

財務ビッグデータの可視化と統計モデリング

地道正行 (関西学院大学 商学部)

概要

本研究では、財務ビッグデータを用いて、企業活動のグローバル化がもたらす負の側面（企業の租税回避、労働者と株主間の付加価値配分、企業の富の偏在等）の実態に関する証拠と課題を提示した。この財務ビッグデータは、Bureau van Dijk (BvD) の 154 カ国・9 万社超の上場企業の（最長）30 年間・91 系列の財務（諸表）データ（Osiris, 300 万行, 1.6GB）、および非上場企業を含む 2,600 万社超の 10 年間・85 系列の世界最大規模の財務（諸表）データ（Orbis, 2.9 億行, 142GB 超）である。これを GPGPU 環境で Apache Spark, PG-Storm と R を連動させて、探索的データ解析（Exploratory Data Analysis）にもとづき、時空間の観点からダイナミックかつインタラクティブなデータ可視化（Data Visualization）を行った。得られた知見に基づき、企業行動を高精度に予測する統計モデリングと実証分析を行うことでその有効性を検証した。これらの結果を社会に広く還元し、企業行動を持続可能な発展に向けて変革することを目指すものである。

1 共同研究に関する情報

1.1 共同研究を実施した拠点名

東京大学 情報基盤センター

1.2 共同研究分野

■超大容量ネットワーク技術分野

1.3 参加研究者の役割分担

地道正行 (関西学院大学 商学部):

- データ前処理, データラングリング
- 探索的データ解析にもとづく統計モデリング

宮本大輔 (東京大学大学院 情報理工学系研究科):

- データ解析環境の構築
- データラングリング

阪 智香 (関西学院大学 商学部):

- 財務データの経済・会計学的考察

永田修一 (関西学院大学 商学部):

- パネルデータ・時系列データ解析の理論構築

2 研究の目的と意義

本研究の目的は、次の 3 つである。

(1) データの前処理とラングリングの高速化:

本研究では、以下の表で与えられるような世界の全

上場企業・非上場企業の財務ビッグデータ*1を GNU parallel 等の並列処理環境を利用し、さらに Apache Spark, PG-Storm 環境とデータ解析環境 R を連動して利用することによって、可視化を行うためのデータ前処理の速度 (velocity) を改善する研究を行うとともに、欠測情報などのデータの品質に関する考察も行う:

データセット名	年月版	データベース	上場・非上場	連結・非連結
DS-Osiris-C-2020	2020/3	Osiris	上場	連結
DS-Osiris-U-2020	2020/3	Osiris	上場	非連結
DS-Osiris-C-2021	2021/3	Osiris	上場	連結
DS-Osiris-U-2021	2021/3	Osiris	上場	非連結
DS-Orbis-C-2019	2019/12	Orbis	上場・非上場	連結
DS-Orbis-U-2019	2019/12	Orbis	上場・非上場	非連結

なお、規模は以下のようなものである:

データセット名	企業数	規模
DS-Osiris-C-2020	96,377 社	約 300 万行, 1.6GB 超
DS-Osiris-U-2020	96,377 社	約 300 万行, 1.6GB 超
DS-Osiris-C-2021	100,542 社	約 300 万行, 1.7GB 超
DS-Osiris-U-2021	100,542 社	約 300 万行, 1.7GB 超
DS-Orbis-C-2019	2,400 万社超	約 2.9 億行, 142GB 超
DS-Orbis-U-2019	2,400 万社超	約 2.9 億行, 142GB 超

さらに、昨年度から開発している学内向けデータ抽出システム SKWAD に日経 NEEDS 財務データや上記の Osiris データ, Orbis データを実装するとともに、非財務情報として FTSE Russell 社の ESG レーティングデータや新国富 (Inclusive Wealth: IW) データも本格的に利用し、その前処理のみならず、財務データと結合することも実施する。

*1 Bureau van Dijk (BvD) のデータベースから抽出

(2) データ可視化:

研究目的 (1) で処理された世界の企業財務ビッグデータと ESG レーティングデータ (DS-FTSE-2020), IW データ (DS-IW-2018) を用い, 時空間の観点からビジュアライゼーション技法を用いたデータ可視化 (data visualization) を行う. データ自身の情報を探索的に引き出し, グローバルな企業活動の実態 (収益性, 付加価値分配, 人的資本, 生産性, 租税回避等) に関する新しい知見と課題を明らかにする.

(3) 統計モデリング:

研究目的 (2) の可視化の情報をもとに, 時間・空間の両面から探索的データ解析 (Exploratory Data Analysis: EDA) を実行することによって, 企業行動を高精度に予測する統計モデリングと実証分析を行うことでその有効性を検証する. 具体的には, 財務データと ESG レーティングデータを用いて, 非対称分布族を考慮した両対数モデルによる株式時価総額の統計モデリングを行う. また, 近年その重要性が指摘されている, 再現可能性を研究全体に対して確保することも, 本研究の目的である (図 1 参照).

本研究では, 財務諸表を対象とする既存研究の中で過去最大規模のデータを扱う. そのためには, 高速な計算機環境・ネットワーク環境といった物理的な資源と, その環境を利用するための技術・知識・経験, さらにそれらを会計学・統計科学・情報科学の専門的観点から総合的に評価・分析できる多面的な知見をもつ人的資源が必要となる. これらの専門知識を有する研究メンバーにより, 東京大学情報基盤センターの強力な JHPCN 環境 (FENNEL) のもと, 財務ビッグデータを処理・可視化・解析する世界で唯一の研究を進めることができることが本研究の意義である.

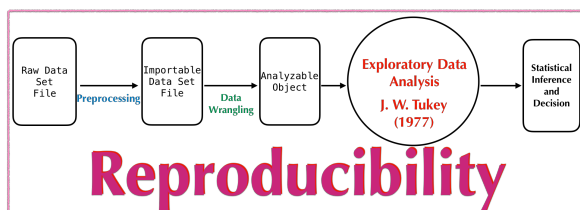


図 1 探索的財務ビッグデータ解析の全工程

3 当拠点公募型研究として実施した意義

本研究で扱う財務データセット (複数のテキストファイル) の内, 最大の規模は合計 140 GB 超であり, 通常の計

算機環境では直接扱うことは難しい. また, これらのファイルは, 抽出したものをそのまま解析することは困難であり, データ形式の変換, 文字・行末コードの変換, 欠測値の置換などの前処理が必要である. これらの工程を高速化するためには, 相応の処理能力を有する計算機環境が必要となる. また, データを分析するためにデータを拠点へ転送する際や, 拠点内でネットワークを介した PG-Strom 等の環境下でラングリングするにあたって, GPGPU 環境と高速なネットワーク資源が必要となる. さらに, 本研究で使用する分析法の一つに, コンピュータ・シミュレーションをベースとした評価法があり, この評価法による研究を進める上で, GPGPU を利用した高速計算を実行可能なコンピュータ環境が必要となる. このような研究を行うにあたり, JHPCN 環境のもと, 高速な計算機環境・ネットワーク環境を利用することによって, グローバルな企業行動について, これまでの研究に引き続き世界初となる証拠と知見を得ることできることが拠点として実施した意義である.

4 前年度までに得られた研究成果の概要

2017~2020 年度に採択された JHPCN 環境 (FENNEL) を利用した研究は, 主に世界の全上場企業の財務データベース Osiris から 2016~2019 年に抽出されたデータセットにもとづくものである. 前年度 (2020 年度) に得られた研究成果は, 主に 2019 年に抽出されたデータにもとづいている.

2020 年度に得られた研究成果の概要を以下に述べる. なお, ここで与えられる成果の総合報告を, JHPCN 第 13 回シンポジウム (国内会議 (16)) にて行った.

研究目的 (1) に対する成果

- 学術論文 (3) の結果を, 国内会議 (1), (2) で報告するとともに, この論文で議論された前処理に対する並列化の方法を, 本年度抽出した Orbis のデータセット (DS-Orbis-C-2019, DS-Orbis-U-2019) に対して適用し, 同様の結果を得ることができた. この結果は, 国内会議 (5) で報告された.
- 前処理した Orbis のデータセット (DS-Orbis-C-2018, DS-Orbis-U-2018) を, 東京大学の FENNEL 環境のもとで, Apache Spark, PG-Strom 環境とデータ解析環境 R を連動して利用することによって, データラングリングを行う速度を改善するための方策について学術論文 (5) で詳細

に議論し、処理に利用したソースコードも与えた。成果としては、従来方法による処理速度よりも 20 倍程度の向上が得られた。なお、この結果は、国内会議 (1), (2) でも報告した。さらに、新たに前処理した Orbis のデータセット (DS-Orbis-C-2019, DS-Orbis-U-2019) についても今年度に用意された PG-Strom 環境でラングリングを行った結果、再現することを確認するとともに、シリアライズの方法として従来形式と Apache Parquet 形式を利用した場合の詳細な比較を行い、従来の方法が 1 時間 10 分かかったのに対して、後者の方法は、10 分程度の時間短縮ができた。これらの結果については、国内会議 (5) で報告した。

- (c) 2015 年のパリ協定や国連の持続可能な開発目標 (Sustainable Development Goals : SDGs) などを受けて、企業の E (環境)・S (社会)・G (ガバナンス) 活動が注目されていることから、企業の ESG 活動に関する開示情報として優れている FTSE Russell ESG レーティングデータ (以下 “DS-FTSE-2020” と略) と会計情報を結合することを検討した。このことを実現するために、まず、2020 年度に抽出した財務データを GNU Parallel を用いて前処理を行いデータセット DS-Osiris-C-2020 を生成した。次に、FTSE Russell 社の ESG レーティングデータのファイルから R を利用して自動抽出することによって、データセット DS-FTSE-2020 を生成した。さらに、異なったデータベースから抽出された企業に関するデータを結合 (join) するために、主キーとして国際証券コード仕様 ISO6166 で定められている全世界共通の証券系コード (ISIN code) を利用し、これらのデータセット DS-Osiris-C-2020 と DS-FTSE-2020 を結合し、新たなデータセット DS-Osiris-C-FTSE-2020 を得た。これら一連の処理工程を make によって自動実行し、再現性を確保した (図 2 参照)。この結果は、学術論文 (7) で公表された。

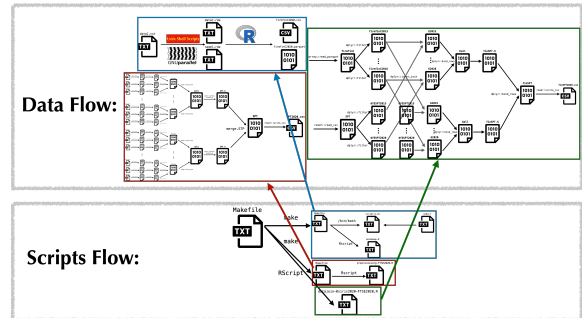


図 2 Osiris データと ESG レーティングデータの結合

- (d) データ抽出環境を整備するための試みとして、システム (SKWAD) を構築した。このシステムに搭載するデータとして、データセット DS-Osiris-C-2020, DS-Osiris-U-2020, DS-Orbis-C-2019, DS-Orbis-U-2019 に加えて、国内の企業を対象とした日経 NEEDS 企業財務データ (2018 年版, 2019 年版, 2020 年版) のデータセット (以後, “DS-NEEDS-2018”, “DS-NEEDS-2019”, “DS-NEEDS-2020” と略) も追加した。このことから、データセット DS-NEEDS-2018, DS-NEEDS-2019, DS-NEEDS-2020 の前処理とデータベース化を行った。これらの処理とシステム構築の詳細を、学術論文 (6) で述べるとともに、国内会議 (6), (9) で発表した。

研究目的 (2) に対する成果

- (a) 研究目的 (1) に対する成果 (c) であるデータセット DS-Osiris-FTSE-2020 を用いて、過去 6 年間における ESG 情報開示・実績と企業価値との関係を検討した。企業の ESG 情報開示・実績について、国別・業種別に可視化し、さらに、ESG 情報開示・実績と企業価値の関係について、対散布図 (図 13) やモーショントラjectory (図 4) を用いた可視化によって明らかにした。この結果は、国内会議 (3), (4), (12) で発表するとともに、学術論文 (1), (2), (4), (8), (9) で公表した。
- (b) 2019 年度に実施した非上場企業の租税回避の蓋然性を確認するための ETR と ROA の散布図を、研究目的 (1) に対する成果 (b) の結果として得られたデータを用いて 2009 年から 2018 年の 10 年分にわたって可視化し、結果が再現することを確認した。この結果は、国内会議 (5) で報告した (図 6 参照)。

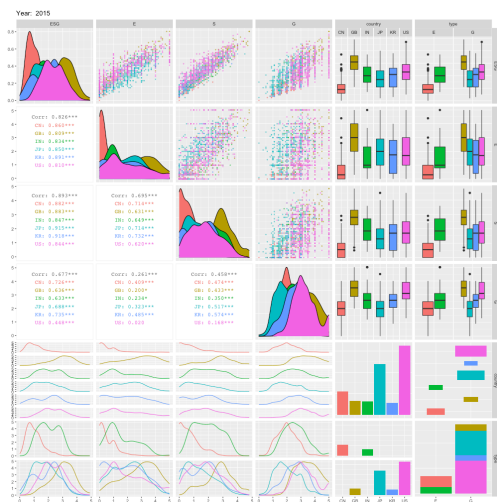


図3 対散布図: ESG, E, S, G, 株式時価総額の推移



図4 モーションチャート: 対数株式時価総額 (y 軸) と ESG スコア (x 軸) の 49 カ国, 2015 ~ 2020 年 (6 年間) の推移

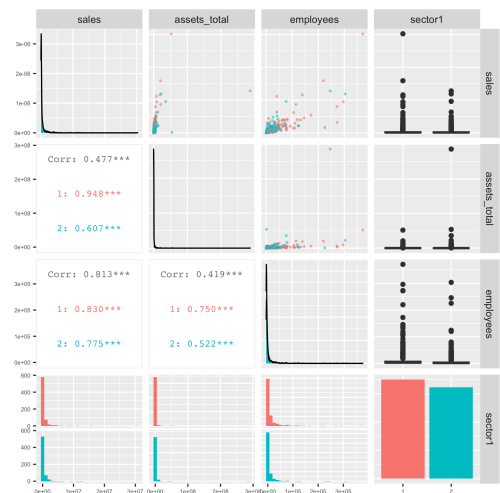


図5 2020年3月決算の東京証券取引所第一部上場企業の売上高, 資産合計, 従業員数の対散布図

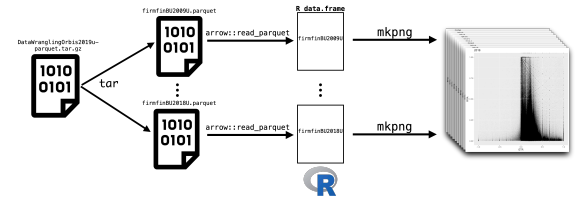


図6 2009年~2018年の ETR (x 軸) と ROA (y 軸) の散布図の描画プロセス

(c) システム SKWAD (研究目的 (1) に対する成果 (d)) から 2019 年 3 月期決算の東京証券取引所第一部上場企業に対する売上高 (sales), 従業員数 (employees), 資産合計 (assets_total), 業種分類 (sector1) などのデータを連結決算ベースで抽出し, 可視化を行った. この結果として得られた知見の主なものとして, 業種によらず売上高 (sales), 資産合計 (assets_total), 従業員数 (employees) の全てが極端に右に歪んだものであることがわかる. また, これらのペアの散布図も原点付近に密集しており, 2次元の意味で歪んだものであることがわかる. さらに, 業種 (1: 製造業, 2: 非製造業) によって, 相関の強弱に違いがあることがわかる. 例えば, 売上高 (sales) と資産合計 (assets_total) の相関は, 製造業 (0.948) と非製造業 (0.607) で極端に異なっていることもわかった (図 5 参照). この結果は, 学術論文 (6) で公表するとともに, 国内会議 (11) でも発表した.

研究目的 (3) に対する成果

- (a) データセット DS-Osiris-FTSE-2020 の可視化 (研究目的 (2) に対する成果 (a)) にもとづいて, 株式時価総額を ESG レーティングで予測するためのモデリングとして, 誤差が非対称分布族に従う片対数モデルを考え, その当てはまりを可視化と赤池情報量基準などで比較・検討した. この結果は, 国内会議 (3), (4) で発表した.
- (b) 現在発展がめざましい, 機械学習の分野での深層学習のモデリングとして TensorFlow とその API である Keras を R から利用してパターン認識を実行することを試みた. この結果は国内会議 (7) で報告された.
- (c) 多重共線性と外れ値が存在する場合に, 配当金支払金額を予測するための線形モデルの回帰係数の推定として, 幾つかの頑健性をもつ縮小推定量を

- 考察し、理論的・実証的な側面から考察を行った。この結果を(査読付き)学術論文(3)で発表した。
- (d) 2019年度の研究目的(3)に対する結果がデータセット DS-Osiris-C-2020 で再現するかを検証した。その結果として、売上高を Box-Cox 変換したものに正規分布を当てはめるよりも、対数変換後に非対称ティータ分布を当てはめることによってモデリングしたものが、当てはまりが良いことが再度確認された。この成果は、国内会議(10)で報告された。
- (e) 研究目的(2)に対する成果(c)として得られた可視化の結果と(査読付き)学術論文(1)で得られた統計モデルを利用することによって、売上高を従業員数、資産合計、業種分類で説明するためのモデリングを行った。その結果として、対数非対称ティータ分布を誤差構造にもつ両対数モデルが検討したモデルの中で赤池情報量規準の意味で最も当てはまりが良いことがわかった(図7参照)。この結果は、国内会議(11)で発表した。

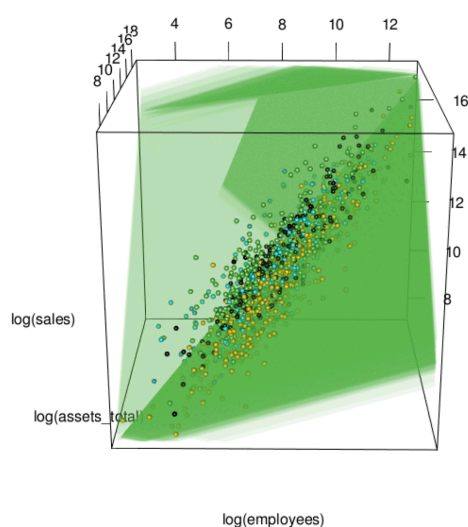


図7 3次元散布図と対数非対称ティータ分布を誤差構造にもつ両対数モデルの回帰平面群

5 今年度の研究成果の詳細

研究目的(1)に対する成果

- (a) 2020年度に得られた財務データセット DS-Osiris-C-2020 と FTSE Russell ESG レーティングデータ (DS-FTSE-2020) の結合に関する結果(学術論文(7))を、国内会議(14), (17), (20),

(22) で発表するとともに、この成果を利用して、今年度新たに抽出したデータセット DS-Osiris-C-2021 と DS-FTSE-2021 を結合し、新たなデータセット DS-Osiris-C-FTSE-2021 も得ることができた。

- (b) 持続可能性とリンクした包括的な指標で、実際に政策決定などにも使われ始めている国連の新国富の指標は、新たに国の富(ストック)をはかる基準であり、このデータは CSV ファイルで提供されている。ここでは、新国富のデータセット DS-IW-2018 と、2018年度に抽出した財務データセット DS-Osiris-C-2018 で提供される指標を年度と国毎に集計したものを結合することによって、年度・国毎の財務データと新国富データの結合を行った(図8参照)。

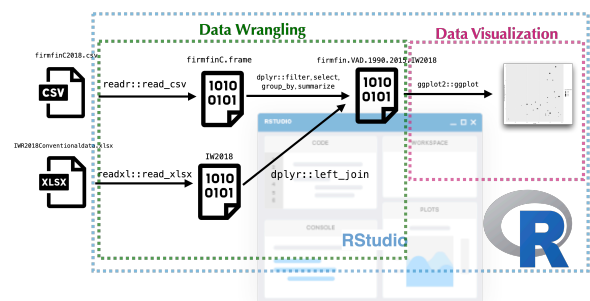


図8 Osiris データと IW データの結合

これらの工程を自動処理するための R のスクリプトを作成した。その際、国名の表記揺れを R のパッケージ (countrycode) を利用することによって ISO コードで統一的に処理する工夫を行った。なお、この結果は、学術論文(4)で公表された。

- (c) データ抽出システム SKWAD で提供する財務データとして、2020年度に抽出された日経 NEEDS 企業財務のデータセット (DS-NEEDS-2020) と、Osiris データセット DS-Osiris-C-2020, DS-Osiris-C-2020, Orbis データセット DS-Orbis-C-2020, DS-Orbis-C-2020 を前処理し、システムに実装した。これらの工程は全て make コマンドによって自動実行する仕様として再現性を確保した。なお、これらの一連の成果は、学術論文(6), (12), (14)で公表するとともに、国内会議(15), (16), (17), (19), (24)で発表した。

- (d) Apache Arrow や vroom パッケージを利用したデータラングリングを、各種の規模の財務データを利用して、読み書きのベンチマークを実施することによって、規模やコンピュータ環境に応じたラングリング法の選択について検討を行った。その結果として、Parquet 形式と vroom パッケージを利用した場合が良いことがわかった。この実験には、次年度から利用する予定の mdx の実験環境*2も利用し、mdx がこれまでの研究を継続するために非常に優れた環境であることも確認できた。以上の結果は国内会議 (25) で発表された。
- (e) データセット DS-Orbis-C-2019, DS-Orbis-U-2019 を利用して JHPCN 環境で構築されたデータベースから財務データを PG-Strom を利用して抽出し、データラングリングを実行することによって世界の非上場企業の財務データを可視化することができるオブジェクト形式へ変換することができた。この結果は、学術論文 (18)) において公表された。

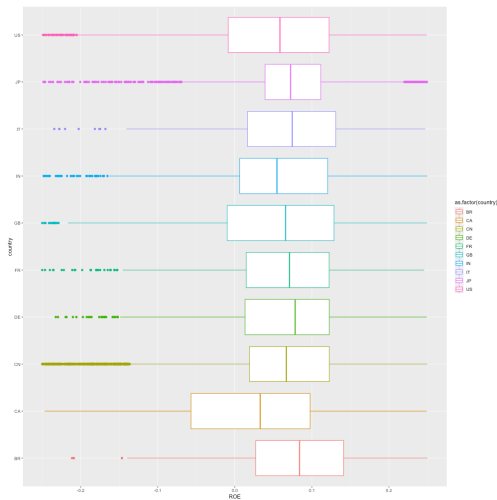


図 9 ROE ($-25\% < ROE < 25\%$) ・ 2018 年度 ・ GDP 上位 10 カ国

研究目的 (2) に対する成果

- (a) データセット DS-Osiris-C-2020 を利用し、世界の全上場企業の 29 年間の財務データを用いた探索的データ解析を実施し、売上や株式時価総額の総額と伸び、収益性分析 (ROE), 安全性分析 (自己資本比率), 生産性分析 (一人当たり売上高), 付加価値分配, 租税回避の蓋然性, 配当性向の実態

を示した。具体的には、各国や各業種の実態を明らかにするとともに、日本企業は収益性や安全性の集中度が高く、アメリカ企業等と比較すると数値が高い企業は少ないものの低い企業も少なく、10 カ国の中位である (図 9)。また、日本企業は租税回避の蓋然性が低く、配当性向については中央値が高く配当ゼロ企業の割合も低いことから (図 10), 政府や投資家への還元に取り組んでいるといえる。一方で、企業の富に関するジニ係数が、他国と同様に上昇している実態も明らかとなった (図 11)。なお、この結果は、学術論文 (10), (11) で公表するとともに、国内会議 (13), (18), (27) において発表した。

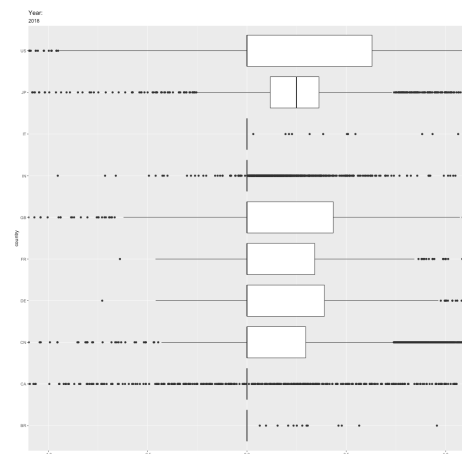


図 10 配当性向 (配当金/当期純利益, $-100\% < \text{配当性向} < 100\%$) ・ 2018 年度 ・ GDP 上位 10 カ国

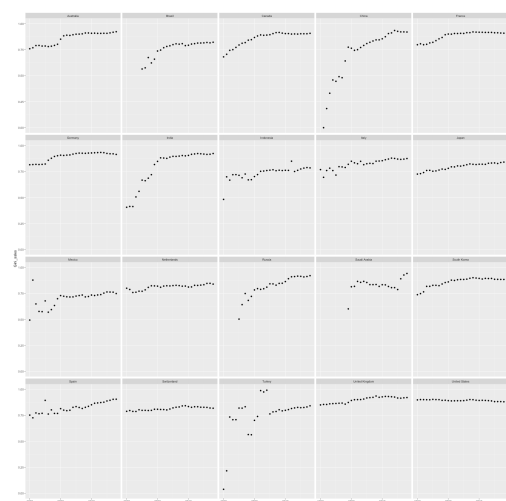


図 11 GDP 上位 20 カ国の全上場企業・29 年間 (1990 ~ 2018 年度) の純資産ジニ係数の推移

- (b) 研究目的 (1) を達成することによって新たに得

*2 プロジェクト名: 「JHPCN-関学地道-東大宮本-実験」

られたデータセット DS-Osiris-C-FTSE-2020 を利用し、以下のような可視化を行った。

- 2018 年度の株式時価総額のヒストグラム (図 12) を描くことによって、分布構造を詳細に検証した。

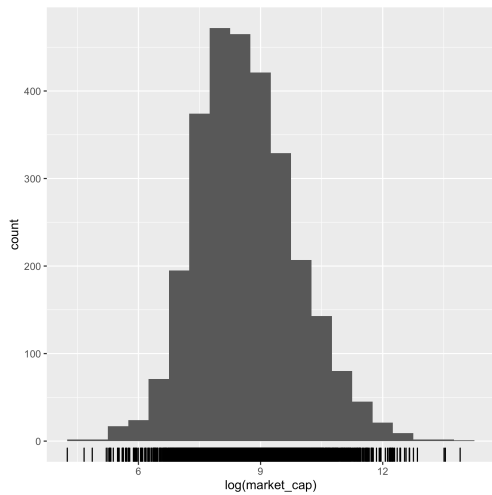


図 12 株式時価総額のヒストグラム: 対数スケール

- 2018 年度の株式時価総額、純資産、当期純利益 (財務データ) とサマリー指標、環境指標、社会指標、ガバナンス指標 (ESG レーティングデータ) の対散布図 (図 13 参照) を描くことによって、各指標の構造 (線形関係、分布等) を検証した。

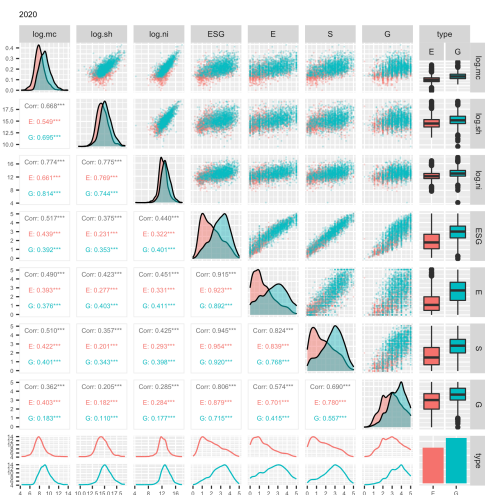


図 13 財務データと ESG レーティングデータの対散布図: 対数スケール

- (c) 新国富は、自然資本 (Natural Capital: 持続的に利用や管理が必要となる資源や自然など)、人的資本 (Human Capital: 教育・健康などの人の豊か

さ)、人工資本 (Produced Capital: 設備・機械などの経済的・物的な豊かさ) の 3 つから成る。これは、現世代と将来の世代が得であろう富 (ストック) を金銭価値で評価し、豊かさのポテンシャル (持続可能性) を計測したものである。研究目的 (1) を達成することによって新たに得られたオブジェクト firmfin.VAD.1990.2015.IW2018 (図 8) を利用し、企業の従業員人件費や利益と人的資本との関連を分析した結果、当期純利益に対して従業員給付が多い国は、新国富に占める人的資本の割合が高いなど、企業の労働分配率が国富 (ストック) にも影響を与えている可能性が示唆された (図 14)。この結果は、(査読付き) 学術論文 (4) で公表された。

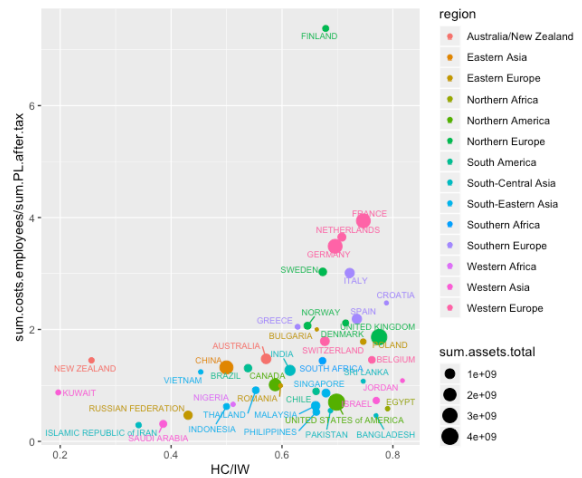


図 14 新国富に占める人的資本の割合 (x 軸) と当期純利益に対する従業員給付の割合 (y 軸) のバブルチャート

- (d) 研究目的 (1) を達成することによって新たに得られたデータ抽出システム SKWAD から、日経 NEEDS 財務データとして、2020 年 3 月期末決算の上場企業 1,441 社の配当金支払額、純資産等の財務データを抽出し、それらの散布図・ヒストグラム等を作成することによって可視化をおこなった。

研究目的 (3) に対する成果

- (a) 研究目的 (2) に対する研究結果であるデータセット DS-Osiris-C-FTSE-2020 に関する 2018 年度の株式時価総額のヒストグラム (図 12) から、対数株式時価総額が非対称正規分布や非対称テール分布等の非対称分布族に属する分布でモデリングできるという結果を得ることができた。

- (b) 上の結果と、研究目的 (2) に対する研究結果である株式時価総額、純資産、当期純利益 (財務データ) とサマリー指標、環境指標、社会指標、ガバナンス指標 (ESG レーティングデータ) の対散布図 (図 13 参照) を援用することによって、非対称分布族に属する誤差分布をもつ株式時価総額の両対数モデリングを行い、財務データと ESG レーティングデータを用いて株式時価総額を予測するモデリングを行った。以上の結果については、学術論文 (13) で公表するとともに、国内会議 (20)、(22)、国際会議 (1) で発表した。
- (c) データセット DS-Osiris-C-2020 と DS-Osiris-C-FTSE-2020 を結合したものを利用して、探索的データ解析を実施し、純資産と利益を所与としても ESG スコアが高いほど企業価値が高いこと、企業価値に対する財務情報の説明力は ESG 情報より大きいものの、ESG 情報の説明力も一定程度確認できること、ESG 情報の説明力が高まりつつあること、さらに、気候変動情報が高いほど企業価値が高いことなどを回帰分析によって確認した。この結果は、学術論文 (16)、(19) で公表するとともに、国内会議 (23) で発表した。
- (d) 研究目的 (2) に対する研究結果である 2020 年 3 月期末決算の上場企業 1,441 社の配当金支払額、純資産等の財務データの散布図・ヒストグラムから得られた知見にもとづいて統計モデリングを行った。結果として、非対称ティー誤差を持つ両対数モデル (応答変数: 配当金支払額 (対数), 説明変数: 純資産 (対数), ROE(粗データ)) が情報量規準の比較等によって最良であることが確認された。この結果は、国内発表 (21)、(26) で発表した。
- (e) 2020 年度の研究目的 (3) に対する結果 (e) を、データを最新のもので再検証し、学術論文 (17) で公表した。

なお、以上の研究において行われる処理は、全て自動的に実行され、また論文も動的文書生成によって作成されており、この意味で再現可能研究の立場から行われている。

また、今年度の新たな試みとして、統計数理研究所 (統計思考院) における人材育成事業であるオンラインワー

クショップ^{*3} において、JHPCN 拠点として採択された 2017 年度から 2020 年度までの期間の研究成果を体系化することによって、総合的に紹介することができた (ワークショップ (1), (2), (3), (4))。

6 今年度の進捗状況と今後の展望

本研究の申請時に設定していた計画では、非上場企業を含めて抽出したデータセットの前処理とラングリングを最新技術 (GNU Parallel, PG-Strom, Apache Arrow) を利用して行い、これまで扱うことが難しかった非上場企業を含めた富の蓄積と偏在、租税回避行動、付加価値分配に関する分析、可視化、モデリングを実行することであったが、今年度の研究成果 (学術論文 12 編, 学会報告 16 回, 著書 1 冊, ワークショップの開催, プレス発表等) に照らしても概ね順調に結果を得ることができているものと思われる。

これまでの進捗を勘案し、各研究目的に対する今後の展望を与える:

(1) 研究目的 (1) に関しては以下の事項を検討する:

- データラングリングについては、ほぼノウハウが確立しているが、環境は日進月歩で進歩しているため、次年度利用する予定の mdx 環境に適した技術動向の研究を引き続き行う。
- これまでは、データにおける欠測値は考慮しない方法で解析してきたが、データにおける欠測値の補完に代表されるデータクリーニングに関して著書 (訳書) (1) で扱われている手法を援用することも、今後の課題である。なお、この研究の発展としてデータを匿名化することによって、真の分布構造をもつ疑似財務データを生成し、学内に構築されたデータ抽出システム (SKWAD) 上へ実装したものを JHPCN 環境へポータリングすることによって、オープンデータとして公開することも展望として考えられる。

(2) 研究目的 (2) に関しては以下の事項を検討する:

- 2017 年度から 2021 年度まで JHPCN プロジェクトとして行ってきた財務データの可視化は、主に世界の企業 (上場・連結決算) に関するデータ (DS-Osiris-C-20**) を利用したものであったが、今年度に得られた結果 (学術論文 (18)) は、非

^{*3} 『探索的ビッグデータ解析と再現可能研究』, <https://sites.google.com/view/ws-ebda-rr-2021/>

上場・単独決算に関する企業の財務データ (DS-Orbis-U-2019) を利用することが可能となったという意味で一つの集大成と呼べるものである。次年度は、この結果を利用して非上場企業の財務データの可視化を行うことを予定しており、この試みから未知の財務データの構造が明らかになることが期待される。

(3) 研究目的 (3) に関しては以下の事項を検討する:

- パネルデータや企業サイズ分布, 確率フロンティア分析に基づくモデリングを検討する。
- 一般化双曲型分布族を利用した財務データのモデリングについて検討する。

なお, 上記の研究では再現可能性を確保することも目的とする。

また, 今年度の新たな試みとして実施したワークショップ (統計数理研究所・統計思考院・人材育成事業) を次年度も引き続き実施する予定である。

7 研究業績一覧 (発表予定も含む)

学術論文 (査読あり)

- (1) Masayuki Jimichi, Daisuke Miyamoto, Chika Saka, Shuichi Nagata, ‘Visualization and statistical modeling of financial big data: Double-log modeling with skew-symmetric error distributions’, *Japanese Journal of Statistics and Data Science*, Vol. 1, Issue 2, pp. 347–371, December 2018. <https://doi.org/10.1007/s42081-018-0019-1>
- (2) 地道正行, 阪 智香, 『探索的財務ビッグデータ解析 – データ可視化による企業活動の実態解明と統計モデリング –』, 日本経営数学会誌, 第 39 巻, 第 1・2 号合併号, pp. 1–23, 2020 年 11 月。
- (3) Jia-Han Shih, Ting-Yu Lin, Masayuki Jimichi, Takeshi Emura, ‘Robust ridge M-estimators with pretest and Stein-rule shrinkage for an intercept term’, *Japanese Journal of Statistics and Data Science*, Available online 19 October 2020, Vol. 4, Issue 1, pp. 107–150, DOI:10.1007/s42081-020-00089-6.
- (4) Shunsuke Managi, Masayuki Jimichi, Chika Saka, Human capital development: Lessons from global corporate data, *Economic Analysis and Policy*,

Available online 6 September 2021, Vol. 72, pp. 268–275, <https://doi.org/10.1016/j.eap.2021.08.013>

国際会議プロシーディングス

(該当なし)

国際会議発表

- (1) Masayuki Jimichi, Chika Saka, Daisuke Miyamoto, Shuichi Nagata, *Statistical Modeling of Market Capitalization with Financial Data and ESG Rating Data*, The 11th Conference of The Asian Regional Section of the International Association for Statistical Computing, Doshisya University, Feb 24, 2022. (査読あり)

国内会議発表

- (1) 地道正行*, 宮本大輔, 阪 智香, 永田修一, 『探索的財務ビッグデータ解析 –PG-Strom によるデータラングリングの並列化–』, 日本計算機統計学会 第 34 回大会, オンライン開催, 2020 年 5 月 31 日 (日)。
- (2) 地道正行*, 宮本大輔, 阪 智香*, 永田修一, 『財務ビッグデータの可視化と統計モデリング』, 学際大規模情報基盤共同利用・共同研究拠点 (JHPCN) 第 12 回 シンポジウム, オンライン開催 (ポスター発表), 2020 年 7 月 9 日 (木)。
- (3) 阪 智香*, 『ESG 情報の開示と企業価値』, 日本経済会計学会, 第 37 回全国大会, 統一論題「非財務情報の開示と企業価値」報告・討論, 早稲田大学, オンライン開催, 2020 年 7 月 25 日 (土)。(招待講演)
- (4) 阪 智香*, 『企業の ESG 情報と企業価値』, 2020 年度国際会計研究会大阪会計研究会第 1 回合同研究会 (オンライン開催), 2020 年 8 月 8 日 (土)。
- (5) 地道正行*, 宮本大輔, 阪 智香, 永田修一, 『探索的財務ビッグデータ解析 –PG-Strom によるデータラングリングの並列化–』, 国際数理科学協会, 2020 年度年会「統計的推測と統計ファイナンス」分科会研究集会, 大阪大学, オンライン開催, 2020 年 8 月 22 日 (土)。
- (6) 地道正行*, 阪 智香, 『財務データ抽出システム KGUSBADES の再構築』, 国際数理科学協会, 2020 年度年会「統計的推測と統計ファイナンス」分科会研究集会, 大阪大学, オンライン開催, 2020 年 8 月 22 日 (土)。
- (7) 劉 澤霖*, 黒澤拓能, 地道正行, 『R と TensorFlow

- による機械学習』, 国際数理科学協会, 2020 年度年会「統計的推測と統計ファイナンス」分科会研究集会, 大阪大学, オンライン開催, 2020 年 8 月 22 日 (土).
- (8) 阪 智香 *, 『開かれた社会関連会計研究とは?』, 日本社会関連会計学会 2020 年度全国大会, 統一論題「開かれた社会関連会計研究を目指して: 国際的、学際的、実践的な貢献の検討」報告・討論, 立命館大学, オンライン開催, 2020 年 11 月 1 日 (日). (招待講演)
- (9) 地道正行 *, 阪 智香, 『学内向け財務データ抽出システムの再構築』, 日本計算機統計学会 第 34 回シンポジウム, オンライン開催, 2020 年 11 月 29 日 (日).
- (10) 地道正行 *, 宮本大輔, 阪 智香, 永田修一, 『変換による財務データの統計解析—売上高の場合—』, 2020 年度日本経営数学会研究大会, オンライン開催 (専修大学神田キャンパス), 2020 年 12 月 19 日 (土). (査読あり)
- (11) 地道正行 *, 宮本大輔, 阪 智香, 永田修一, 『R による探索的財務データ解析』, 統計数理研究所共同研究集会 2020 年度「データ解析環境 R の整備と利用」, 統計数理研究所, オンライン開催, 2020 年 12 月 19 日 (土).
- (12) 阪 智香 *, 『ESG と財務会計: ESG と企業価値、今後に向けて』, 日本会計研究学会, 第 70 回関西部会, 統一論題「ESG と会計」報告・討論, 関西学院大学, オンライン開催, 2020 年 12 月 19 日 (土). (招待講演)
- (13) 阪 智香 *, 『SDGs の観点から企業行動の実態を明らかにする—探索的財務データ解析—』 日本学術会議 SDGs と経営実践・経営学・経営学教育を検討する分科会, 2021 年 4 月 29 日 (木).
- (14) 地道正行 *, 阪 智香 *, 『財務データと ESG レーティングデータの前処理と結合』, 日本計算機統計学会 第 35 回大会, オンライン開催, 2021 年 6 月 3 日 (木).
- (15) 地道正行 *, 阪 智香, 『財務データ抽出システム SKWAD』, 日本経営数学会第 43 回 (通算 63 回) 研究大会, オンライン開催 (専修大学神田キャンパス), 2021 年 6 月 12 日 (土). (査読あり)
- (16) 地道正行 *, 宮本大輔, 阪 智香, 永田修一, 『財務ビッグデータの可視化と統計モデリング』, 学際大規模情報基盤共同利用・共同研究拠点 (JHPCN) 第 13 回シンポジウム, オンライン開催, 口頭発表, 2021 年 7 月 8 日 (木).
- (17) 地道正行, 宮本大輔, 阪 智香, 永田修一『財務ビッグデータの可視化と統計モデリング』, 学際大規模情報基盤共同利用・共同研究拠点 (JHPCN) 第 13 回シンポジウム, オンライン開催, ポスター発表, 7 月 9 日 (金).
- (18) 阪 智香 *, 『サステナブル投資の視点: 財務・非財務情報の探索的データ解析』, 日本学術会議 環境学委員会 環境政策・環境計画分科会 サステナブル投資小委員会, 2021 年 8 月 6 日 (金).
- (19) 地道正行 *, 阪 智香『財務データ抽出システム SKWAD』, 国際数理科学協会, 2021 年度年会「統計的推測と統計ファイナンス」分科会研究集会, 関西学院大学, オンライン開催, 2021 年 8 月 22 日 (土).
- (20) 地道正行 *, 宮本大輔, 阪 智香, 永田修一『財務データと ESG レーティングデータによる株式時価総額の統計モデリング』, 国際数理科学協会, 2021 年度年会「統計的推測と統計ファイナンス」分科会研究集会, 関西学院大学, オンライン開催, 2021 年 8 月 22 日 (日).
- (21) 田中 凜 *, 阪 智香, 地道正行, 『日本の上場企業における企業価値に関する統計モデリング』, 国際数理科学協会, 2021 年度年会「統計的推測と統計ファイナンス」分科会研究集会, 関西学院大学, オンライン開催, 2021 年 8 月 22 日 (日).
- (22) 地道正行 *, 宮本大輔, 阪 智香 *, 永田修一, 『財務データと ESG レーティングデータによる株式時価総額の統計モデリング』, 2021 年度統計関連学会連合大会, オンライン開催, 2021 年 9 月 7 日 (火). (査読あり)
- (23) 阪 智香 *, 『グローバルな財務・ESG データ分析からみえる課題』, 国際会計研究学会第 38 回研究大会, 統一論題報告, オンライン開催, 2021 年 10 月 9 日 (土). (招待講演)
- (24) 地道正行 *, 阪 智香, 『財務データ抽出システム SKWAD: Osiris, Orbis データの利用』, 日本計算機統計学会 第 35 回シンポジウム, オンライン開催, 2021 年 11 月 27 日 (土).
- (25) 地道正行, 宮本大輔, 阪 智香, 永田修一『R による財務データの前処理再考』, 統計数理研究所共同研究集会 2021 年度「データ解析環境 R の整備と利用」, 統計数理研究所, オンライン開催, 2021 年 12 月 18 日 (土).
- (26) 田中 凜 *, 阪 智香, 地道正行『企業価値の統計モデリング』, 第 16 回日本統計学会春季集会慶應義塾大学三田キャンパス, ハイブリッド開催, 2022 年 3 月 4 日 (土).

- (27) 阪智香*, 『世界の上場企業・非上場企業の会計ビッグデータ解析』, 日本会計研究学会, 第71回関西部会, 統一論題「AI・DXの進展と会計研究・教育」報告, 京都先端科学大学, オンライン開催, 2022年3月4日(金). (招待講演)

公開したライブラリなど

(該当なし)

その他

学術論文

- (1) 阪智香, 國部克彦, 地道正行, 『探索的データ解析に基づく世界企業の付加価値分配』, 神戸大学ディスカッションペーパー, 2019-28, pp. 1-35, 2020年1月.
- (2) 大鹿智基, 阪智香, 地道正行, 『「社会にとってよい企業」への市場の評価とサステナビリティ』, 企業会計, 第72巻, 第1号, pp. 74-80, 中央経済社, 2020年1月.
- (3) 地道正行, 『探索的財務ビッグデータ解析 –前処理の並列化–』, 商学論究, 第67巻, 第3号, pp. 1-19, 関西学院大学商学研究会, 2020年3月. <http://hdl.handle.net/10236/00028408>
- (4) 阪智香, 國部克彦, 地道正行, 『会計と不平等 –付加価値分配率の探索的データ解析–』, 国民経済雑誌, 第221巻, 第4号, pp. 1-20, 2020年4月. http://www.lib.kobe-u.ac.jp/handle_kernel/E0042013
- (5) 地道正行, 『探索的財務ビッグデータ解析 –PG-Stromによるデータラングリングの並列化–』, 商学論究, 第67巻, 第1号, pp. 1-34, 関西学院大学商学研究会, 2020年9月. <http://hdl.handle.net/10236/00029031>
- (6) 地道正行, 『財務データ抽出システムの再構築 –NEEDS 企業財務データを中心に–』, 商学論究, 第68巻, 第3号, pp. 1-78, 関西学院大学商学研究会, 2021年3月. <http://hdl.handle.net/10236/00029257>
- (7) 地道正行, 阪智香, 『財務データと ESG レーティングデータの前処理と結合』, 商学論究, 第68巻, 第3号, pp. 79-116, 関西学院大学商学研究会, 2021年3月. <http://hdl.handle.net/10236/00029258>
- (8) 阪智香, 『ESG 情報と企業』, 商学論究, 第68巻, 第4号, pp. 149-170, 関西学院大学商学研究会, 2021年3月. <http://hdl.handle.net/10236/00029270>
- (9) 阪智香, 『ESG と企業価値 –将来に向けての視点–』, 会計, 第199巻, 第4号, pp. 12-24, 2021年4月. (日本会計研究学会, 統一論題報告, 招待論文)
- (10) 阪智香, 『企業の格差と分配 –探索的財務データ解析–』, ディスクロージャー & IR, 第17巻, pp. 154-63, 宝印刷, 2021年5月.
- (11) 阪智香, 『アカデミック・フォーサイト –世界の上場企業の財務データ解析–』, 会計・監査ジャーナル, 第792巻, pp. 76-84, 中央経済社, 2021年7月.
- (12) 地道正行, 『財務データ抽出システムの再構築 –Osiris データの利用–』, 商学論究, 第69巻, 第1号, pp. 71-109, 関西学院大学商学研究会, 2021年7月. <http://hdl.handle.net/10236/00029776>
- (13) 地道正行, 阪智香, 『財務データと ESG レーティングデータによる株式時価総額の統計モデリング』, 商学論究, 第69巻, 第2号, pp. 1-64, 関西学院大学商学研究会, 2021年12月. <http://hdl.handle.net/10236/00029974>
- (14) 地道正行, 『財務データ抽出システムの再構築 –Orbis データの利用–』, 商学論究, 第69巻, 第2号, pp. 65-109, 関西学院大学商学研究会, 2021年12月. <http://hdl.handle.net/10236/00029975>
- (15) 阪智香, 『開かれた社会関連会計研究とは?』, 社会関連会計研究, 第33巻, pp. 105-116, 2021年12月. (日本社会関連会計学会, 統一論題報告, 招待論文)
- (16) 阪智香, 『気候変動と企業価値』, 青山アカウンティングレビュー, 第11号, pp. 61-64, 2022年2月.
- (17) 地道正行, 『R による探索的財務データ解析と再現可能研究 –NEEDS 企業財務データの利用–』, 商学論究, 第69巻, 第3・4号, pp. 1-82, 関西学院大学商学研究会, 2022年3月. <http://hdl.handle.net/10236/00030066>
- (18) 地道正行, 阪智香, 『探索的財務ビッグデータ解析と再現可能研究 –非上場企業のデータラングリング–』, 商学論究, 第69巻, 第3・4号, pp. 83-120, 関西学院大学商学研究会, 2022年3月. <http://hdl.handle.net/10236/00030067>
- (19) 阪智香, 『グローバルな財務・ESG データ分析からみえる課題』, 国際会計研究学会年報 2021 年度 (近刊). (日本会計研究学会, 統一論題報告, 招待論文)

ワークショップ

- (1) 地道正行, 『探索的財務データ解析と再現可能研究』, 統計数理研究所統計思考院, 2021 年度人材育成事業, オンラインワークショップ『探索的ビッグデータ解析と再現可能研究』, 2021 年 8 月 29 日 (日).
- (2) 地道正行, 『探索的財務ビッグデータ解析と再現可能研究』, 統計数理研究所統計思考院, 2021 年度人材育成事業, オンラインワークショップ『探索的ビッグデータ解析と再現可能研究』, 2021 年 8 月 29 日 (日).
- (3) 阪 智香, 『財務ビッグデータの会計学の視点からの可視化』, 統計数理研究所統計思考院, 2021 年度人材育成事業, オンラインワークショップ『探索的ビッグデータ解析と再現可能研究』, 2021 年 8 月 29 日 (日).
- (4) 宮本大輔, 『財務ビッグデータの前処理・データラングリング高速化』, 統計数理研究所統計思考院, 2021 年度人材育成事業, オンラインワークショップ『探索的ビッグデータ解析と再現可能研究』, 2021 年 8 月 29 日 (日).

訳書

- (1) Loo, Mark van der, Jonge, Edwin de 著, 地道正行, 高橋雅夫, 藤野友和, 安川武彦, 和田かず美 共訳, 『統計的データクリーニングの理論と実践 -R によるデータ編集/欠測補完システム-』 共立出版, 2022 年 2 月, ISBN: 9784320114630

マニュアル・プレスリリース等

1. 地道正行, 『SKWAD ユーザマニュアル -NEEDS 企業財務データの抽出-』, pp. 1-88, 関西学院大学リポジトリ, 2021 年 4 月 27 日, <http://hdl.handle.net/10236/00029654>
2. 産経新聞, 2021 年 9 月 22 日夕刊の記事, 【経済 24 時】「ESG コスト増も成長後押し」において, 本共同研究内容が紹介された.
3. 阪 智香, 『探索的財務データ解析にみる税負担削減行動』, 近畿税理士界, 第 693 号, p. 7, 2022 年 1 月.
4. 阪 智香, 『企業活動と SDGs』, 日経 SDGs フェス大阪関西「2025 年大阪・関西万博に向けて」, グランフロント大阪, 2022 年 2 月 17 日 (木).
5. 阪 智香, 『我が国のサステナビリティ開示のあり方を考える』, 第 5 回国際会計人材ネットワークシンポジウム, パネリスト報告, オンライン開催, 2022 年 3 月 23 日 (水).