかの傾向がみられるか確認した（図表４－２－７）。

（図表４－２－７　各指標の再現データと財務データの増加率の比較）

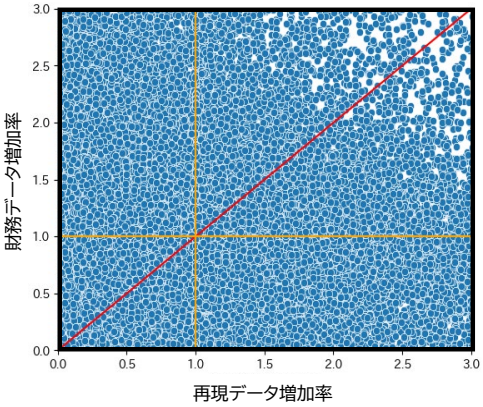
（１）売上高の増加率の比較



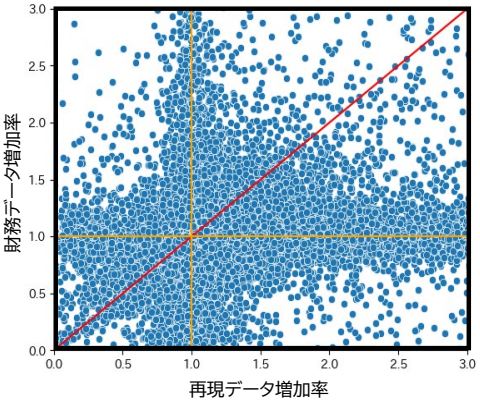
（２）費用の増加率の比較



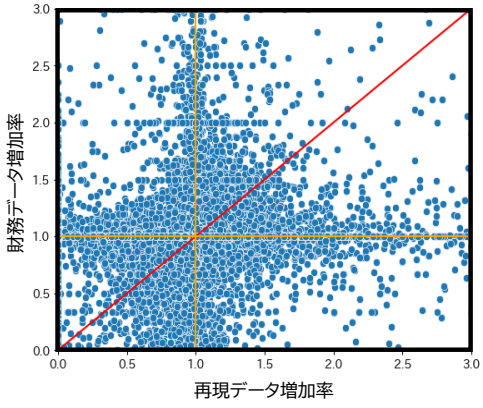
（３）利益の増加率の比較



（４）人件費の増加率の比較



（５）従業員数の増加率の比較



（備考）再現データ（横軸）、財務データ（縦軸）に関して増加率（t 年度の値/t-1 年度の値）を計算。増加率が 1.0 であれば前年度と変化がないことを示す。赤線が 45 度線（45 度線上の点は、再現データが r 増加した際、財務データも r 増加していることを意味する）、橙線が 1.0 を示す。

図表から見て分かるように、財務諸表データの各指標の増加率を再現データが安定して再現できているといった関係にはなく、増加率の観点からも再現データが財務諸表データを再現できていないといった形となった。

（３）　　複数条件を用いた企業の絞り込みによる財務諸表の再現

①　絞り込み条件の選定

前節でみたように、全ての銀行口座データを用いて再現データを作成した場合、売上高等の水準の再現率は著しく低い企業が多かったが、再現度合いの高い企業も一定数存在した。そこで、全企業において財務諸表データの再現を目指すのではなく、再現された売上高等の数値が信頼できる企業群を得ることを目的に、複数の条件を用いて企業を 540 社に絞り込み、絞り込んだ企業において再現データ（540 社）と財務データを比較した。絞り込みに当たっては、複数の指標の再現度合が高い企業に絞り込むべく、売上高、人件費、従業員数に関する絞り込みを複数パターン実施するとともに（手順①）、「普段から銀行口座を用いているか」という観点からの条件を追加するため、口座の利用頻度、クレジット支払いや賃貸料支払など口座の利用目的などから、「特定用途条件」を設定し、更なる絞り込みを行った（手順②）。売上高、人件費、従業員数に関する絞り込み条件、及び利用目的に関する絞り込み条件は下図（図表４－３－１・図表４－３－２）の通りである。

（図表４－３－１　絞り込み条件　手順①）

手順①で利用した抽出条件	選択肢
抽出に使った事業年度	[2019] [2019,2020] [2019，2020，2021]
従業員数増加率の|再現-実績|の上限	条件なし、0.1、0.2
従業員数の（再現/実績）の（下限，上限）	条件なし、(0.9，1.1)、（0.8，1.2）
人件費増加率の|再現-実績|の上限	条件なし、0.1、0.2
人件費の（再現/実績）の（下限，上限）	条件なし、(0.5，1.0)、（0.6，0.9)、(0.6，1.0）
売上高増加率の|再現-実績|の上限	条件なし、0.1、0.2
売上高の（再現/実績）の（下限，上限）	条件なし、(0.9，1.1)、（0.8，1.2）

（図表４－３－２　絞り込み条件　手順②）

手順②で利用した抽出条件	選択肢
抽出に使った事業年度	－[23]
帝国データバンクデータのうち みずほ銀行が第一取引銀行かどうか*	Yes、No
事業年度内の月平均入金取引数の下限	0、10
事業年度内の月別最小入金取引数の下限	0、1
事業年度内の月平均出金取引数の下限	0、10
事業年度内の月別最小出金取引数の下限	0、1
同名他行口座の取引がない	Yes、No
年間合計融資回数下限	0、1
月間最小賃貸料支払回数下限	0、1
月間最小クレジット支払回数下限	0、1
月間最小電気水道ガス支払回数下限	0、1
月間最小税金支払回数下限	0、1

なお、従業員数、売上高に関しては、それぞれ再現比率１を中心に（0.9, 1.1）ないし（0.8, 1.2）のレンジを抽出条件としているが、人件費に関しては（0.5, 1.0）、（0.6, 0.9）、（0.6, 1.0）を抽出条件としている。これは前述の通り、銀行口座データの人件費に、派遣労働者や委託事業の人件費、社会保険料以外の福利厚生費が含まれていないことから、「給与額で算出するデータ上の人件費＜財務諸表上の人件費」となることを勘案したものである。

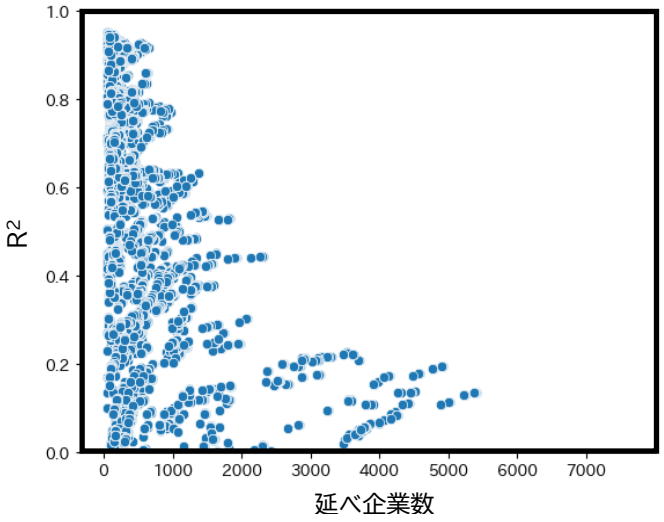
②　各条件による絞り込みの結果

上記の条件に応じて場合分けを行い、売上高、従業員数、人件費の３指標について財務データと再現データの決定係数が高く、サンプルサイズを十分に確保できる絞り込みのパターンを探す。手順①については、抽出に用いる事業年度が３通り、従業員増加率の上限も３通り、従業員数の（上限, 下限）の条件も３通り等となるので、最終的に$3^6 \times 4 = 2916$ 通りの絞り込みパターンが得られる。この時点で「売上高、従業員数、人件費のいずれかの R^2 が 0.7 以上」かつ「サンプルサイズが 500 社以上」となった条件に限定し、手順②の基準を適用して更に絞り込みを行った。最終的な、売上高、従業員数、人件費それぞれの R^2 の平均とサンプルサイズの関係をプロットすると下図（図表４－３－３）の通りと

[23] 手順①で使用した事業年度と同じ。

なった。

（図表４－３－３　売上高、従業員数、人件費の R^2 の平均とサンプルサイズの散布図）

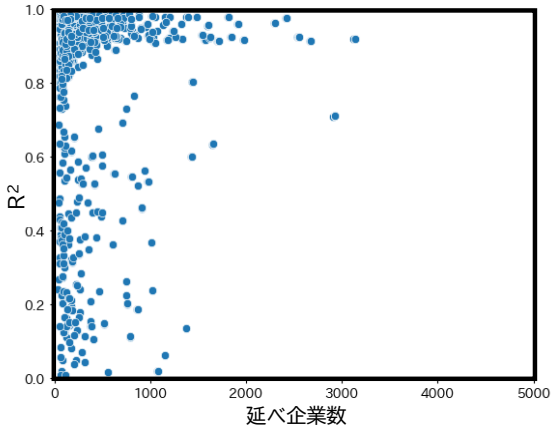


図からみて分かるように、売上高、従業員数、人件費すべてで銀行口座データを用いて財務諸表のデータを再現できている（R^2 の平均が１に近い）企業は、銀行口座データ全体と比べるとごく少数となり、銀行口座データ内の中小企業約 50 万社の中、財務諸表データを結合でき、銀行口座データを用いて売上高、従業員数、人件費をいずれも高い水準で再現できる企業はおよそ 1000 社以下である事がわかる。

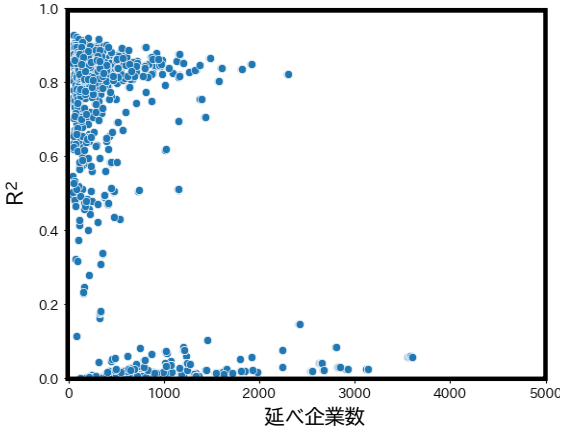
なお、売上高、人件費、従業員数に関する絞り込みに関して、再現データと財務データの売上高、人件費、従業員数のそれぞれの値を比べた際の決定係数 R^2（再現データと財務データの値にどれほど関係性があるか）とサンプルサイズの関係を見ると、図表４－３－４でみられるように、個別の指標毎では必ずしもトレードオフがあるわけではない。そもそも個別の指標を再現できているかを抽出条件にしているという要因もあるが、高い再現水準においてもサンプルサイズが大きい条件の組み合わせは、売上高のみ、人件費のみ、従業員数のみであれば一定数存在している。

（図表４－３－４　売上高・人件費・従業員数それぞれの R^2 とサンプルサイズの散布図（手順①段階））

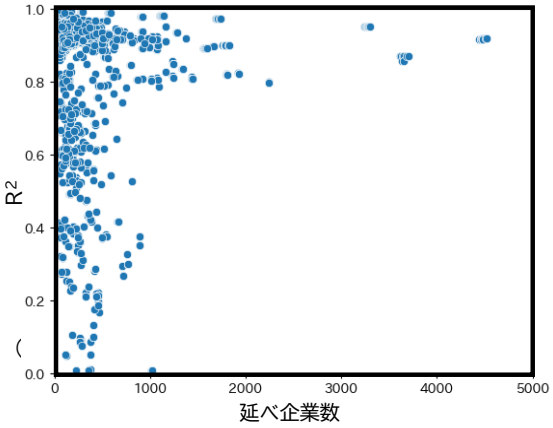
①売上高のR^2とサンプルサイズの関係



②人件費のR^2とサンプルサイズの関係



③従業員数のR^2とサンプルサイズの関係



③　複数条件を用いた絞り込みによる抽出企業の決定

絞り込みの結果を踏まえ、従業員数、人件費、売上のR^2の平均が 0.9 以上かつ、企業数が多い絞り込み条件を抽出すると、下記（図表４－３－５）の通りとなった。

（図表４－３－５　複数条件を用いた抽出結果）

	抽出に用いた年度	従業員数再現比率の（下限，上限）	人件費再現比率の（下限，上限）	売上再現比率の（下限，上限）	特定用途条件	従業員数 R2	人件費 R2	売上高 R2	企業数
1	[2019，2020]	(0.8，1.2)	(0.5，1.0)	(0.8，1.2)	なし	0.977	0.829	0.942	661
2	[2019，2020]	(0.8，1.2)	(0.5，1.0)	(0.8，1.2)	クレジット	0.981	0.831	0.942	628
3	[2019，2020]	(0.8，1.2)	(0.6，1.0)	(0.8，1.2)	なし	0.975	0.865	0.931	566
4	[2019]	(0.8，1.2)	(0.6，1.0)	(0.8，1.2)	賃貸料	0.929	0.882	0.921	544
5	[2019，2020]	(0.8，1.2)	(0.6，1.0)	(0.8，1.2)	クレジット	0.979	0.869	0.930	541
6	**[2019]**	**(0.8，1.2)**	**(0.6，1.0)**	**(0.8，1.2)**	**クレジット 賃貸料**	**0.929**	**0.882**	**0.921**	**540**
7	[2019]	(0.8，1.2)	(0.6，1.0)	(0.9，1.1)	なし	0.928	0.872	0.922	527
8	[2019]	(0.8，1.2)	(0.6，0.9)	(0.8，1.2)	賃貸料	0.922	0.872	0.920	517
9	[2019]	(0.8，1.2)	(0.6，0.9)	(0.8，1.2)	クレジット 賃貸料	0.922	0.872	0.920	513
10	[2019]	(0.8，1.2)	(0.6，0.9)	(0.9，1.1)	なし	0.925	0.863	0.921	504

（備考）「年度内の月別最小入金取引数の下限：１」、「年度内の月別最小出金取引数の下限：１」はすべての条件で適用している。その他表に含まれていない特定用途条件は付けられていない。

この条件の中で、サンプルサイズが十分に確保され、かつ「普段から銀行口座を用いているか」という点を満たすという観点から、有意に選択する形になるが、抽出に用いた事業年度が 2019 年の単年度のみ、従業員数・人件費・売上の再現比率の上限・下限がそれぞれ（0.8, 1.2）、（0.6, 1.0）、（0.8, 1.2）（手順①）かつ「１か月に最低一回以上入金がある」「１か月に最低一回以上出金がある」「１か月に最低一回以上クレジットカードの支払がある」「１か月に最低一回以上賃貸料の支払がある」（手順②[24]）、抽出条件６に合致する 540 社に絞り込んで分析を行うこととした。

[24] 手順②に関して、サンプルサイズが著しく小さくなってしまう等の要因で、「帝国データバンクデータのうちみずほ銀行が第一取引銀行かどうか」「年度内の月平均入金取引数（の下限）」「年度内の月別最小入金取引数（の下限）」等は「特定用途条件」として採用するには至らなかった。帝国データバンクデータについては、銀行口座データ内の財務データを補完する目的で財務諸表データが無い企業のデータを購入したため、そもそも「帝国データバンクデータがある」＝銀行口座データ内に財務データがないサンプルでは、銀行口座データを用いた際の再現度合いが低い企業が多くなった。帝国データバンクデータのような外部データを用い、「第一取引銀行」のデータか否かを峻別することの有効性については、今後も検討を進める必要があると考えられる。

なお、従業員数・人件費・売上の抽出に用いた事業年度が 2019 年単年度である抽出条件を優先させた考え方は以下の通りである。2019 年～2021 年の区間のみで当てはまりの良い企業を抽出したいのではなく、2019 年～2021 年度の区間外においても普遍的に当てはまりの良い企業を抽出することを目的としており、抽出期間を長く取って 2019 年度・2020 年度・2021 年度の区間で高い従業員数・人件費・売上の再現度合いの企業を抽出した場合、2019 年度～2021 年度の範囲で当てはまりが良いことは確実に言えるが、将来のデータにおいても当てはまりが必然的に高いとは言い切れない。一方、2019 年単年度における従業員数・人件費・売上の再現度合いの条件で企業を抽出し、それでもなお期間を通じた R^2が高い場合、2019 年単年度を基準に選択した企業が、2020 年・2021 年両年度でも再現度合いが高いことが分かり、将来的にも当てはまりが良いであろうことが推測できるため、決定係数R^2が同程度であれば、抽出条件に用いる期間が短い方が、将来的にも再現度合いが高い、普遍的な抽出基準となっていると考えられる。

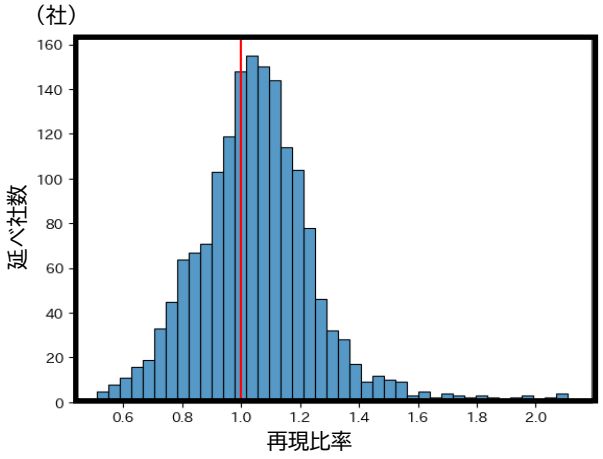
（４）　　540 社絞り込み後の再現データ

この節では、前節の絞り込みルールを用い、540 社に絞り込んだ際の再現データ（540 社）と財務データの比較を行う。

①　再現売上高（540 社）による財務データの再現度合い

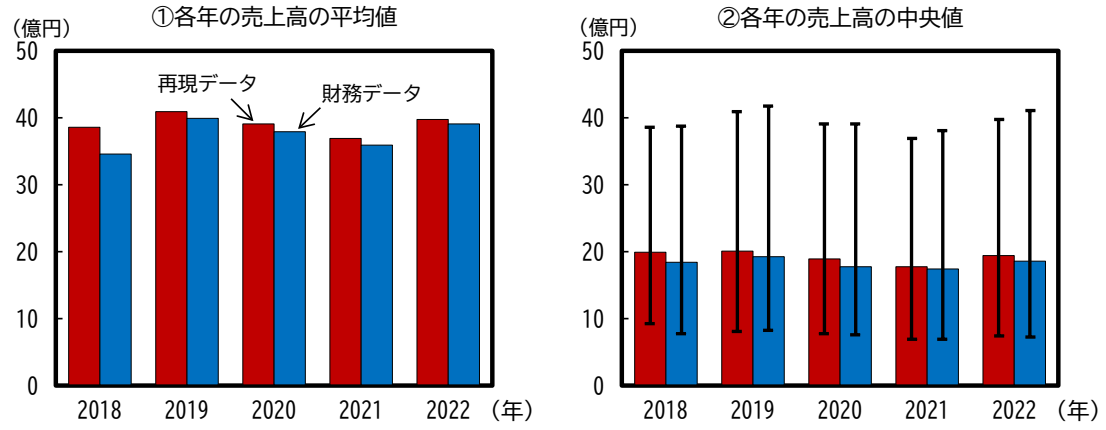
再現売上高（540 社）による財務データ再現度合い、再現増加率の分布は以下の通りとなった。なお、再現度合いは抽出の条件に用いた2019年度以外の期間の再現比率の分布、増加率の分布は 2020 年度以降での再現増加率の分布とする。後述の人件費及び従業員数も同様である。

（図表４－４－１　再現売上高（540 社）の再現比率）



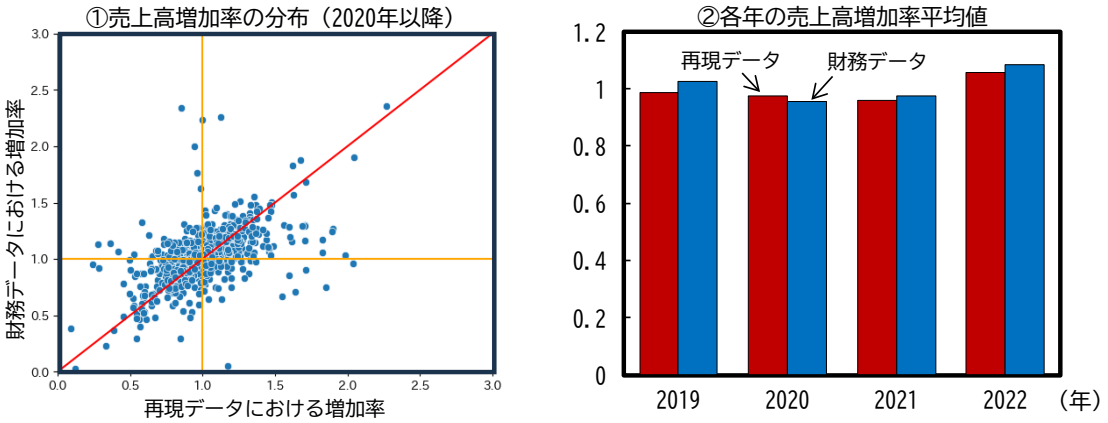
（備考）2019 年度以外での期間での再現比率の分布。

（図表４－４－２　再現売上高（540 社）と財務データの平均値及び中央値）



（備考）②について、エラーバーは上端が第３四分位点、下端が第１四分位点を示す。

（図表４－４－３　再現売上高（540 社）と財務データの増加率）

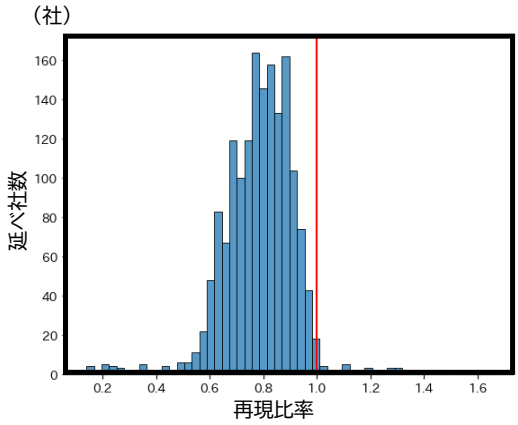


図表から分かるように、再現売上高（540 社）では、全企業を対象とした再現売上高と比べて、財務データの再現度合いが大きく高まっている。2019 年度以外の期間においても再現売上高（540 社）は財務データを１を中心とした正規分布に近い内容で再現できており、増加率でみても、再現できている企業が多い。再現売上高（540 社）の平均値と財務データの売上高の水準には差異が見られるが、売上高の増加率の動きは似通ったものとなっている。中央値・第１四分位点・第３四分位点に関しても、再現売上高（540 社）と財務データ（売上高）の間の動きに類似が見られる。

②　再現人件費（540 社）による財務データの再現度合い

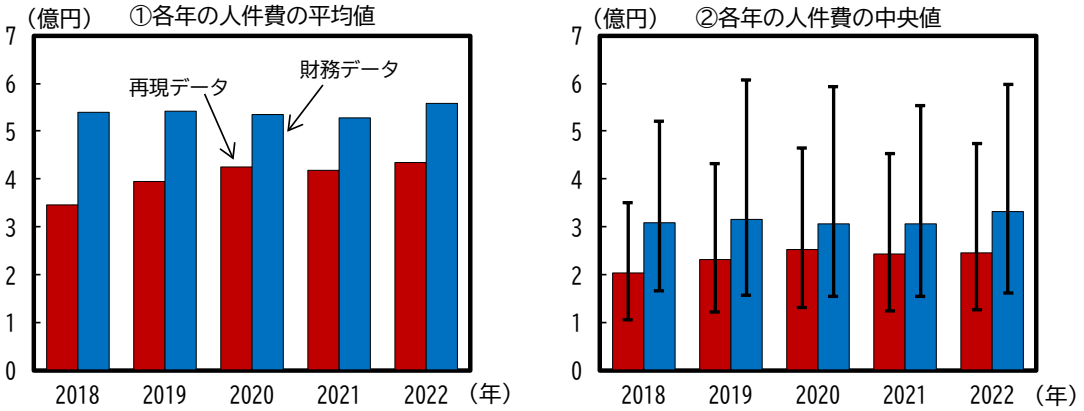
人件費に関しても、再現人件費（540 社）と財務データを同様に比較した。

（図表４－４－４　再現人件費（540 社）の再現比率）



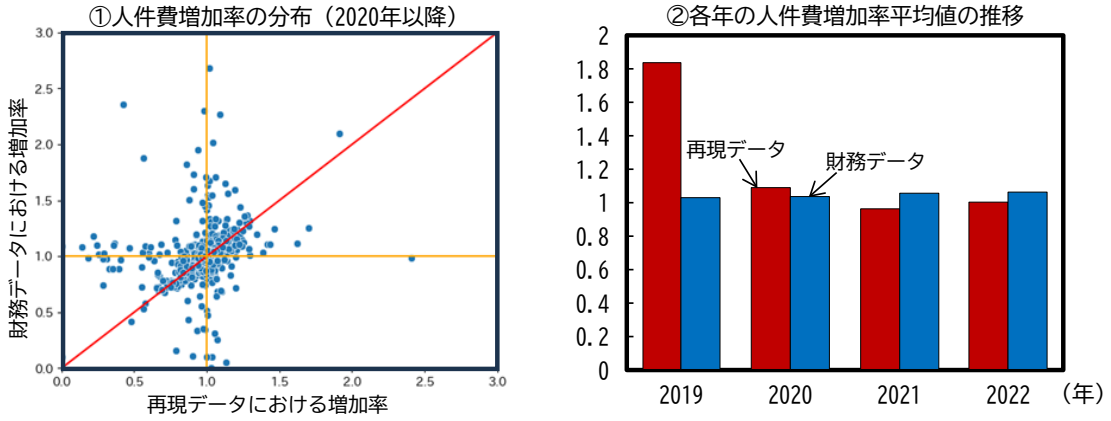
（備考）2019 年度以外での期間での再現比率の分布

（図表４－４－５　再現人件費（540 社）と財務データの平均値・中央値の比較）



（備考）②について、エラーバーは上端が第３四分位点、下端が第１四分位点を示す。

（図表４－４－６　再現人件費（540 社）と財務データの増加率）

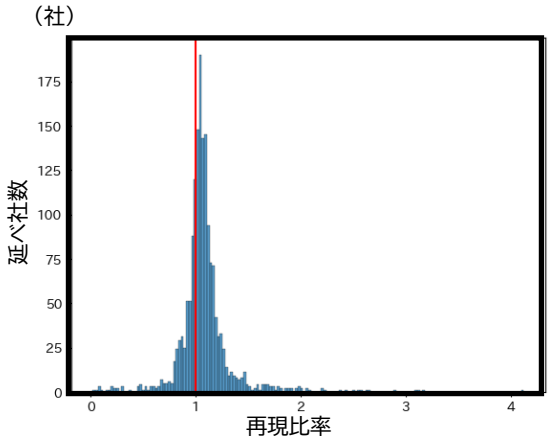


図から見られるように、再現人件費（540 社）は財務データの人件費を 0.8[25]を中心とする正規分布に近い形で再現できており、人件費に関しても絞り込みにより再現度合いが向上したと考えられる。増加率で見た場合も、比較的良好な再現度合いとなっている。

③　再現従業員数（540 社）による財務データ（従業員数）の再現比率

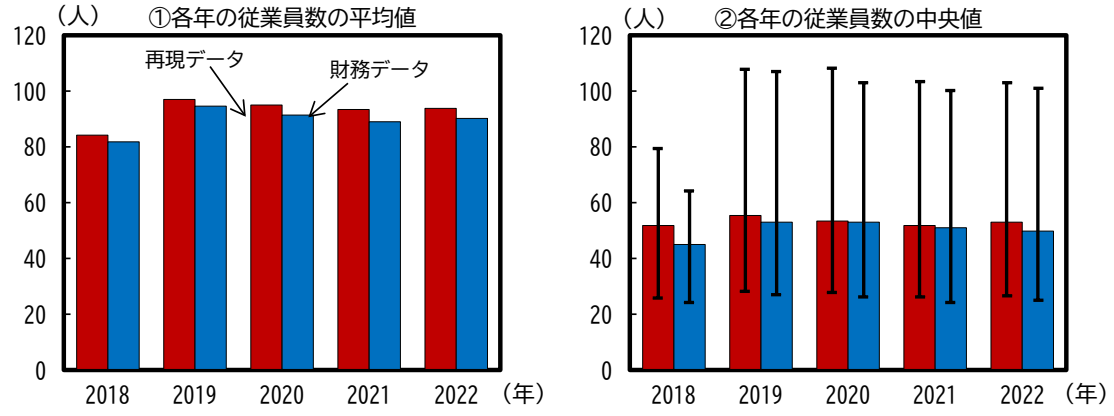
従業員数に関して、再現データ（540 社）と財務データを比較すると以下の通りとなった。

（図表４－４－７　再現従業員数（540 社）の再現比率）



（備考）2019 年度以外での期間での再現比率の分布

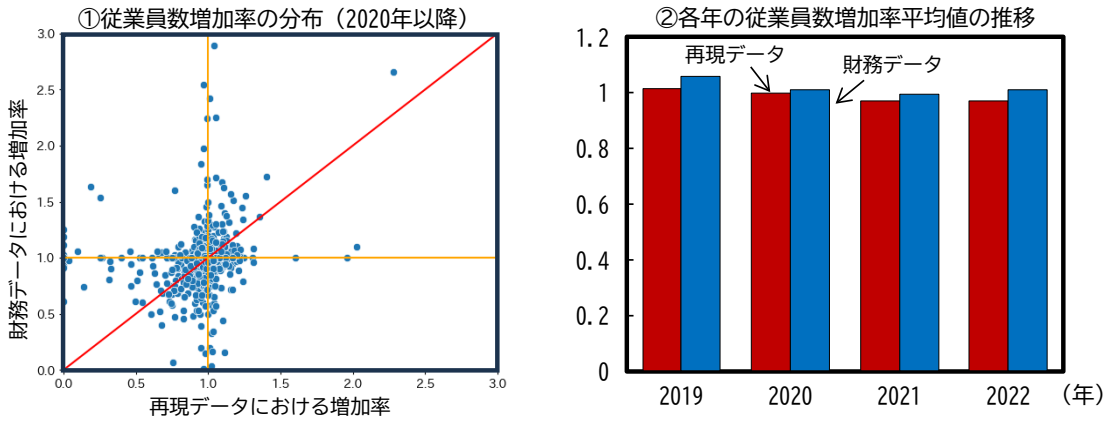
（図表４－４－８　再現従業員数（540 社）と財務データの平均値・中央値の比較）



（備考）②について、エラーバーは上端が第３四分位点、下端が第１四分位点を示す。

[25] 前述の通り、銀行口座データの人件費は派遣労働者や委託事業の人件費、社会保険料以外の福利厚生費を含んでいないため、銀行口座データは財務データによる人件費を１より小さい割合で再現すると考えられる。

（図表４－４－９　再現従業員数（540 社）と財務データの増加率）



図表から分かるように、再現従業員数においても財務データの再現度合いは１を中心とした分布の形に近くなった。再現従業員数に関しては、一度給与の振込口座を高い再現でカウントできた場合、給与振込口座の変更といった事象が起こらなければ、それ以降の給与振込数＝再現従業員数を継続して把握できると考えられ、より裾野が狭い分布になっていると考えられる。

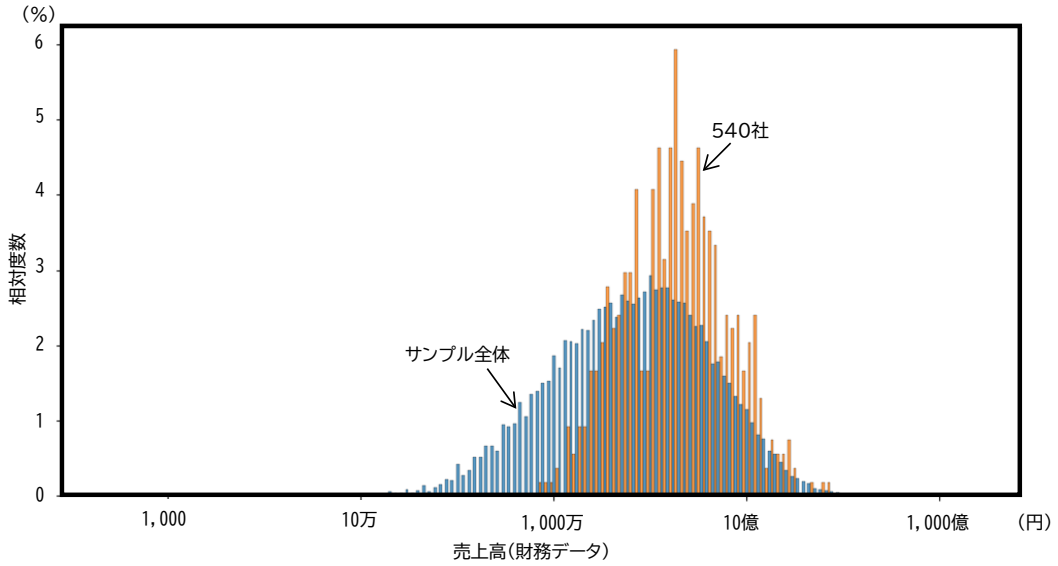
④　絞り込んだ企業 540 社の特徴

ここまで、540 社に絞り込んだ際の再現データの再現度合い向上についてみてきたが、この小節では絞り込んだ 540 社の特徴について考察する。絞り込みの結果抽出された 540 社とサンプル企業全体の財務データを比較すると、540 社では売上高・資本金で見て相対的に規模の大きな企業が多く（図表４－４－10）、都道府県としては東京都所在の企業が（図表４－４－11）、業種としては製造業・卸売業が多い（図表４－４－12）。製造業や卸売業において、540 社内に含まれる再現度合いが高い企業が多くなった理由としては、製造業・卸売業ともに、取引における現金の使用比率が低く、即時性が高いインターネットバンキングによる振込の利用比率が高いなど [26]、業種の特徴として銀行口座の履歴に残りやすい取引が多いことが可能性として挙げられる。

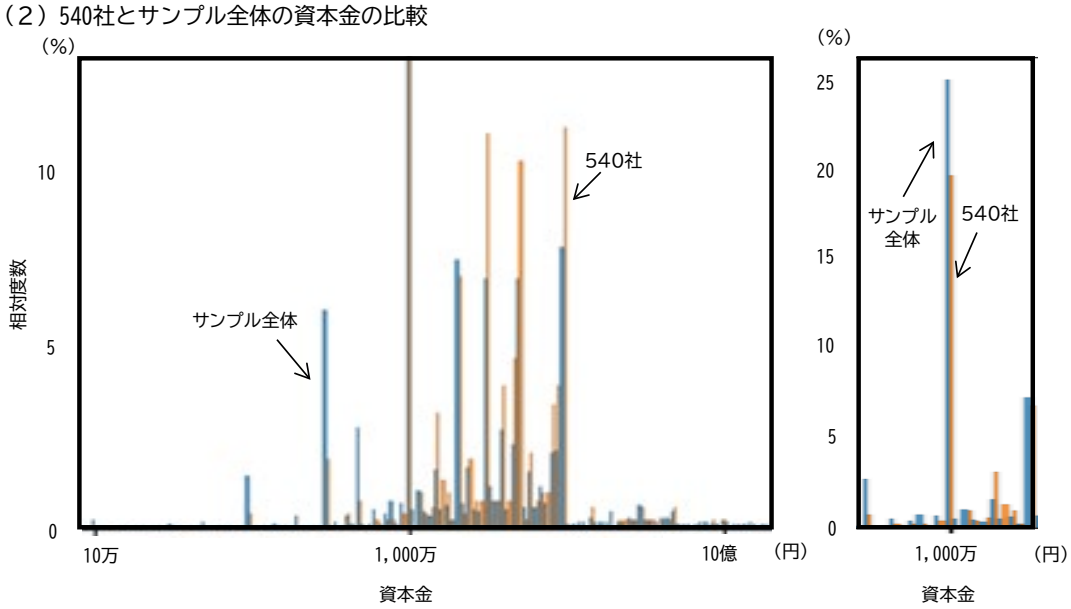
[26] 三菱 UFJ リサーチ＆コンサルティング（2023）によると、振込（インターネットバンキング、ファームバンキング）の使用割合は全業種で 48.4%、製造業では 57.6%、卸売業では 65.0%となっており、製造業や卸売業では即時性の高いインターネットバンキングの活用が進んでいることが示唆される（ただし、振込（窓口、ＡＴＭ）を見ると、製造業・卸売業（それぞれ 46.8%、41.6%）は業種全体（55.6%）を下回る）。販売先との取引（受取）を見ても、業種全体で振込の利用割合が 82.1%であるのに対し、製造業では 86.8%、卸売業では 89.9%と、銀行口座の使用頻度が高いと推測される。

（図表４－４－10　年間売上高及び資本金に関する 540 社とサンプル全体の比較）

（１）540社とサンプル全体の年間売上高の比較

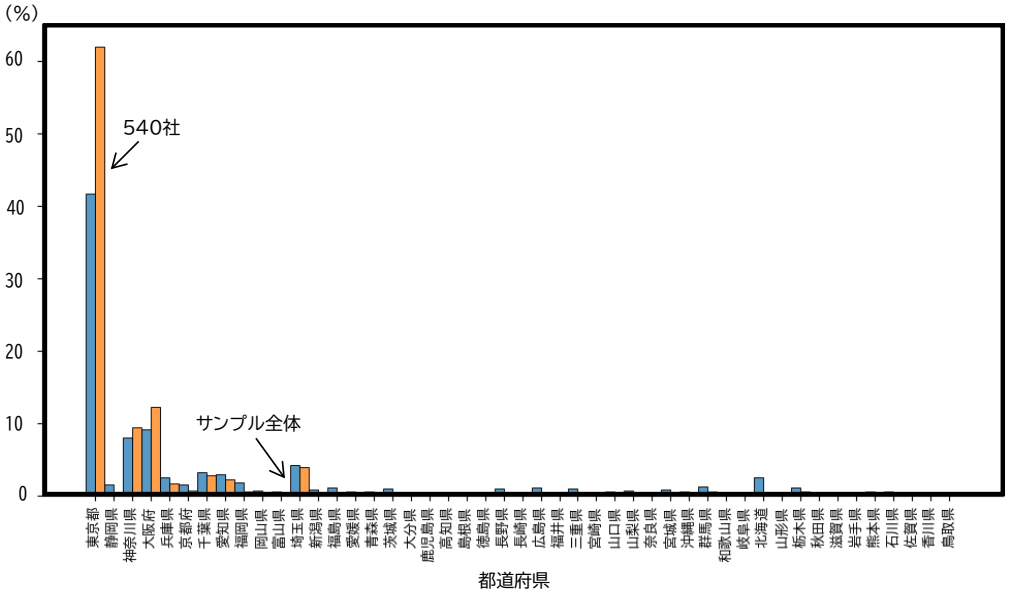


（２）540社とサンプル全体の資本金の比較

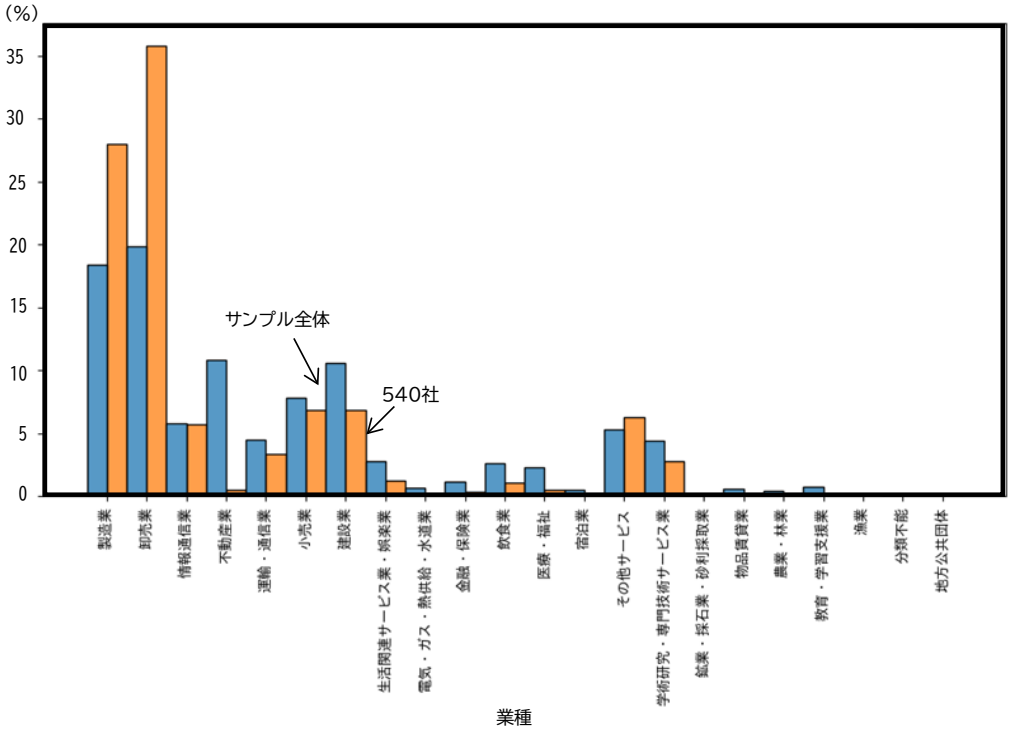


（備考）売上高、資本金額が対数スケールとなっていることに留意。図表作成の都合上、「サンプル全体」に再現度合いが高い 540 社は入っていないが、540 社がサンプル全体に占める割合はごくわずかなため、全体の分布に影響はない（以下同様）。資本金は 1000 万円に多くの企業が集中しているため、1000 万円周辺を軸の単位を変更して再掲している。

（図表４－４－11　所在地に関する 540 社とサンプル全体の比較）



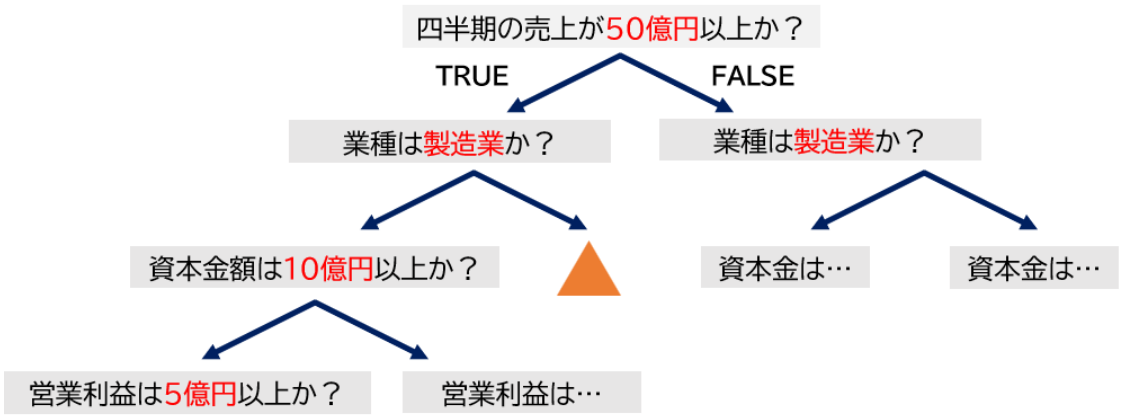
（図表４－４－12　業種に関する 540 社とサンプル全体の比較）



（５）　“Isolation Forest”　を利用した絞り込み手法の検討

ここまで、複数条件を用いた絞り込み手法の検討、及び絞りこまれた 540 社の再現データの特徴についてみてきたが、本節では複数条件を用いた 540 社への絞り込みに加えて、機械学習を利用した絞り込みルール「Isolation Forest」について検討する。Isolation Forest は、例えば企業サンプルを四半期の売上に関して「売上が 50 億円以上か、以下か」というように２つのカテゴリに分類する決定木（図表４－５－１）を利用した機械学習手法の一つである。具体的には、データをサンプリングした上で、大量の決定木を作成し、特定のデータが決定木を用いて分類した結果、「孤立」するまでに要した判定の回数の平均を使用して異常値スコアを算出する。

（図表４－５－１　Isolation Forest における「決定木」のイメージ）



（備考）赤字で示した金額や業種のカテゴリは、Isolation Forest の過程によってランダムに変わり、例えば売上高に関して本図のように 50 億円を閾値として分類する場合もあれば、32 億 4576 万円、96 億 61 万 117 円といった値を閾値として分類することもある。仮に「売上が 50 億円以上」かつ「製造業」の条件を満たす、図中△に分類される企業が１社しかなかった場合、その企業は距離（判定の回数）２で「孤立」したとみなすことができる。

大きな特徴として、「事前」に外れ値のルールを定めて外れ値を計算する[27]一般的な異常値処理のやり方と異なり、ランダムな条件による決定木分析を繰り返すことによって、事後的に外れ値のルールを機械学習によって算定するという手法の違いがあり、分析対象とするデータの特徴が事前に予想できない場合[28]や、多次元データの処理[29]に関して強みが

[27] 具体的には、中央値から標準偏差の２倍以上乖離した値（２σ以上外れた値）を外れ値とする、全サンプルの上下１パーセンタイルの値を外れ値として除外する、等。

[28] データの特徴が正規分布となること等が予想できる場合は、代表値（中央値・平均値など）と標準偏差を用いた伝統的な外れ値処理の手法が機能しやすいと考えられる。

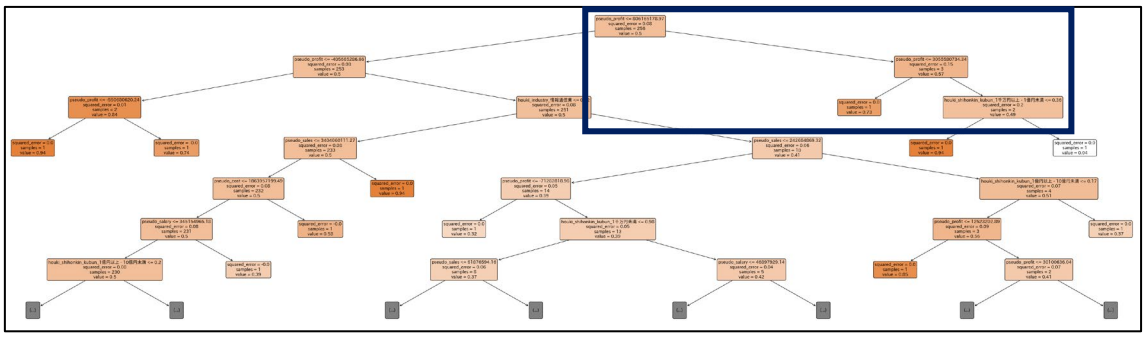
[29] 銀行口座データのように、「資本金」「売上高」「費用」「業種」「営業利益」等、数多くの変数を用いて異常値を認定しようとする場合、従来的な外れ値処理の手法では「資本金」と「売上高」のどちらを優先して異常値認定に用いるべきか、といった判断を事前に行う必要があるが、Isolation Forest の場合は、どちらの項目を外れ値認定に用いるべきか、処理プロセスの過程で事後的に判別でき、見落としや恣意性を排除しやすい。

ある手法である[30]（Liu et al.（2008、2012））。

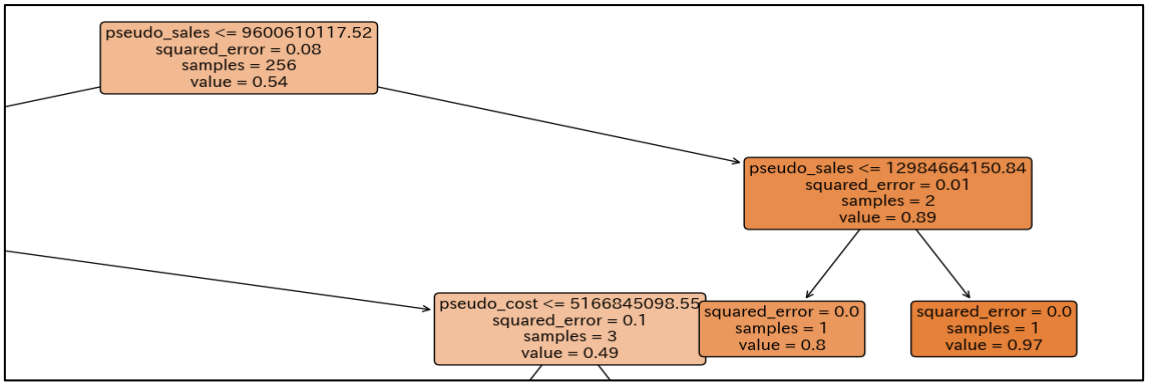
この Isolation Forest を用いて、銀行口座データの異常値処理を行った。今回、全サンプルからランダムに 256 個のデータを抽出し[31]、毎回抽出データを変えながら、決定木分析を繰り返し、異常値を判別しやすい条件を機械学習で選定した（図表４－５－２）。

（図表４－５－２　生成された銀行口座データの Isolation Forest における決定木の一例[32]）

①決定木の一例の全体図



②上記決定木の一部（青枠内）を拡大



再現利益・再現売上高・再現費用・資本金（財務データ利用）・業種といった項目において決定木分析を行った後、解釈ができる異常値処理の条件として、各四半期ごとのデータにおいて、

条件１：再現利益　-405,665,286 未満　または　806,165,178 以上

条件２：推定売上　9,600,610,117 以上

---

[30] また、処理するデータ量が増えた際の計算量の増加が線形である（指数関数的に増えない）という実装上の強みもある。Liu et al. (2008, 2012) をはじめとして、決定木分析のプロセスを２変数データを例に、xy 平面上の座標に対して直線 x = a,b,c… 及び y = α, β, γ …を引いていき、直線で区切られた領域において「孤立」した値を異常値とする、と説明されることが多いが、本稿では銀行口座データ分析に当たって実際に作成した決定木のイメージと合わせるため、実際に「木」の構造を図示した。

[31] 全サンプルを対象に一度に決定木分析を行うのではなく、サブサンプルの抽出と決定木分析を繰り返す理由として、サンプル全体に対する過学習によって未知のデータへの当てはまりが悪くなることを防ぐことが挙げられる。

[32] Isolation Forest においてはこうした決定木を大量に生成して機械学習手法を用いて分析するため、ここに掲載したのはあくまで大量の決定木の中の一つの例である。

の条件１または２のどちらかを満たすデータを異常値として除去する方針とした。上記の方針に基づくと、約６万４千社分存在する四半期データにおいて、平均して[33]1,232 社が異常値として除外された。複数条件を用いた 540 社への絞り込み時と比べ、多くのデータを残しつつ、後述するように法人企業統計と比較した際にも推計精度を上げることが可能となっており、機械学習の手法を用いたビッグデータ分析の有用性が示されたと考えられる。

５．業績等に関するマクロ的動向の確認

（１）　分析方法

銀行口座データから、全業種と主な業種別の売上高、営業利益、人件費を再現し、法人企業統計等の公的統計と比較し、経済指標の時系列変動がどの程度再現できるかを検証する。法人企業統計との比較に当たっては、銀行口座データの企業規模、業種別のサンプルの偏りを補正するため、同統計の母集団に合わせる形で拡大推計を行う。

全体サンプルを用いて作成したデータとの比較の他、資本金規模を限定したサンプル、４．で検討した複数条件を用いた絞り込みで抽出した 540 社のサンプル、Isolation Forest を用いて異常値を除いたサンプルとの比較も行い、サンプルの抽出方法を工夫することで公的統計との誤差がどの程度縮小するかを検討する。全業種での比較の他、製造業、卸売業・小売業、不動産業・物品賃貸業、宿泊業・飲食サービス業を対象に業種別の比較も行い、業種により法人企業統計等との整合性がどの程度異なるかも確認する。また、試みとして、再現データの業種や指標を絞り、月次統計である鉱工業生産指数、商業動態統計、総雇用者所得との比較も行う。

①　拡大推計

みずほ銀行の口座データは中小企業の定義に基づき抽出したが、結果として資本金規模が大きい企業が多いサンプルとなっていた。このため、再現データの拡大推計は、法人企業統計の全規模の母集団に合わせる形で行い、比較の対象とする法人企業統計の各指標も全規模の数値を用いる[34]。

４．で得られた再現データについて、各資本金・業種別の月次集計値を四半期で合計する。その際、第一四半期（グラフ上では Q1 と表記）は 1～3 月期とする（以下、Q2：４～６月、Q3：7～9 月、Q4：10～12 月）。

銀行口座データで利用可能な業種区分は日本銀行の金融統計調査の業種であるが、これ

[33] 毎四半期ごとに外れ値の判定を行っているため、例えば 2020 年 1-３月期に「異常値」と認定された企業 A のデータが、2020 年４-６月期には「異常値ではない」と認定されることもあり、「異常値」の数は毎四半期ごとに変わりうる。

[34] 確認のため、法人企業統計の資本金 10 億円未満、1 億円未満の数値との比較も行ったが、基本的な結論は変わらなかった。

を法人企業統計調査の業種と対応させ、16 の業種を設定した（具体的な業種は図表５－１－１を参照)。拡大推計は、資本金区分別・業種別に法人企業統計調査の母集団法人数に合わせる形で重みづけ推計を行う。具体的には「再現データの各区分での集計値÷同集計法人数×法人企業統計調査の母集団法人数」で拡大推計値を計算する。法人企業統計の四半期調査では、資本金 1000 万円未満の企業は対象としていないため、拡大推計は資本金 1000 万円以上 10 億円未満、1 億円以上 10 億円未満、10 億円以上の 3 区分で実施する。また、比較は金融業・保険業を除く系列同士で比較する。

図表５－１－１に、再現データのサンプルと、法人企業統計調査の母集団の業種別、資本金区分別の企業数を示した。拡大推計には使用しないが、金融業、保険業や資本金 1 千万未満の企業数も示している。表内の赤の帯は業種別の、緑の帯は資本金区分別の、青の帯は業種別×資本金区分別の企業数の相対的な大きさを表す。

再現データのサンプルサイズは、全体で約 6.4 万社である。再現データは法人企業統計の母集団と比較して企業数の分布が大きく異なる訳ではないが、

- 資本金規模が大きい企業が多い[35]
- 業種別には建設業が少なく、製造業、卸売業・小売業が多い

といった特徴がみられる。この点は３．（３）で経済センサスを用いて確認した結果と概ね整合的である。

[35] 再現データでは資本金 10 億円以上、1～10 億円の区分に相当数の企業があり、一方、1 千万円未満の企業は少ない。

（図表５－１－１　再現データと法人企業統計調査の業種別、資本金区分別の企業数）

再現データのサンプル企業数

	10億円以上	1億円以上 - 10億円未満	1千万円以上 - 1億円未満	1千万円未満	総計
農林水産業	18		127	193	338
鉱業、採石業、砂利採取業	5	9	23	43	80
建設業	19	394	3,300	3,081	6,794
製造業	306	1,455	6,733	3,322	11,816
電気・ガス・熱供給・水道業	30	64	115	243	452
情報通信業	107	567	1,937	1,117	3,728
運輸業、郵便業	68	258	1,501	1,039	2,866
卸売業・小売業	105	1,069	10,896	5,751	17,821
金融業、保険業	39	69	149	488	745
不動産業、物品賃貸業	121	483	3,409	3,288	7,301
学術研究、専門・技術サービス業	16	107	1,444	1,281	2,848
宿泊業、飲食サービス業	19	62	852	1,106	2,039
生活関連サービス業、娯楽業	11	68	829	899	1,807
教育、学習支援業	4	6	127	360	497
医療、福祉業	8	59	277	1,142	1,486
その他のサービス業	34	181	1,814	1,387	3,416
総計	893	4,868	33,533	24,740	64,034

法人企業統計調査の母集団企業数

	10億円以上	1億円以上 - 10億円未満	1千万円以上 - 1億円未満	1千万円未満	総計
農林水産業	5	169	5,806	28,758	34,738
鉱業、採石業、砂利採取業	31	41	1,824	1,286	3,182
建設業	231	1,429	146,066	344,120	491,846
製造業	1,811	6,007	140,246	172,040	320,104
電気・ガス・熱供給・水道業	132	501	1,724	11,851	14,208
情報通信業	447	3,347	36,197	90,583	130,574
運輸業、郵便業	250	1,168	36,441	42,953	80,812
卸売業・小売業	643	4,682	205,815	381,585	592,725
金融業、保険業	830	1,496	8,742	59,711	70,779
不動産業、物品賃貸業	366	2,921	116,269	275,848	395,404
学術研究、専門・技術サービス業	511	2,054	67,248	243,908	313,721
宿泊業、飲食サービス業	54	592	26,249	127,598	154,493
生活関連サービス業、娯楽業	62	976	29,255	115,443	145,736
教育、学習支援業	17	124	5,064	20,723	25,928
医療、福祉業	14	323	8,204	59,060	67,601
その他のサービス業	156	1,385	42,038	104,120	147,699
総計	5,560	27,215	877,188	2,079,587	2,989,550

（備考）赤の帯は業種別の、緑の帯は資本金区分別の、青の帯は業種別×資本金区分別の企業数の相対的な大きさを表す。

② 再現データと法人企業統計の比較方法

比較する指標は、売上高、営業利益、人件費とする。全業種（金融業・保険業は除く）での比較の他、再現データで比較的サンプルが多い製造業、卸売業・小売業、不動産業・物品賃貸業、サンプルサイズは少ないが新型コロナウイルス感染症の影響が大きく注目度の高い宿泊業・飲食サービス業を対象に、業種別の比較も行う。

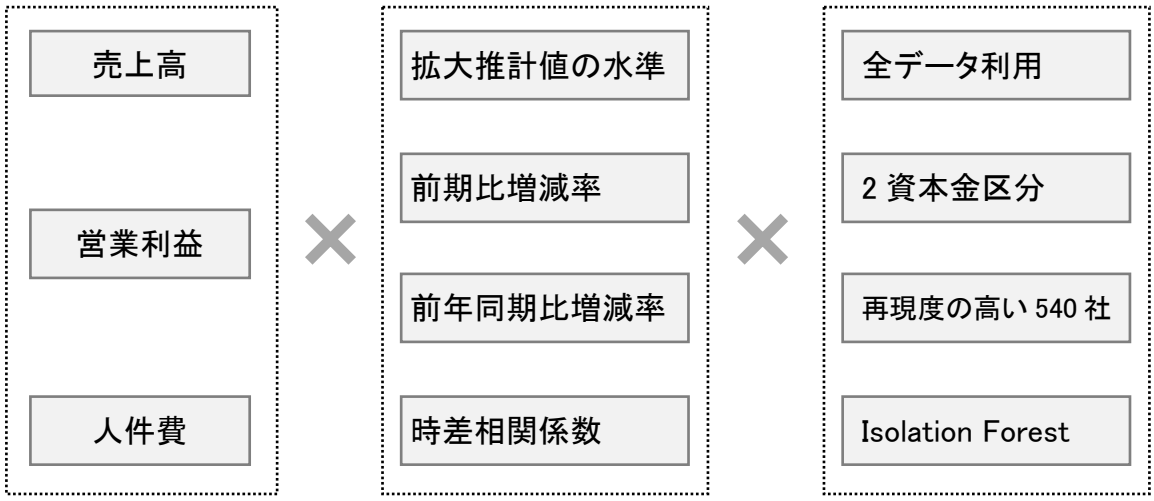
全サンプルを用いた場合の他、以下のパターンでサンプルを絞った場合について比較する。

- サンプルの多い資本金 1 千万円～1 億円、1 億円～10 億円の 2 区分の合算値（サンプルサイズ：約 3.8 万社）
- ４．で検討した再現度の高い 540 社
- Isolation Forest で異常値を各月で除去したサンプル

各指標について拡大推計値の水準、増減率（前期比、前年同期比）、時差相関係数をそれぞれ比較した。

図表５－１－２に比較を行う組み合わせを示した。以下では主な結果を抜粋して紹介する。

（図表５－１－２　再現データと法人企業統計の比較方法）



（２）　　全業種・全サンプルの比較

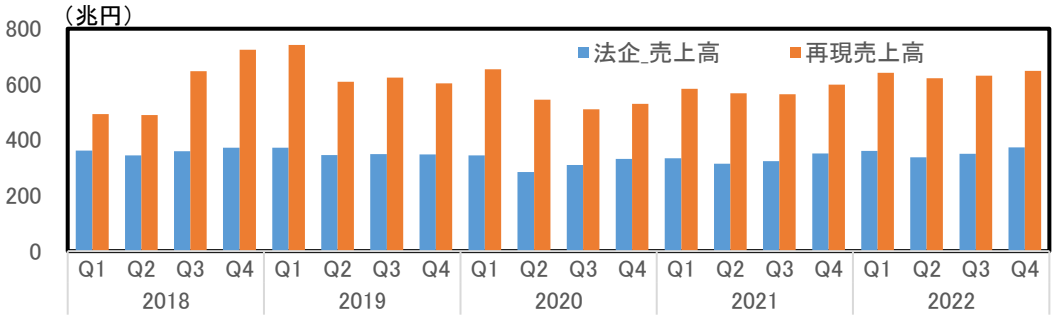
①　水準の比較

全業種・全サンプルデータを用いた再現売上高、再現営業利益、再現人件費の拡大推計値の水準を法人企業統計と比較すると（図表５－２－１）、

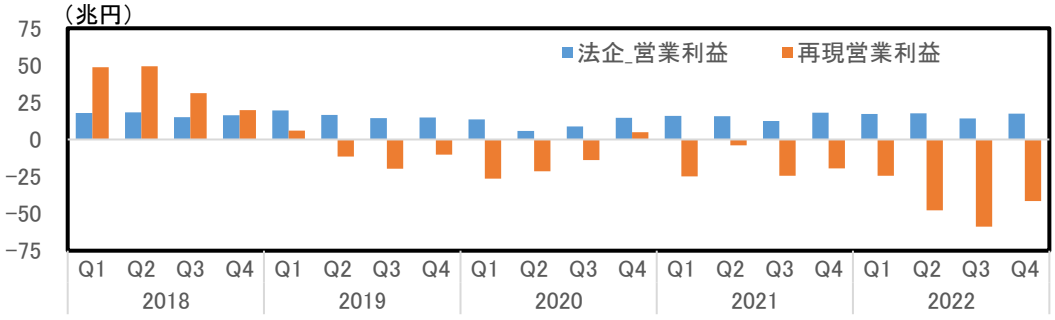
- 売上高は再現データの方が大幅に高い水準となった。
- 営業利益は、再現データではマイナス（損失）となる期が多く、プラス（利益）が続く法人企業統計とは異なる結果となった。
- 人件費は再現データの方が大幅に低い水準となった。

（図表５－２－１　再現データと法人企業統計の比較（水準））

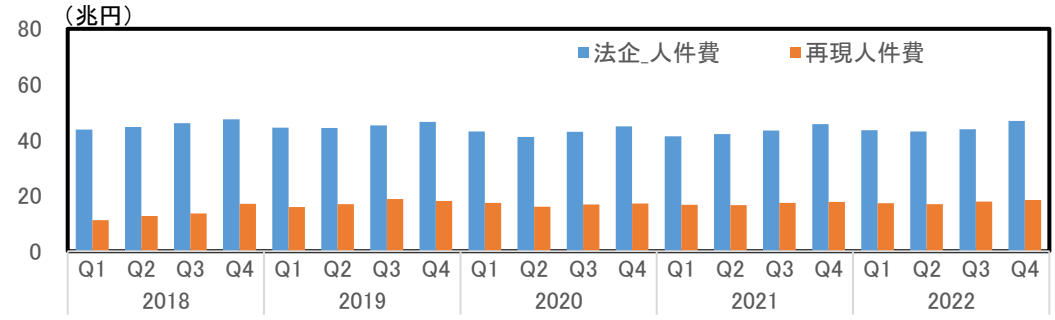
売上高



営業利益



人件費

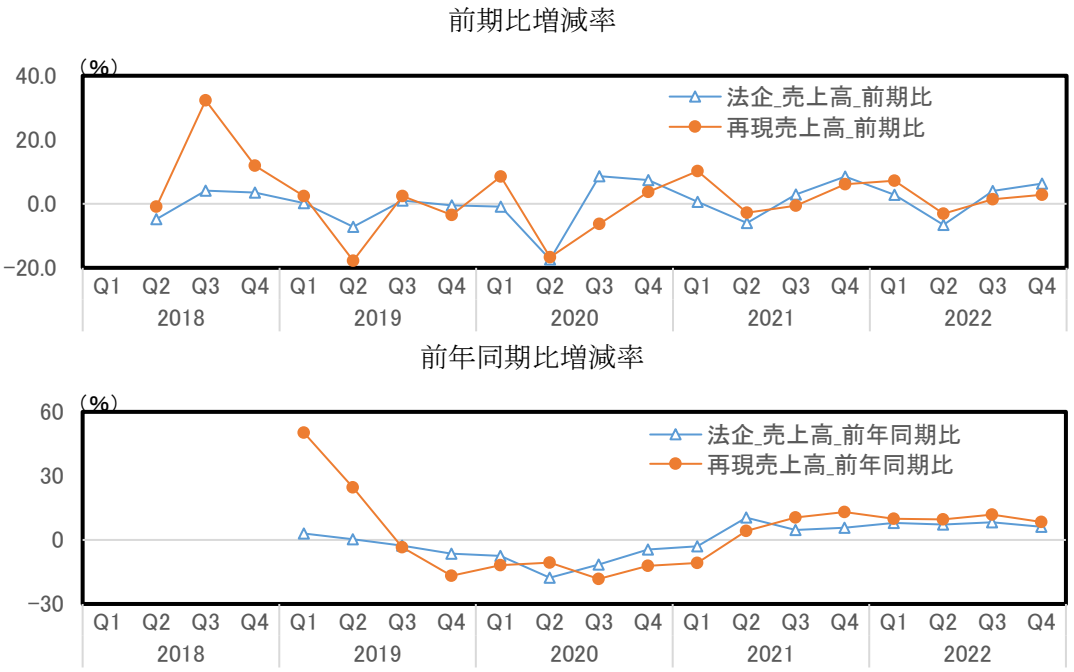


②　動向の比較：売上高

売上高について再現データと法人企業統計の動向を前期比増減率、前年同期比増減率により比較すると（図表５－２－２）、両者とも、多くの期間で類似した動きをしている。時差相関係数[36]も当期が最も高く、概ね一致したタイミングで推移していることを示している。ただし、銀行口座データは 2018 年のサンプルが少ないことから、2019 年の前年同期比は大きな変動を示している。

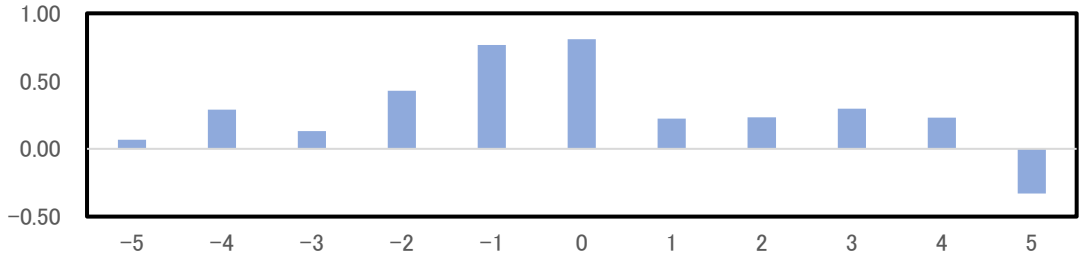
法人企業統計の前年同期比増減率は、新型コロナウイルス感染症の影響により 2020 年４－６月期に大きなマイナスとなり、その後回復に向かっているが、銀行口座データにおいても、タイミングはやや遅れているものの、落ち込みと回復の動きがみられる。銀行口座データと財務データの比較の際に検討したように、銀行口座データによる再現売上高は現金の動きであるキャッシュフローを捉えていると考えられるため、取引時点の記録である会計上の動きを捉える法人企業統計に対して若干のラグを持つと考えられる。

（図表５－２－２　再現データと法人企業統計の比較（売上高の動向））



36　再現データはサンプルの少ない 2018 年を除いた 2019 年から 2022 年の 16 四半期の値を、法人企業統計は 2017 年 10～12 月期から 2024 年 1～3 月期の値を用い、再現データを固定して法人企業統計を当期を中心に過去、将来に 1 四半期ずつずらしながら相関係数を計算した。グラフ横軸の数値は法人企業統計を過去（-1、-2…）、将来（1、2…）にずらした期の数を表す。

時差相関係数

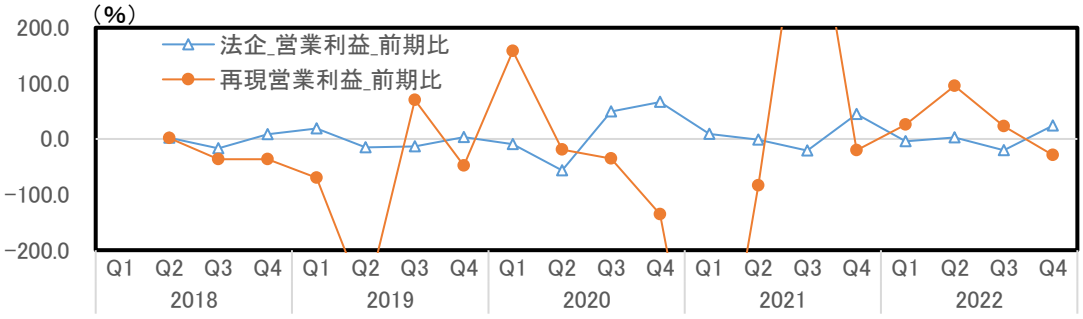


③　動向の比較：営業利益

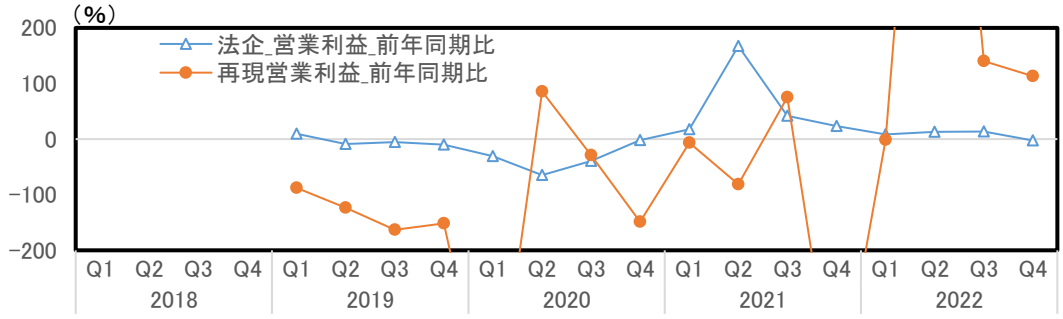
営業利益の動向については（図表５－２－３）、再現データは前期比増減率、前年同期比増減率とも法人企業統計と比べて変動が非常に大きく、時差相関係数も当期はゼロに近く、他の期もマイナスの値が多いなど、連動した動きは見られない。

（図表５－２－３　再現データと法人企業統計の比較（営業利益の動向））

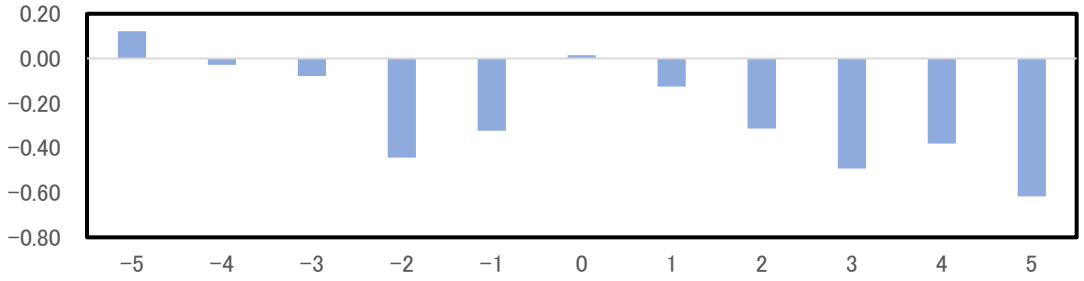
前期比増減率



前年同期比増減率



時差相関係数



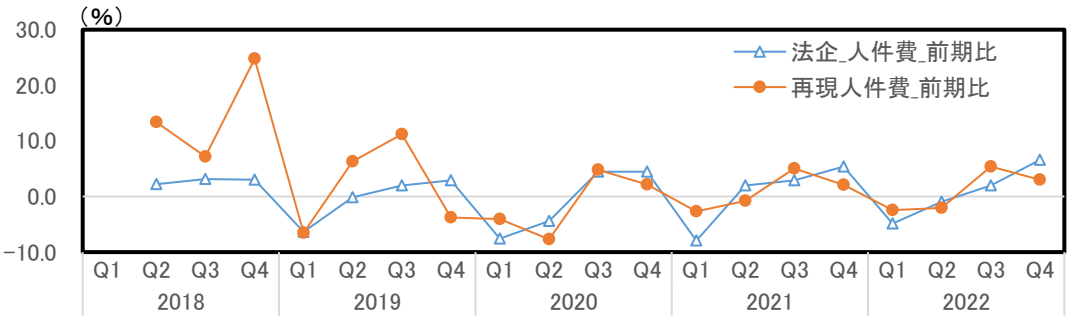
④　動向の比較：人件費

人件費の動向を前期比増減率、前年同期比増減率により比較すると、両者とも、多くの期間で類似した動きをしている。時差相関係数も当期が最も高く、概ね一致したタイミングで推移していることを示している（図表５－２－４）。

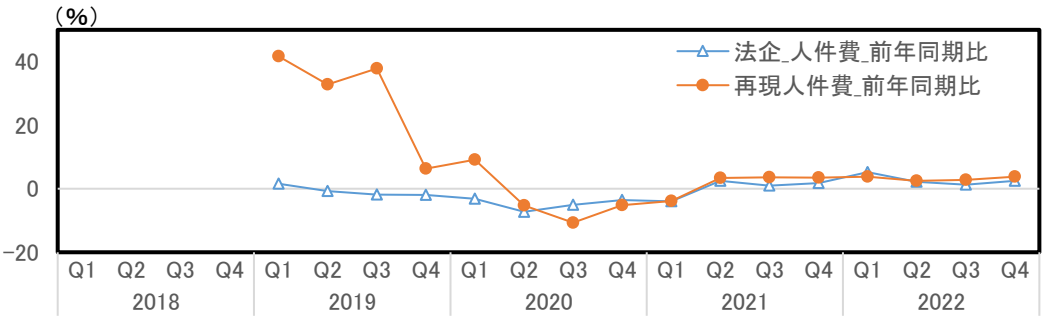
再現データは 2018 年のサンプルが少ない影響もあり、人件費は特に法人企業統計との乖離が大きく、2019 年の前年同期比増減率が大きく変動している。このため、以降の検討では 2020 年以降のデータを用いる。

（図表５－２－４　再現データと法人企業統計の比較（人件費の動向））

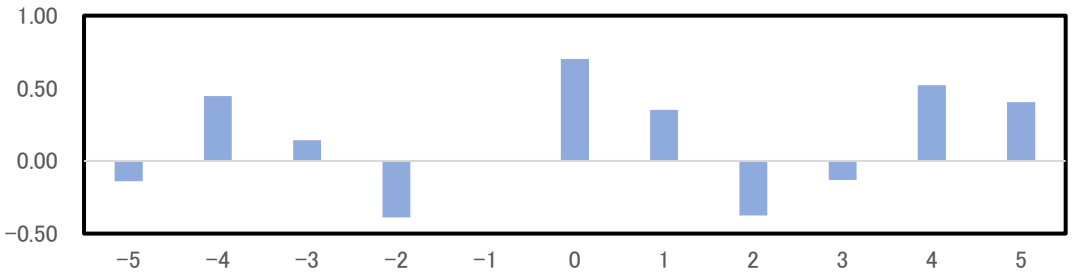
前期比増減率



前年同期比増減率



時差相関係数



（３）　抽出条件別の比較（全業種）

サンプルの絞り込み方法によって法人企業統計との誤差がどの程度縮小するか検討するため、以下の４つの抽出パターンごとに法人企業統計との比較を行った。法人企業統計との誤差は、平均絶対誤差（MAE、Mean Absolute Error）で評価した。

- パターン１：全サンプル
- パターン２：２資本金区分の合算値（資本金１千万円～１億円、１億円～10 億円）
- パターン３：４．で検討した再現度の高い 540 社のサンプル
- パターン４：Isolation Forest により異常値を除いたサンプル

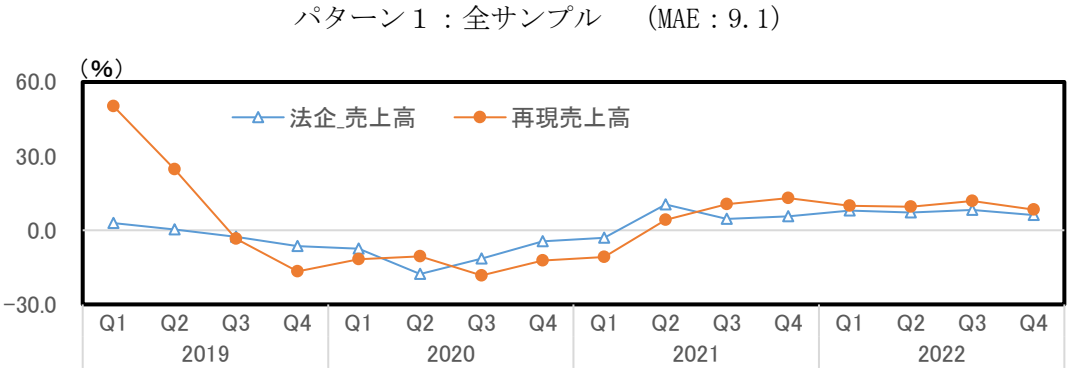
拡大推計値の水準、前期比増減率、前年同期比増減率、時差相関係数のそれぞれの比較を行ったが、以下では、時系列の動向を見る際に重要と思われる、前年同期比増減率の結果を紹介する。

①　売上高

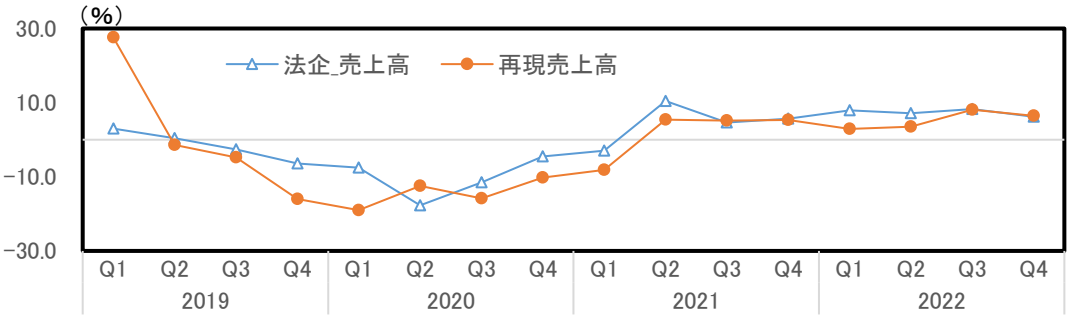
売上高の前年同期比増減率について、４つのパターンごとに作成した再現データと法人企業統計を比較すると、全サンプルを使用した場合より、サンプルを絞り込んだ方が法人企業統計との誤差が小さくなる結果となった（図表５－３－１）。特に、再現度の高い 540 社を使用したパターン３、Isolation Forest により異常値を除いたパターン４で、MAE が顕著に縮小している。

新型コロナウイルス感染症の影響による 2020 年の売上の落ち込みやその後の回復も、パターン３、４ではより法人企業統計に近い動きになっている。ただし、例えばパターン３で、売上が最も減少した期は、法人企業統計では 2020 年４～６月期であるのに対し、キャッシュフローの動きをとらえる再現売上高では７～９月期になっており、指標の変動のタイミングにはやや相違がみられる。

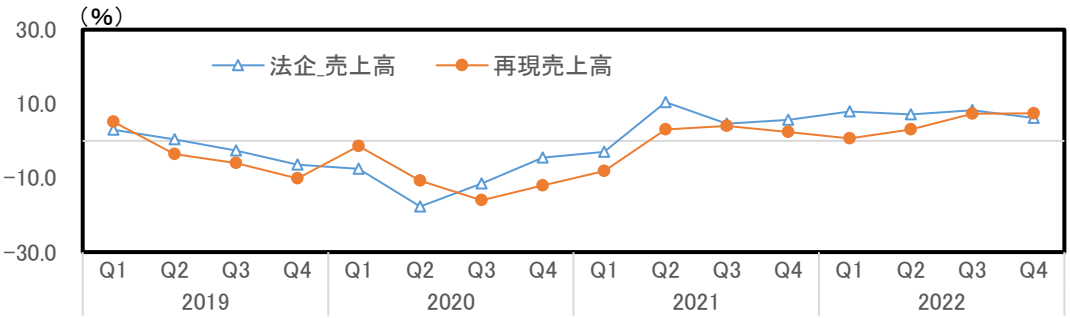
（図表５－３－１　売上高の前年同期比増減率、４パターンの比較）



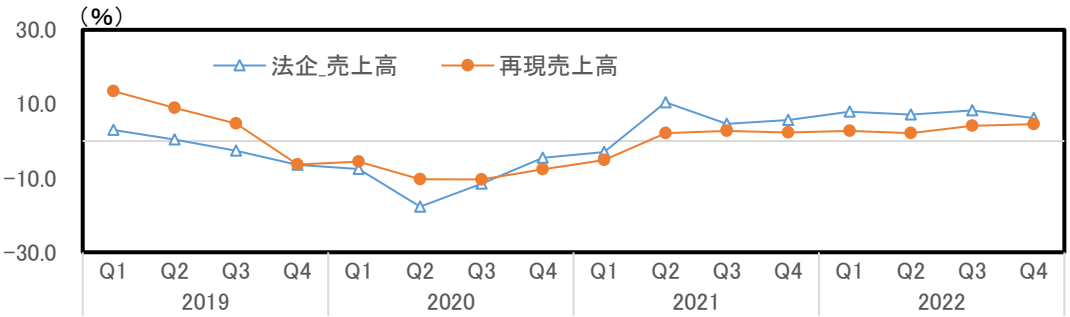
パターン２：２資本金区分　（MAE：5.3）



パターン３：再現度の高い 540 社　（MAE：4.3）



パターン４：Isolation Forest　（MAE：4.5）

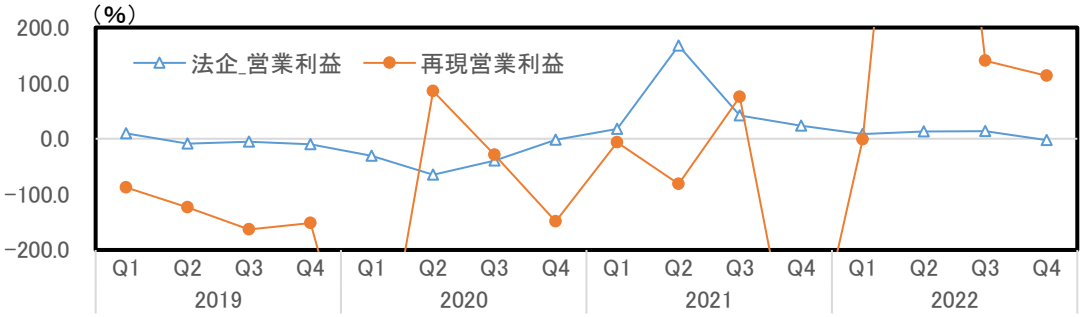


②　営業利益

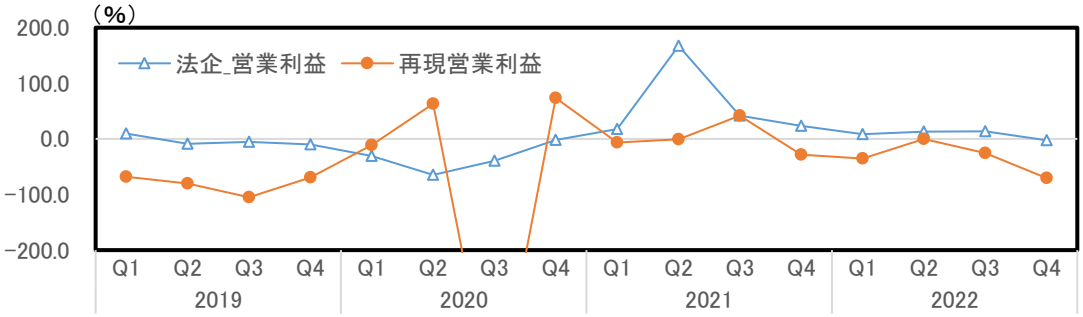
営業利益の前年同期比増減率については、抽出方法を変えても法人企業統計と類似した動きを再現することは困難であったが、Isolation Forest を適用したパターン４で法人企業統計との誤差が顕著に改善した（図表５－３－２）。営業利益では、売上高や人件費と比較して変動が大きく、増減率が極端に大きい（小さい）値をとるサンプルがあり、それが全体の増減率にも影響している可能性がある。Isolation Forest により異常値を除くことで、他の抽出方法に比べれば法人企業統計と近い推移のグラフとなっているが、増減の方向性を的確に捉えるには至らなかった。

（図表５－３－２　営業利益の前年同期比増減率、４パターンの比較）

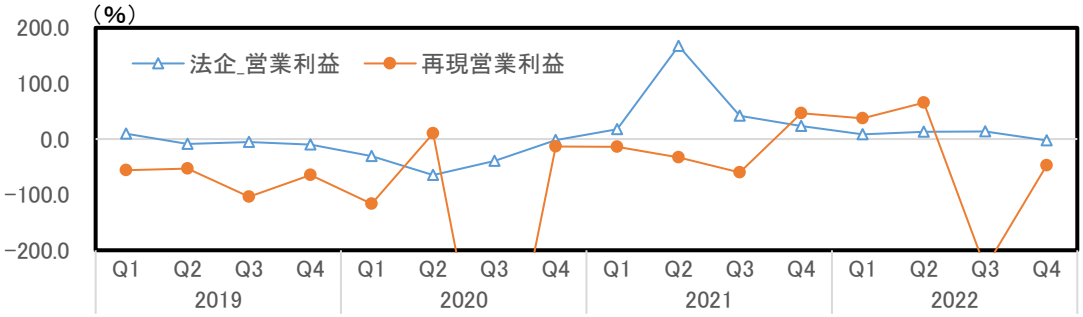
パターン１：全サンプル　（MAE：218.0）



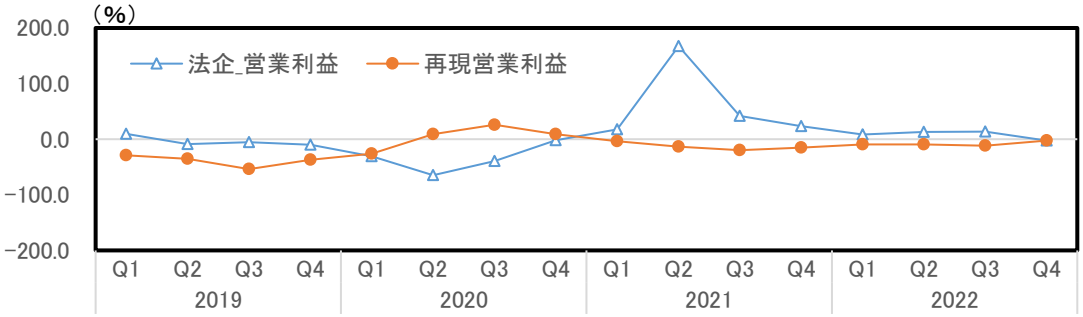
パターン２：２資本金区分　（MAE：87.5）



パターン３：再現度の高い 540 社　（MAE：105.1）



パターン４：Isolation Forest　（MAE：41.5）

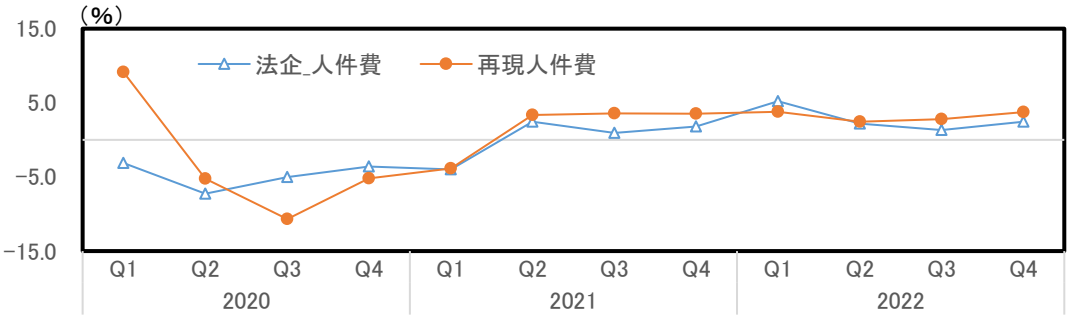


③　人件費

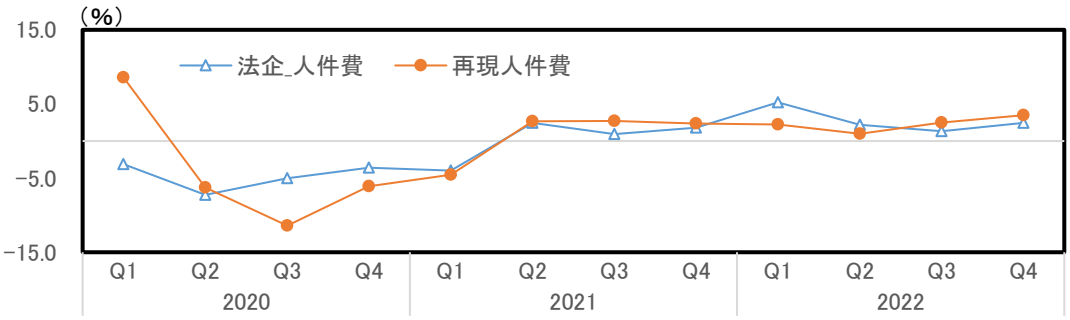
人件費はサンプルの関係で 2019 年の前年同期比が大きく変動しているため、2020 年以降の推移で比較している。人件費は、パターン１から４で同程度の誤差となった。2020 年以降の前年同期比増減率では全サンプルでも法人企業統計と比較的近い推移をしており、サンプルの抽出方法を変えても誤差は変わらなかった（図表５－３－３）。

（図表５－３－３　人件費の前年同期比増減率、４パターンの比較）

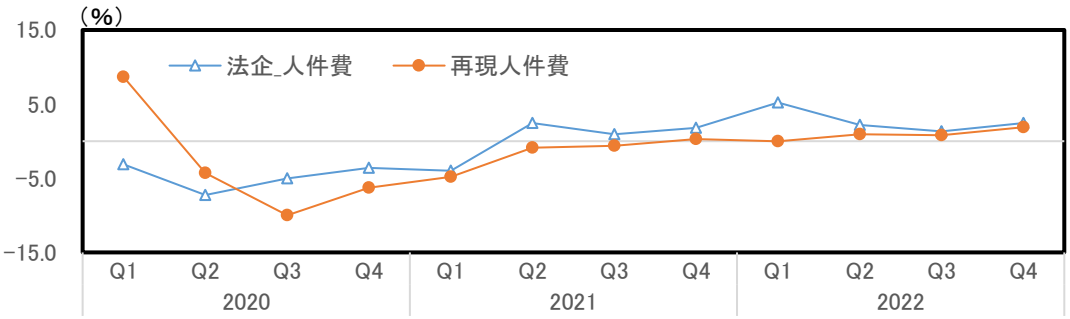
パターン１：全サンプル　（MAE：2.6）



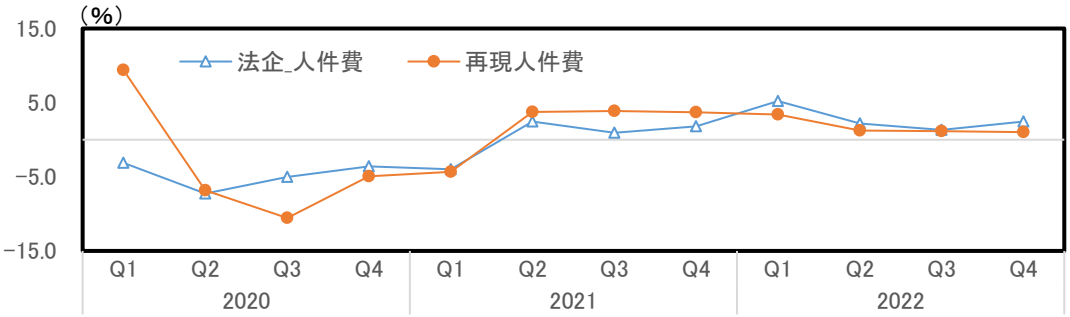
パターン２：２資本金区分　（MAE：2.6）



パターン３：再現度の高い 540 社　（MAE：3.1）



パターン４：Isolation Forest　（MAE：2.5）

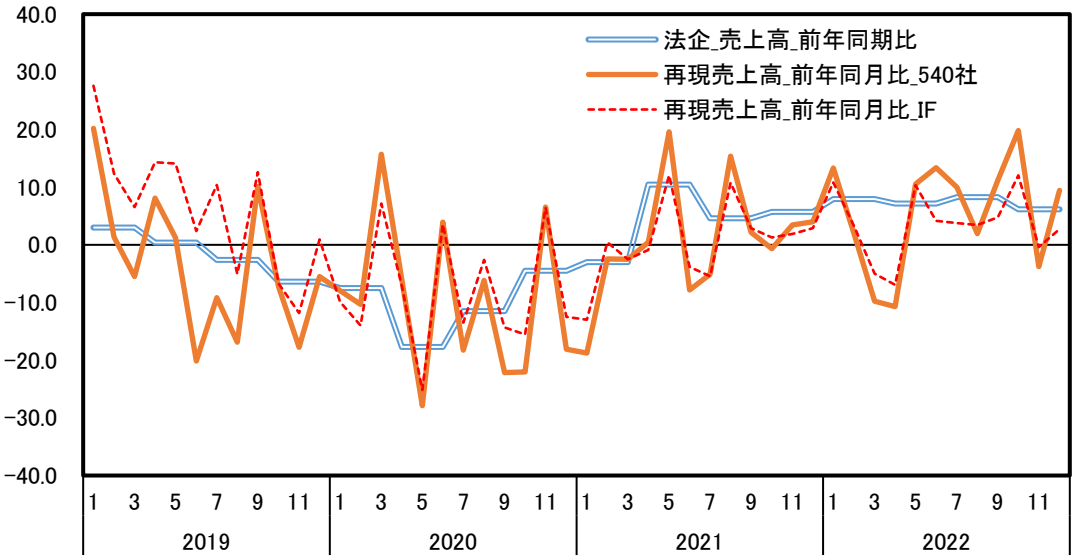


④　月次の推移（売上高、前年同月比増減率）

企業業績のデータは年や四半期が中心であるため、銀行口座データを用いて、より高頻度のデータが作成できれば有用な情報となる可能性がある。上記では法人企業統計と比較するため四半期データを作成したが、経済動向把握のために重要であり、法人企業統計と整合性が高かった再現売上高の前年同期比増減率について、MAE も小さかったパターン３、４の月次データを作成した。

結果は前年同月比増減率であるにも関わらず四半期データと比べると変動が非常に大きく、傾向を読み取りにくかった（図表５－３－４）。月次ベースで有用なデータを作成するには、変動の原因等に関する検討が必要と考えられる。

（図表５－３－４　売上高前年同月比増減率の月次推移）



（４）　　抽出条件別の比較（売上高の業種別比較）

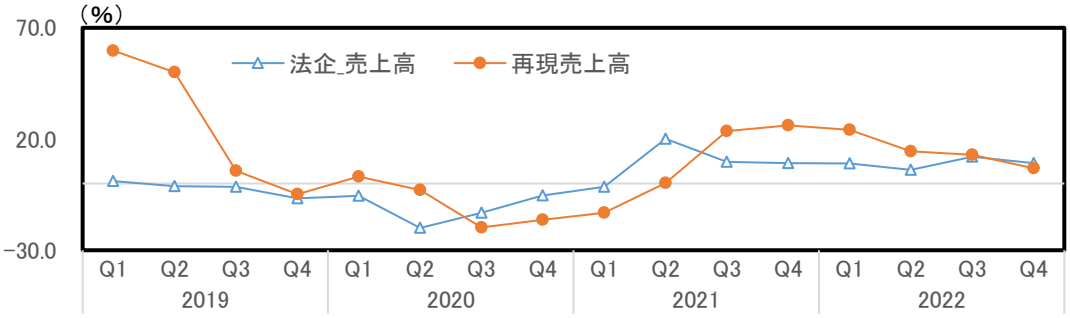
再現データと法人企業統計の整合性に業種別に違いがあるかどうかを確認するため、業種別の比較を行う。以下では、経済動向を見る上で重要と考えられる売上高の前年同期比増減率について、（３）で用いた４つのサンプル抽出パターンごとの比較の結果を紹介する。

①　　製造業

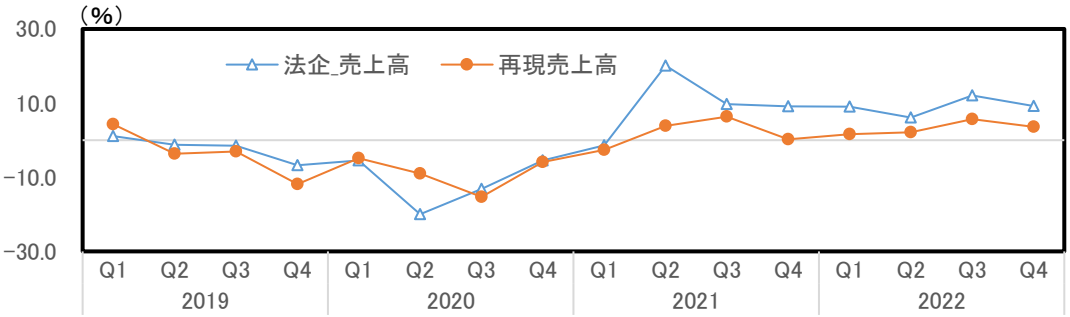
製造業の全サンプル（パターン１）の誤差は全業種より大きい（全業種の MAE：9.1%、製造業の MAE：15.7%）が、2020 年の売上高の落ち込みやその後の回復といった動きはある程度再現できている（図表５－４－１）。サンプル抽出方法については、資本金区分を２区分に絞ったパターン２、Isolation Forest により異常値を除いたパターン４で、MAE が顕著に縮小している。パターン２で改善が見られたのは、10 億円以上の区分で再現売上高が非常に大きいサンプルが多く、推計結果に影響を及ぼしている可能性が考えられる。キャッシュフローを捉える再現売上高の 2020 年の落ち込み、改善のタイミングが法人企業統計より遅れる点は、サンプルの抽出方法を変えても同様であった。

（図表５－４－１　製造業売上高の前年同期比増減率、４パターンの比較）

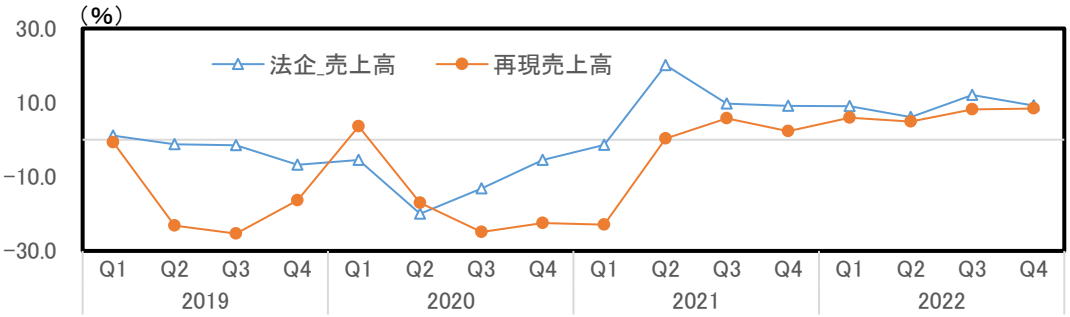
パターン１：全サンプル　（MAE：15.7）



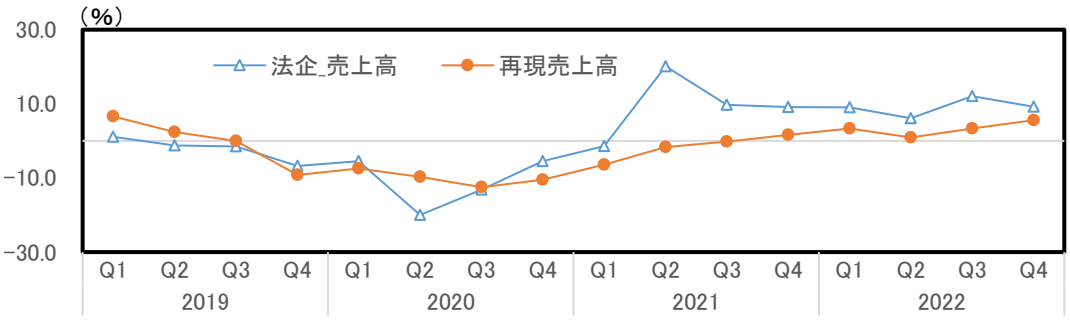
パターン２：２資本金区分　（MAE：5.0）



パターン３：再現度の高い 540 社　（MAE：10.0）



パターン４：Isolation Forest　（MAE：6.1）



②　卸売業・小売業

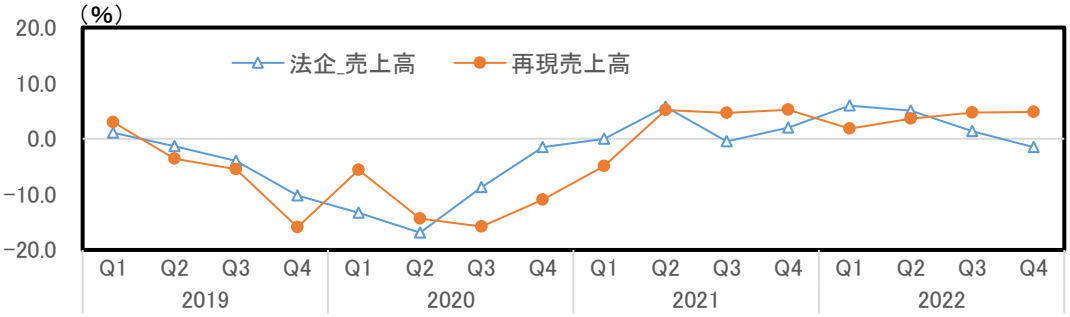
卸売業・小売業では、全サンプル（パターン１）の場合でも全業種に比べて誤差が小さい（全業種の MAE：9.1%、卸売業・小売業の MAE：4.2%）。小売業で現金決済の比率が高いため、取引と入出金のタイミングが近いことが背景として考えられる（図表５－４－２）。

サンプルの抽出パターン別では、Isolation Forest を使ったパターン４で誤差が最も小さかった。異常値が推計精度に影響を及ぼしており、それらが除外できたことで誤差が縮小した可能性がある。また、パターン４では、再現データで新型コロナウイルス感染症の影響による落ち込みが最も大きかった期が 2020 年４－６月期であり、法人企業統計と一致している。

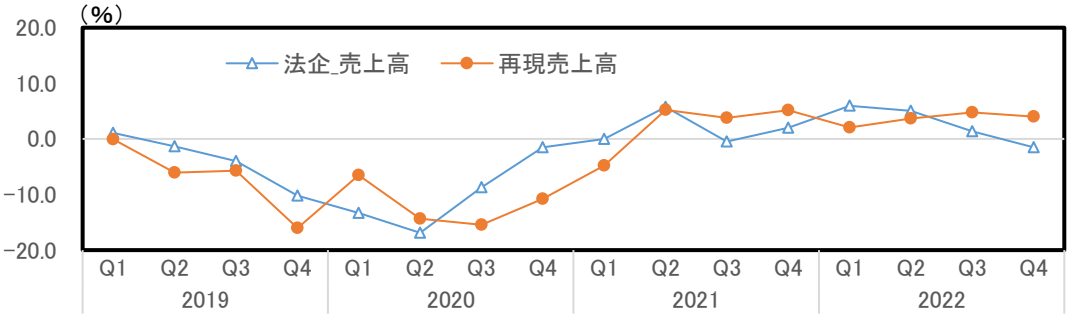
以上から、卸売業・小売業では、他の業種に比べ、銀行口座データによって経済変動をより適切に把握できる可能性が示唆される。

（図表５－４－２　卸売業・小売業売上高の前年同期比増減率、４パターンの比較）

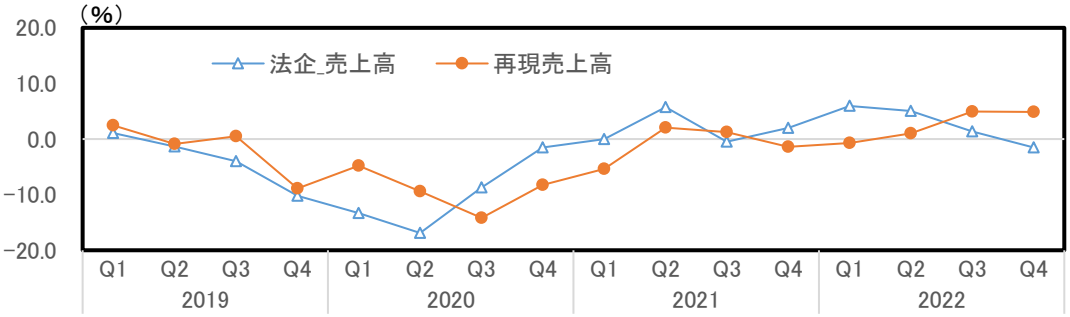
パターン１：全サンプル　（MAE：4.2）



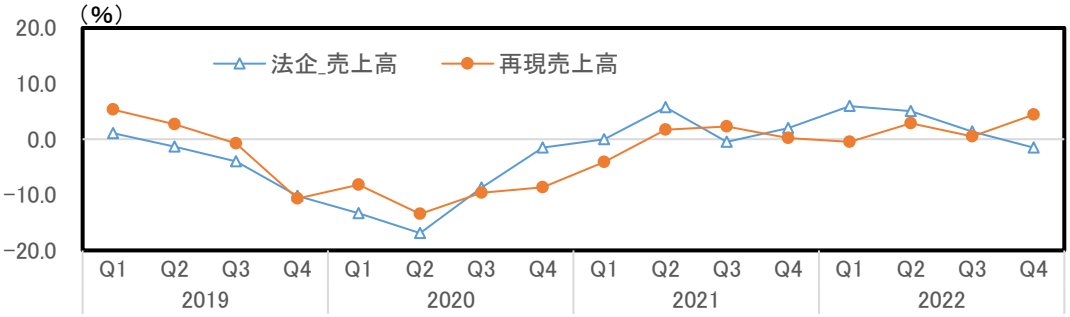
パターン２：２資本金区分　（MAE：4.1）



パターン３：再現度の高い 540 社　（MAE：4.4）



パターン４：Isolation Forest　（MAE：3.5）



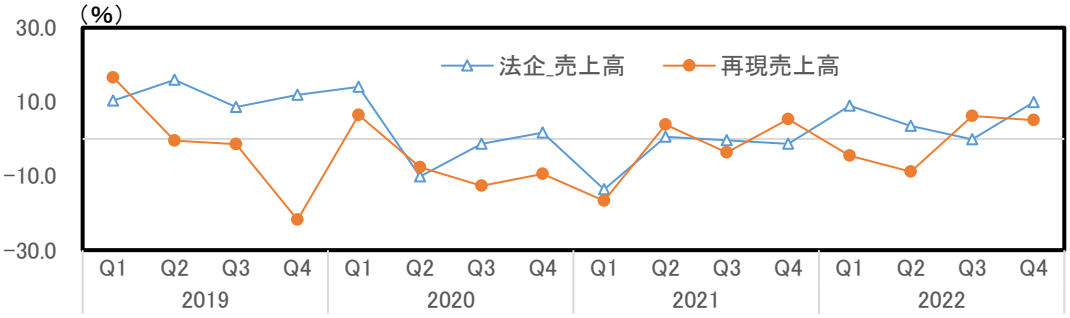
③　不動産業・物品賃貸業

不動産業・物品賃貸業の全サンプル（パターン１）の誤差は全業種と同程度である（全業種の MAE：9.1%、不動産業・物品賃貸業の MAE：9.5%）が、前年同期比増減率の上昇、下落の動きは必ずしも一致しておらず、法人企業統計の動向を再現できているかは不明瞭であった（図表５－４－３）。

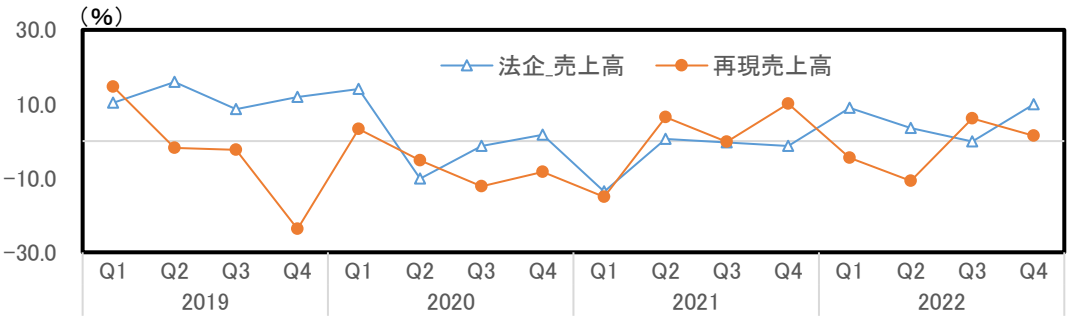
サンプル抽出方法を変えたパターン２、パターン３では、法人企業統計との誤差は特に改善しなかった。４．で検討した再現度の高い 540 社によるパターン３では、業種を絞るとサンプルサイズが非常に小さくなったため、増減率の変動が非常に大きく、MAE は全サンプルより大幅に悪化した。

（図表５－４－３　不動産業・物品賃貸業売上高の前年同期比増減率、４パターンの比較）

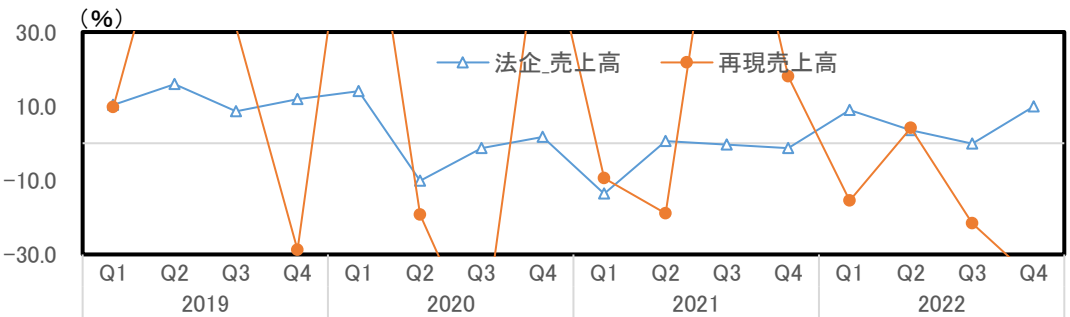
パターン１：全サンプル　（MAE：9.5）



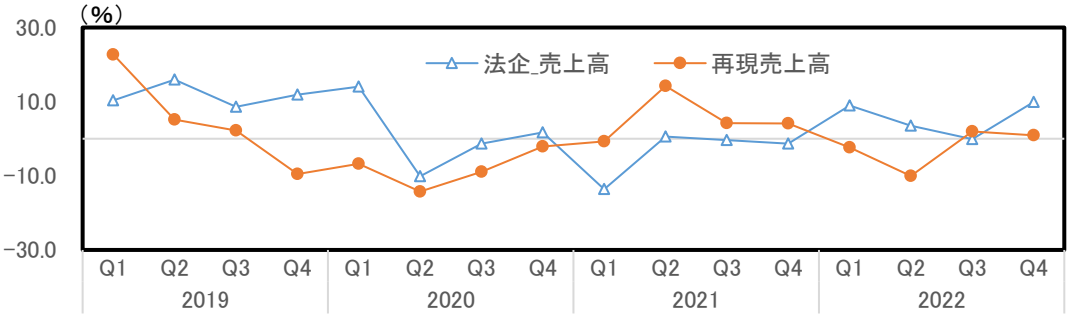
パターン２：２資本金区分　（MAE：10.4）



パターン３：再現度の高い 540 社　（MAE：34.2）



パターン４：Isolation Forest　（MAE：10.0）



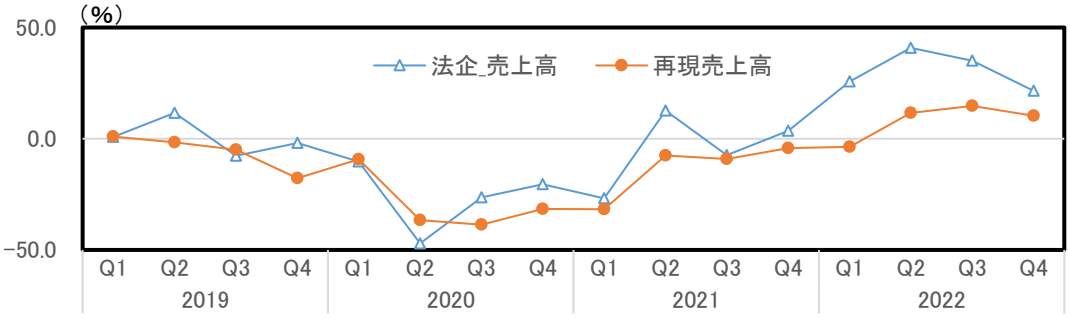
④　宿泊業・飲食サービス業

宿泊業・飲食サービス業の全サンプル（パターン１）の誤差は全業種よりやや大きいが（全業種の MAE：9.1%、宿泊業・飲食サービス業の MAE：12.0%）、2020 年の売上高の落ち込みやその後の回復の動きは比較的似通っている（図表５－４－４）。宿泊・飲食サービス業でも、現金決済の比率が高いことが背景として考えられる。

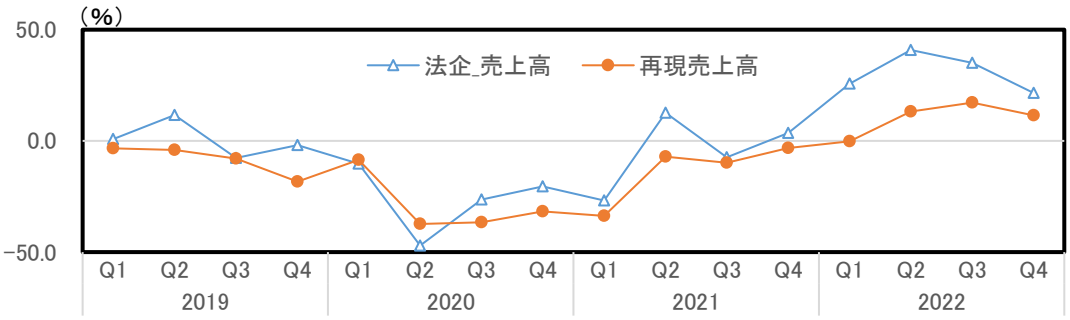
サンプル抽出方法を変えたパターン２、パターン４では、法人企業統計との誤差は特に改善しなかった。４．で検討した再現度の高い 540 社によるパターン３では、MAE は全サンプルより悪化した。業種を絞ったためサンプルサイズが非常に小さくなった影響も考えられる。このため、精度には問題があるが、新型コロナウイルス感染症の影響による落ち込みが最も大きかった期が 2020 年４－６月期と法人企業統計と一致するなど、興味深い動きもみられる。

（図表５－４－４　宿泊・飲食サービス業売上高の前年同期比増減率、4 パターンの比較）

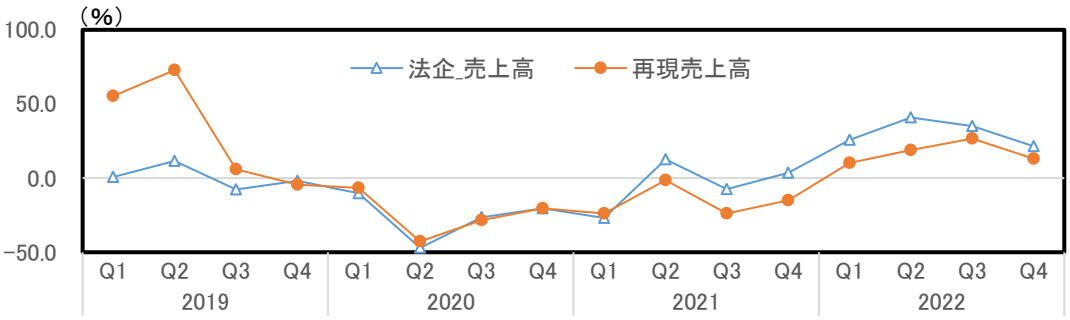
パターン 1：全サンプル　（MAE：12.0）



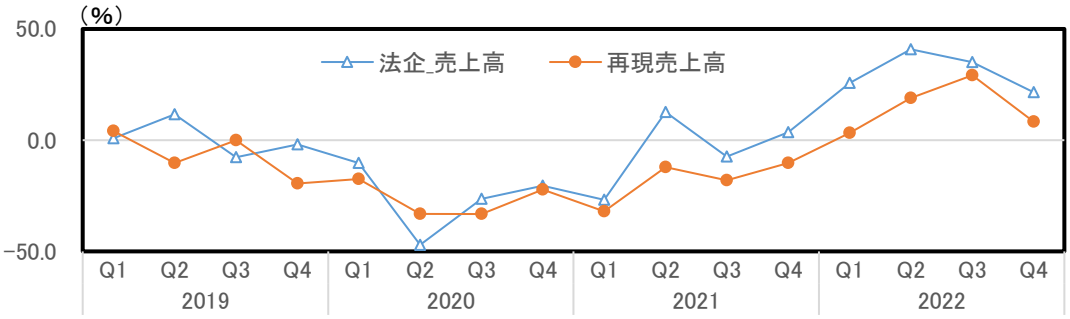
パターン 2：2 資本金区分　（MAE：11.7）



パターン 3：再現度の高い 540 社　（MAE：15.5）



パターン 4：Isolation Forest　（MAE：12.4）



（備考）パターン３のみグラフのスケールが異なることに留意

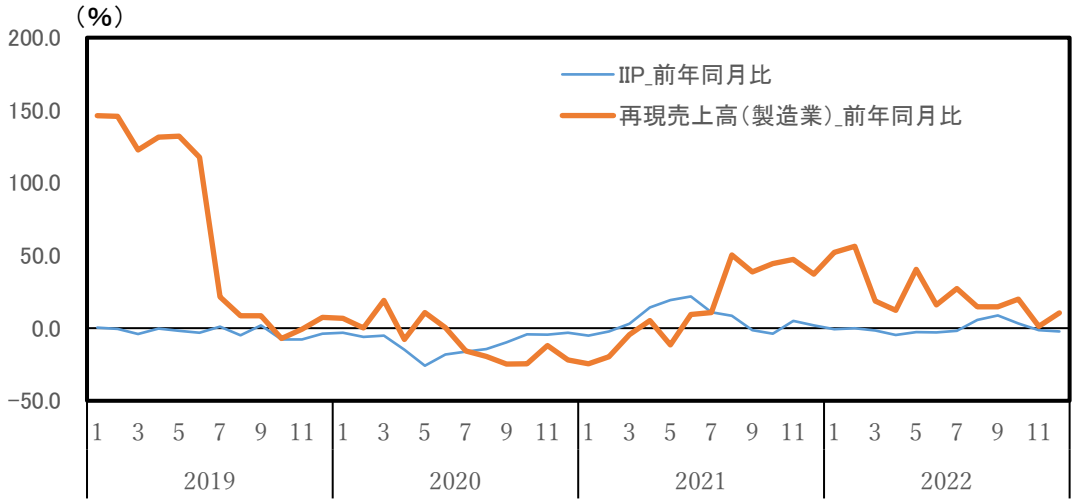
（５）　　他指標との比較

再現データから月次指標を作成し、他の経済指標との比較を行う。経済指標としては、鉱工業生産指数、商業動態統計、総雇用者所得を用い、再現データのうち、概念の近い指標と比較する。具体的には、鉱工業生産指数は製造業の再現売上高と、商業動態統計は卸売業・小売業の再現売上高と、総雇用者所得は全業種の再現人件費と比較する。なお、比較対象データの資本金別・業種別のサンプル構成の情報が得られないなど、再現データの拡大推計を行うことが困難であるため、拡大推計は行わない。

①　　鉱工業生産指数

再現データによる製造業売上高の前年月比増減率を鉱工業生産指数と比較すると、再現データは鉱工業生産より変動が大きく、やや遅れて変動していることが示唆された[37]（図表５－５－１）。

（図表５－５－１　再現データ（製造業、売上高）と鉱工業生産指数の比較）

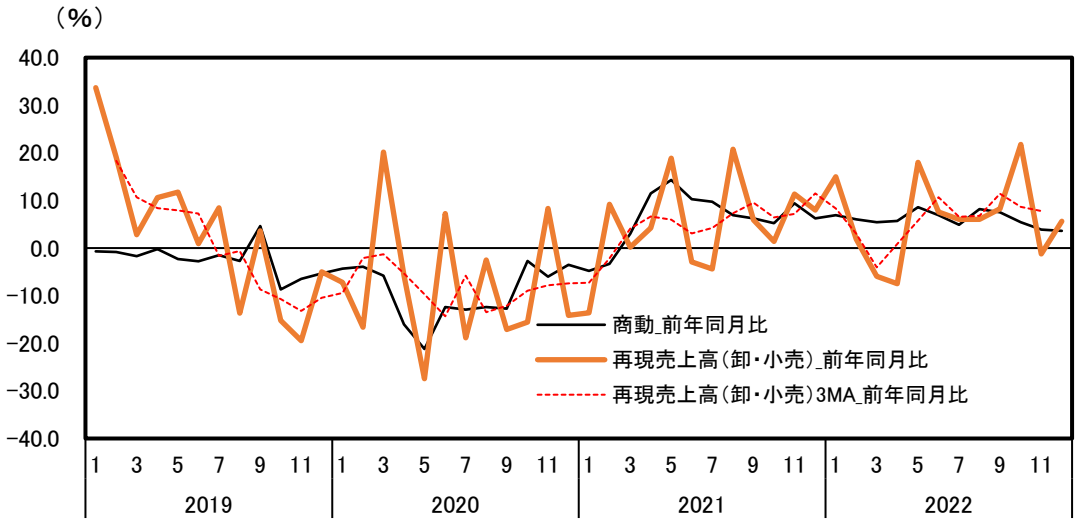


②　　商業動態統計

再現データによる卸売業・小売業売上高の前年同月比増減率を商業動態統計と比較すると、再現データは月ごとに大きく変動しているが、便宜的に３か月移動平均の前年同月比増減率をみると、2019 年後半以降は概ね近い推移をしている（図表５－５－２）。

[37] 2019～2021 年の数値を用いて製造業の再現売上高と鉱工業生産指数の時差相関係数を計算すると、再現売上高は鉱工業生産の５か月前の値との相関が最も高かった。

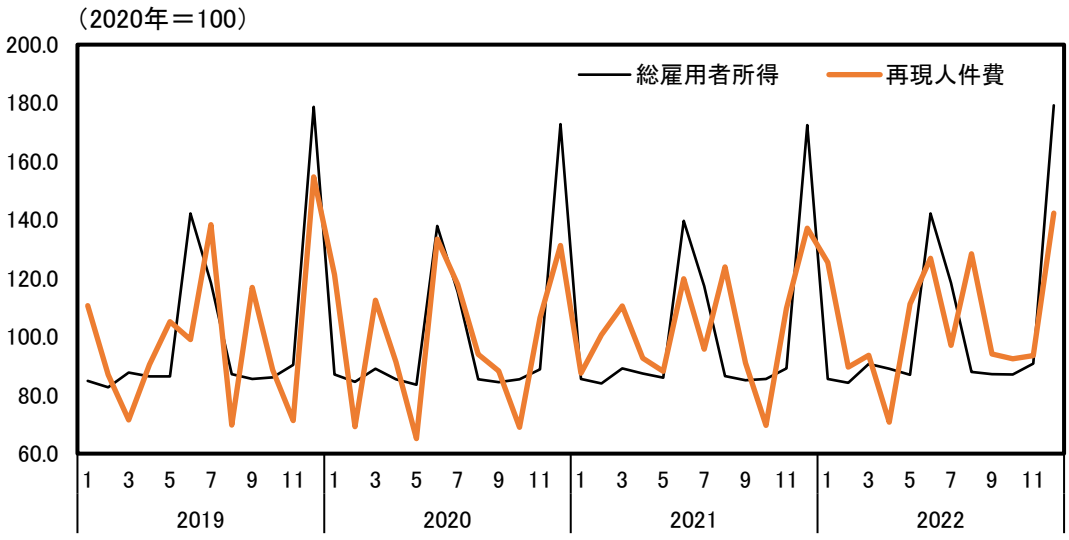
（図表５－５－２　再現データ（卸売・小売業、売上高）と商業動態統計の比較）



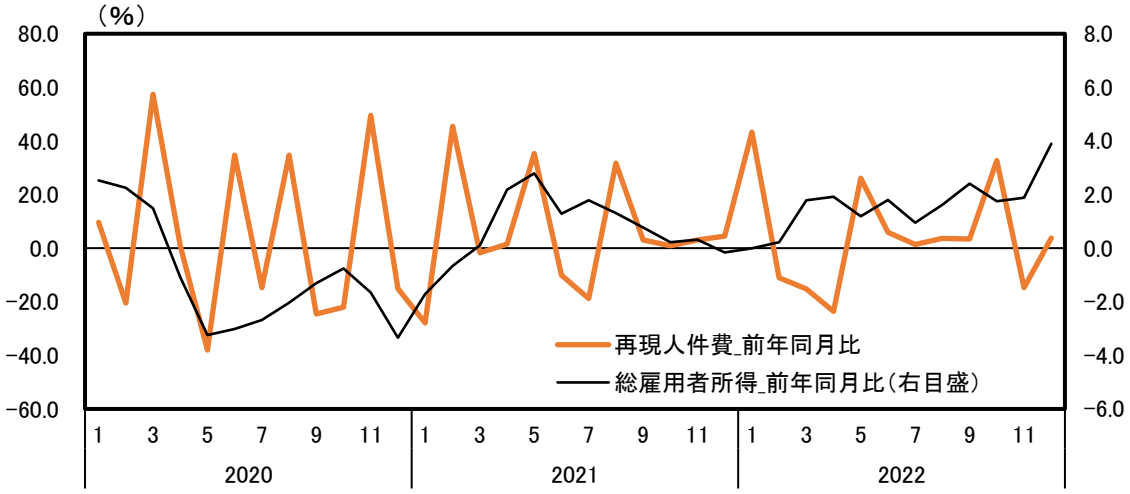
③　総雇用者所得

再現データによる人件費と総雇用者所得の名目値の水準を比較すると、賞与等で大きく上下する動きを概ね再現できていることが分かる（図表５－５－３）。一方、前年同月比増減率は大きく上下している上、総雇用者所得で見られる新型コロナウイルス感染症の影響による落ち込みや回復の動きは明瞭でない（図表５－５－４）。

（図表５－５－３　再現データ（全業種、人件費）と総雇用者所得の比較
（2020 年＝100 とした指数））



（図表５－５－４　再現データ（全業種、人件費）と総雇用者所得の比較
（前年同月比増減率））



６．まとめ

（銀行口座データの分布の特徴）

本稿では、リアルタイムデータを用いた経済動向分析の一環として、みずほ銀行の銀行口座データから中小企業のサンプルを抽出し、売上高、費用、営業利益、人件費、従業員数など主要な財務指標を再現し、実際の財務データと比較するとともに、経済動向分析データとしての有用性を検証した。

分析を進めるにあたり、まず、銀行口座データの企業数等の分布が実際の企業分布と整合的かどうか検証するため、経済センサスと比較した。銀行口座データの企業数の分布は経済センサスと比較して、①業種別の分布は似通っており、②所在地別の分布は東京都が多く、③資本金規模別で見ると規模の大きい企業が多く、④従業員規模別で見た場合も規模が大きい企業が多く、⑤個人企業の分布は大きく異なっている（例えば銀行口座データの企業数は不動産業・物品賃貸業に偏っている）、といった特徴があった。

（銀行口座データによる再現値と財務データの比較）

次に銀行口座データから、主要な財務指標等を再現する。まず日々の取引の記録である銀行口座データを企業ごとに集約した上で、一定期間の入金を売上高、出金を費用として集計する。その際、仕訳ルールを適用し、金融取引や税・補助金の受払や同一企業間の取引は除く。売上高から費用を差し引き、営業利益とみなす。また、給与支払いの合計を人件費、月ごとの給与振込件数の合計を従業員数とみなす。

こうして再現したデータを銀行が保有する財務諸表の数値と比較し、銀行口座データが財務データをどの程度再現できるか検証した。対象は 2018 年度から 2022 年度の財務諸表データがある約 6.3 万社である。

主要な財務指標等の再現比率（再現データ÷財務データ）の分布をみると、売上高、費用はゼロ付近でほとんど再現できない企業が多数を占めたが、1 付近に山があり、概ね正確に再現できる企業も一定数存在することが確認できた。一方、営業利益の再現比率はゼロ近傍が山になるのみで、ほとんど再現できない結果となった。人件費の再現比率はゼロ付近が大半だが、1.0 より少し小さい 0.7～0.8 あたりに山がある分布となった。財務データの人件費には派遣労働者や委託事業の人件費や社会保険料以外の福利厚生費が含まれるが、再現データは給与・賞与、社会保険料の支払いに限定されるため、再現比率が 1 より小さいことには妥当性がある。従業員数の再現比率もゼロ付近が大半だが、1.0 付近の割合が比較的目立つ分布となった。人件費、従業員数についても、概ね正確に再現できる企業が一定数存在することが確認できた。

（複数条件を用いた企業の絞り込み）

全サンプルを用いた検証を踏まえ、複数の絞り込み条件を設定して再現比率の高い企業を抽出することを試みた。具体的には売上高、人件費、従業員数の再現比率が 1（人件費は 0.7 ～0.8）に近い、各指標の再現データの増加率の財務データとの誤差が小さいことな

どを条件として設定した。更に、実証分析に使用できるデータセットを得ることを展望して、「普段から銀行口座を用いている」（他行の口座はあまり使用していない）企業を抽出するための「特定用途条件」（賃貸料やクレジットカード支払いに使用されていることなど）も設定した。これらの条件に基づきサンプルの抽出を行い（複数条件の組み合わせで約3千パターン、更に特定用途条件を適用して絞り込み)、再現データと財務データの決定係数が高く、サンプルサイズの大きい組み合わせを探索した。

抽出されたサンプルを確認すると、銀行口座データを用いて売上高、従業員数、人件費をいずれも高い水準で再現できるサンプルの企業数はおおよそ 1000 社以下に限られることが分かった。このうち、抽出に用いた年度が 2019 年の単年のみ、従業員数、人件費、売上の再現比率の上限・下限がそれぞれ（0.8, 1.2)、(0.6, 1.0)、(0.8, 1.2）かつ「1か月に最低一回以上クレジットカードの支払がある」「1か月に最低一回以上賃貸料の支払がある」、という抽出条件に合致する540社のサンプルを以降の分析に使用することとした。このサンプルは、全サンプルに比較して売上高や資本金の規模が相対的に大きい企業や、都道府県別には東京都所在の企業、業種としては製造業、卸売業が多いといった特徴がみられた。

（Isolation Forest を用いた企業の絞り込み）

再現データ集計値の時系列推移を法人企業統計等と比較する際、再現度の高いサンプルへの複数条件で絞り込んだ540社に加え、Isolation Forestを用いて異常値処理を行った再現データも比較の対象とした。Isolation Forestは、決定木を利用した機械学習手法の一つであり、データをサンプリングした上で、異常値を判定するための大量の決定木を作成して評価し、最終的な基準を設定する。具体的には再現売上高と再現営業利益について異常値の基準を設定したが、各四半期で異常値と判定されたサンプルは平均 1,200 件程度であり、複数条件を用いた絞り込みに比べてより多くのサンプルを対象とした分析が可能である。

（業績等に関するマクロ動向の確認）

銀行口座データから再現した企業業績等に関する指標について時系列データを作成し、主に法人企業統計と比較した。再現データは資本金規模の大きい企業が多いサンプルであったため、法人企業統計との比較の際は、全規模の母集団に合わせて拡大推計を行い、比較対象とする指標も全規模の数値を用いた。

まず、再現データの拡大推計値の水準を法人企業統計と比較すると、売上高は法人企業統計より過大、人件費では過小となり、営業利益は正負が一致しなかった。銀行口座データと法人企業統計のデータの性質や母集団の違いなどから、水準を再現することは困難と考えられる。

再現データの売上高、営業利益、人件費の動向を、前期比増減率、前年同期比増減率、時差相関係数により法人企業統計と比較すると、売上高、人件費については概ね類似した動きが確認されたが、営業利益については連動した動きはみられなかった。再現売上高に

ついて、タイミングはやや遅れているものの、新型コロナウイルス感染症の影響による2020年の落ち込みと回復の動きを捉えることができた。キャッシュフローを捉える再現売上高は、会計上の動きを捉える法人企業統計に対して若干のラグを持つと考えられる。なお、再現データは2018年のサンプルが少ないことから、2019年の前年同期比増減率は売上高、人件費でも乖離が大きかった。

前年同期比増減率について、サンプルの抽出パターンを変えることで法人企業統計との整合性が改善するか確認すると、

- 売上高については、再現度の高い540社のサンプル、Isolation Forestにより異常値を除いたサンプルで、法人企業統計との誤差が顕著に縮小した。
- 営業利益では、Isolation Forestにより異常値を除いた場合に誤差が改善したが、法人企業統計と類似した動きを再現することは困難であった。営業利益については、極端な値をとるサンプルの存在が、法人企業統計の動向の再現を困難にしている可能性が示唆される。
- 人件費では、2020年以降は全サンプルでも法人企業統計と近い動きを再現できており、サンプルの抽出方法を変えても結果は変わらなかった。

売上高の前年同期比増減率について業種別の比較を行うと、卸売業・小売業では他の業種に比べて法人企業統計との誤差が小さく、業種により整合性が異なることが確認された。

- 製造業では法人企業統計との誤差は全業種より大きいが、2020年の売上高の落ち込みやその後の回復といった動きはある程度再現できた。また、資本金区分を2区分に絞ったサンプル、Isolation Forestにより異常値を除いたサンプルで、誤差が顕著に改善した。
- 卸売業・小売業では全業種に比べて誤差が小さく、再現データで経済変動をより適切に把握できる可能性が示唆された。小売業で現金決済の比率が高いため、取引と入出金のタイミングが近いことが背景として考えられる。
- 不動産業・物品賃貸業では、前年同期比増減率の上昇、下落の動きは必ずしも一致しておらず、法人企業統計の動向を再現できているかは不明瞭であった。
- 宿泊業・飲食サービス業では、2020年の売上高の落ち込みやその後の回復の動きは比較的似通っている。宿泊・飲食サービス業でも、現金決済の比率が高いことが背景として考えられる。

再現データから作成した月次の指標を概念の近い他の月次統計と比較したが、再現データは月ごとの変動が大きく、一致した動きはあまり見られなかったが、以下のような点が確認できた。

- 製造業の再現売上高を鉱工業生産指数と比較すると、再現データは鉱工業生産指数よりやや遅れて変動していることが示唆される。
- 卸売業・小売業の再現売上高を商業動態統計と比較すると、2019年後半以降は概ね近い推移をしている。

・　再現人件費を総雇用者所得と比較すると、賞与等で大きく上下する動きを概ね再現できたが、総雇用者所得で見られる新型コロナウイルス感染症の影響による落ち込みや回復の動きは明瞭でない。

以上から、銀行口座データから再現した売上高、人件費について、企業業績等の動向を業種別に一定の精度で把握できることが分かった。標本抽出を工夫して信頼性の高いサンプルを集めることや、異常値を除くことで誤差を縮小し、精度を高められる可能性があることも確認できた。なお、銀行口座データはキャッシュフローを捉えるものであるため、会計上の動きを捉える法人企業統計等と変動のタイミングがやや異なる点に留意する必要がある。この点についえて、現金決済の比率が高いと考えられる卸売業・小売業や宿泊業・飲食サービス業では変動のタイミングが近いことも分かった。

標本抽出方法について、本稿で検討した再現度の高いサンプルへの絞り込み条件は、実証分析への使用を想定して条件を厳しくしたためサンプルサイズが小さくなり、業種別の分析には限界があったが、時系列変動の把握だけであれば条件を緩和して多くのサンプルサイズを確保する方法も考えられる。また、Isolation Forest などで異常値を除くだけでも法人企業統計との誤差が改善する場合があるので、異常値処理の工夫も引き続き検討する必要がある。

銀行口座データを用いて企業業績等のより有用な分析を行うため、企業統計で重視される営業利益の再現や、標本抽出方法の更なる改善方法を検討することが望まれる。

参考文献

井上祐介・川村健史・小寺信也（2019）「位置データを用いた滞在人口の分析　－働き方改革の進展－」経済財政分析ディスカッション・ペーパー・シリーズ、DP/19-3

宇南山卓（2011）「家計調査の課題と改善に向けて」統計と日本経済、1(1)、pp.3-28

大久保友博・高橋耕史・稲次春彦・高橋優豊（2022）「『オルタナティブデータ消費指標』の開発：オルタナティブデータを用いた個人消費のナウキャスティング」日本銀行ワー キングペーパーシリーズ、No. 22-J-9

小林周平・鈴木源一朗（2022）「経済動向分析における家計簿アプリデータの活用」経済財政分析ディスカッション・ペーパー・シリーズ、DP/22-3

小林周平・鈴木源一朗（2023a）「経済動向分析における家計簿アプリデータの更なる活用」経済財政分析ディスカッション・ペーパー・シリーズ、DP/23-2

小林周平・鈴木源一郎・吉中孝（2023b）「給付の所得制限周辺の世帯に限定したデータを用いた子育て世帯への臨時特別給付の消費増加効果の推計手法」経済財政分析ディスカッション・ペーパー・シリーズ、DP/23-4

財務省財務総合政策研究所（2020）「法人企業統計の一部早期化に係る検証（中間報告）」第４回国民経済計算体系的整備部会 QE タスクフォース会合提出資料　資料１―１

財務省財務総合政策研究所（2021）「法人企業統計の一部早期化に係る検証（中間報告２）」第 29 回国民経済計算体系的整備部会提出資料　資料３－１
財務省財務総合政策研究所（2022）「法人企業統計の一部早期化に係る検証（中間報告３）」第 31 回国民経済計算体系的整備部会提出資料　資料４－１
鈴木源一郎・森成弥（2023）「クレジットカードデータを用いた個人消費動向把握の精度向上の取組」経済財政分析ディスカッション・ペーパー・シリーズ、DP/23-2
中小企業庁（2022）「2022 年版　中小企業白書」
都竹直樹・岩上順子・栗山博雅（2024）「給与計算代行サービスデータの活用検討」経済財政分析ディスカッション・ペーパー・シリーズ、DP/24-2
内閣府（2022）「月例経済報告等に関する関係閣僚会議資料（令和５年４月 25 日）」（https://www5.cao.go.jp/keizai3/getsurei/2023/04kaigi.pdf#page=5）
内閣府（2023）「月例経済報告等に関する関係閣僚会議資料（令和６年５月 27 日）」（https://www5.cao.go.jp/keizai3/getsurei/2024/05kaigi.pdf#page=6９
内閣府（2023）「月例経済報告等に関する関係閣僚会議資料（令和６年６月 27 日）」（https://www5.cao.go.jp/keizai3/getsurei/2024/06kaigi.pdf#page=20）
内閣府政策統括官（経済財政分析担当）（2023）「特別定額給付金が家計消費に与えた影響－リアルタイムに記録される家計簿アプリデータを活用した分析－」政策課題分析シリーズ 22
三菱 UFJ リサーチ&コンサルティング（2023）「産業界における手形・小切手の利用実態等に関する調査　最終報告書」https://www.zenginkyo.or.jp/fileadmin/res/abstract/council/tegata_denshi/tegata_denshi2021_12_3.pdf
Baker, S, R., Farrokhnia, R, A., Meyer, S., Pagel, M., Yannelis, C. (2020) “Income, Liquidity, and the Consumption Responce to the 2020 Economic Stimulus Payments”. *National Bureau of Economic Research Working Paper Series*, No. 27097, DOI: 10.3386/w27097
Carvalho, V, M., Garcia, J, R., Hansen, S., Ortiz, A., Rodrigo, T., Rodriguez Mora, J, V., Ruiz, P. (2021) “Tracking the COVID-19 crisis with high-resolution transaction data”. *R. Soc. Open Sci*, 8:210218, DOI: 10.1098/rsos.210218
Cox, N., Ganong, P., Noel, P., Vavra, J., Wong, A., Farrell, D., Greig, F., Deadman, E. (2020). “Initial Impacts of the Pandemic on Consumer Behavior: Evidence from Linked Income, Spending, and Savings Data”. *Brookings Papers on Economic Activity,* 35–69, DOI: 10.1353/eca.2020.0006
Karger, E., Rajan, A. “Heterogeneity in the Marginal Propensity to Consume: Evidence from Covid-19 Stimulus Payments”. WP-2020-15, Federal Reserve Bank of Chicago, DOI: 10.21033/wp-2020-15

Kawaguchi, K., Kodama, N., Kumanomido, H., Tanaka, M. (2023) “Using Manager’s Expectations for Ex-Ante Policy Evaluation: Evidence from the Covid-19 Crisis”. *Journal of Economics & Management Strategy,* 32, 714-732, DOI:10.1111/jems.12515

Kubota, S., Onishi, K., Toyama, Y. (2021) “Consumption responses to COVID-19 payments: Evidence from a natural experiment and bank account data”. *Journal of Economic Behavior & Organization*, 188:1-17, DOI: 10.1016/j.jebo.2021.05.006.

Liu, F, T., Ting, K, M., Zhou, Z. (2008) “Isolation Forest” *2008 Eighth IEEE International Conference on Data Mining*, pp.413-422, DOI: 10.1109/ICDM.2008.17

Liu, F. T., Ting, K, M., Zhou, Z. (2012) “Isolation-Based Anomaly Detection”. ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1), pp.1-39, DOI: 10.1145/133360.2133363

Raschka, S., Mirjalili, V., (2017) *Python Machine Learning.*, Second Edition, Packt Publishing.