

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

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、世界 154 カ国の全上場企業・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-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-Orbis-C-2019	2,600 万社超	約 2.9 億行, 142GB 超
DS-Orbis-U-2019	2,600 万社超	約 2.9 億行, 142GB 超

(2) データ可視化:

研究目的 (1) で処理された世界 154 カ国の企業財務ビッグデータを用い、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 (付加価値配分と人的資本, 生産性, 租税回避) 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。なお、企業サイズ分布 (Firm Size Distribution: FSD) や確率フロンティア

*1 Bureau van Dijk (BvD) のデータベースから抽出

分析 (Stochastic Frontier Analysis: SFA) に基づく企業分布生成理論モデルに関して検証を行うとともに、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).

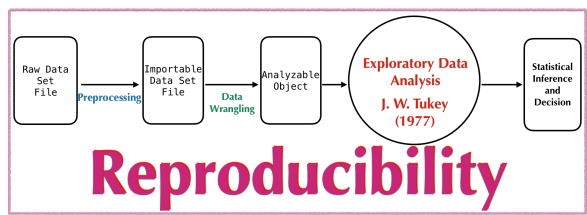


図 1 探索的財務ビッグデータ解析の全工程

本研究が扱う財務データは、既存研究のデータ量を遙かに凌駕する過去最大規模のものであり、これを扱うためには、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。

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が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

2017～2019 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 2016～2018 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 (2019 年度) に得られた研究成果は、主に 2018 年に抽出された以下のデータにもとづいている。

データセット名	抽出年度	データベース	上場・非上場	連結・非連結
DS-Osiris-C-2018	2018	Osiris	上場	連結
DS-Osiris-U-2018	2018	Osiris	上場	非連結
DS-Orbis-C-2018	2018	Orbis	上場・非上場	連結
DS-Orbis-U-2018	2018	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-2018	9 万社超	約 300 万行, 1.4GB 超
DS-Osiris-U-2018	9 万社超	約 300 万行, 1.4GB 超
DS-Orbis-C-2018	2,400 万社超	約 2.4 億行, 124GB 超
DS-Orbis-U-2018	2,400 万社超	約 2.4 億行, 124GB 超

2 節で述べた研究目的それぞれについて、2019 年度に得られた研究成果の概要を述べる。なお、ここで与えられる成果の総合報告を、JHPCN 第 12 回シンポジウム (国内会議 (14)) にて行った。

研究目的 (1) に対する成果

- (a) データセット DS-Osiris-C-2018, DS-Osiris-U-2018, DS-Orbis-C-2018, DS-Orbis-U-2018 の前処理に対して GNU parallel を利用して並列化することによって、処理速度の向上を検証した結果、従来の方法と比べて 5 倍程度の処理速度の向上が得られた (図 2参照)。



図 2 財務ビッグデータ (データセット DS-Orbis-U-2018) の並列化前処理の全工程

この結果を、処理に利用したソースコードとともに学術論文 (8) で公表すると共に、国内会議 (5), (7), (8), (9), (11), (12) で報告した。

- (b) (a) で前処理されたデータファイルを、東京大学情報基盤センター内に実験的に設置された環境でデータベース化し PG-Strom を利用してデータラングリングしたところ、それまで利用していた Spark 環境よりも高速化できることが検証できた (図 3)。なお、この結果については、国内会議 (12) にて報告を行った。

研究目的 (2) に対する成果

- (a) データセット DS-Osiris-C-2018 における 2015

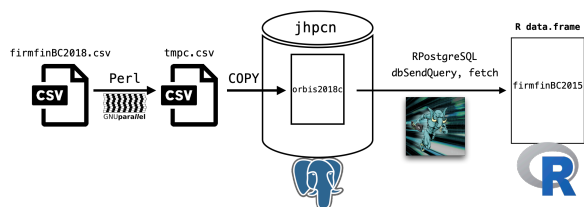


図3 Orbis データの PG-Strom によるデータラングリング

年の売上高データについて、Box-Cox 変換後に正規分布を当てはめた結果を、動的でインタラクティブ性をもつデータ可視化によって検証するための Web アプリケーションを RStudio Shiny と R を用いて構築する方法について、論文 (学術論文 (3)) で報告した (図 4)。

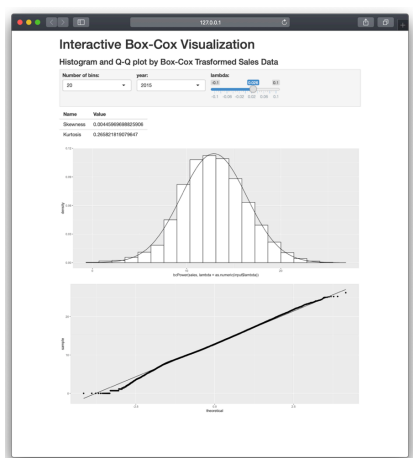


図4 RStudio Shiny による Box-Cox 変換のインタラクティブなデータ可視化

- (b) データセット DS-Osiris-C-2018, DS-Osiris-U-2018 を用いて世界の全上場企業の付加価値分配を可視化することによって、従業員賃金が削られて、投資家利益が増加していること、また、企業の租税回避が横行している実態を明らかにした。地域別分析や国別分析からも同様の傾向を確認した。この結果については学術論文 (4), (11) で公表すると共に、国際会議 (2), 国内会議 (3), (6) で報告した。
- (c) データセット DS-Osiris-C-2018, DS-Orbis-U-2018 を用いて非上場企業の租税回避の蓋然性の確認した。具体的には、付加価値分配のうち、政府への分配 (税金支払)、特に企業の租税回避に焦点を当てた。上場企業については 2018 年度に実施したため、非上場企業を含む世界 154 カ国の全企業について、実効税率 (=支払税金/税引前利益, ETR) を x 軸、総資本利益率 (ROA) を y 軸にプロットした散布図 (10 年分) を作成した。税

率ゼロ (x 軸真ん中) に企業が集中し、全く税金を払わない企業が利益率の相当高い企業にもあり (y 軸上限は ROA 100%), 租税回避の蓋然性が確認できる (図 ?? 参照)。この結果については学術論文 (2), (4) で公表するとともに、国内会議 (10) で報告した。

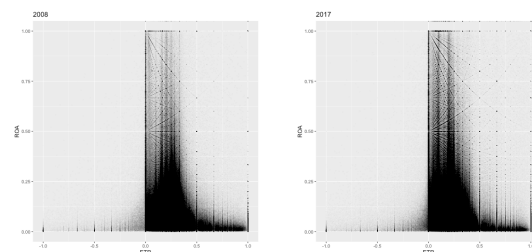


図5 ETR (x 軸) と ROA (y 軸) の散布図: 2008 年, 2017 年

- (d) データセット DS-Osiris-C-2018, DS-Orbis-U-2018 を用いて企業資本 (総資産) と環境資本・人的資本・人工資本との関係等を調査した。具体的には、モーションチャートによって、(フローである GDP の補足として) ストックを測る国連の新国富 (Inclusive Wealth) の指標を用い、140 カ国の国毎の企業のストック (総資産) 合計との関係 (過去 20 年間の推移) を示した (図 6 参照)。人工資本や人的資本 (図 6 の中段, 下段) は、企業総資産の増加とともに増えているが、自然資本 (図 6 上段) は減少している。この図では G7 諸国に軌跡をつけているが全体も同じ傾向である。このことから経済活動によって不可避免的に生じる環境問題への対応は、最重要課題のひとつであると言える。一方、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る (図 7)。また、49 カ国の企業の ESG 活動を評価した FTSE Russell ESG Rating (2018 年度データ) と Osiris 財務データを用いた分析では、ESG 総合スコアと、E (環境)・S (社会)・G (ガバナンス) それぞれのスコアのいずれを用いた分析でも、スコアの高い企業の企業価値 (株式時価総額) が高いことが観察された。これは企業の ESG 活動に取り組む企業への good news でもある。SDGs 達成に向けて、経済社会の持続可能性を確保するためには、企業活動の実態をグローバルレベルで解明し、その課題を解決していく必要がある。世界を正しく知るための証拠の提示は、ステークホルダーの

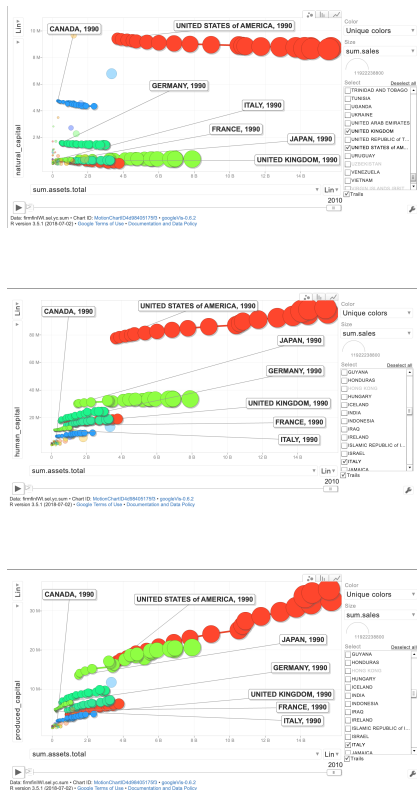


図6 140カ国企業資本を x 軸, 環境資本 (上段)・人的資本 (中段)・人工資本 (下段) を y 軸にとったモーションチャート: G7 の 1990 年から 2010 年までの推移の軌跡

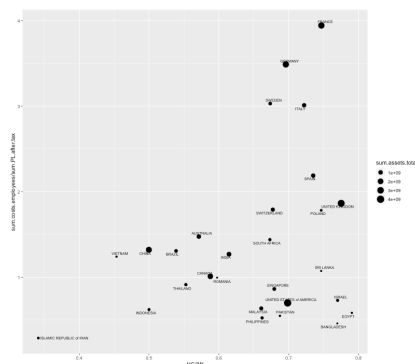


図7 新国富に占める人的資本の割合 (x 軸) と当期純利益に対する従業員給付の割合 (y 軸) のバブルチャート: 資産合計の総和を円の大きさにマッピング

民主的な参加やガバナンスに不可欠である。これらの結果については学術論文 (5) で公表した。

研究目的 (3) に対する成果

- (a) データセット DS-Osiris-C-2018 における 2015 年の売上高データについて、これまで検討してきた対数変換後のデータに非対称分布族を当てはめた結果に対して、さらに Box-Cox 変換後に正規分布を当てはめた結果との比較を行い、これまで検討してきた結果が優位性を保つことがわかった。この結果を、学術論文 (3) として公表し、Sweave を利用し、 $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ と R を協調して作成す

ることによって、再現性を確保した。

- (b) 学術論文 (1) 探索的データ解析を実行することによって得られた統計モデルを、新たなデータセット DS-Osiris-C-2018 に当てはめ結果が再現すること、及び、全ての工程が再現することを確認した (図 8 も参照)。この結果については、学術論文 (11) で報告すると共に、学会報告を行った (国際会議 (1), 国内会議 (2), (7), (8))。

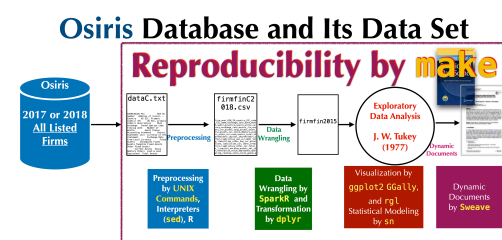


図8 再現可能研究: Osiris データの前処理・ラングリング, 探索的データ解析, 動的文書生成を `make` コマンドによって自動実行

5 今年度の研究成果の詳細

2 節で述べた研究目的それぞれについて、今年度に得られた研究成果の詳細を述べる。

研究目的 (1) に対する成果

- (a) 学術論文 (8) の結果を、国内会議 (13), (14) で報告すると共に、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 (17) で報告された。
- (b) 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 (10) 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 (13), (14) 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、シリアル化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 (17) で報告した (図 9 参照)。

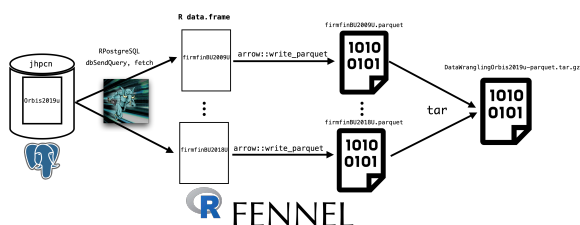


図 9 PG-Strom 環境におけるラングリングと Apache Parquet によるシリアル化

- (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-Russell-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、まず本年度抽出したデータセット DS-Osiris-C-2020 (財務データ) を GNU Parallel を用いて前処理したものと、データセット DS-FTSE-Russell-2020 を R を利用して前処理した後、R を利用してこれらのデータセットを主キーとして ISIN コード (国際証券コード) を用いて結合することによって新たなデータセット (以後、“DS-Osiris-FTSE-2020”と略) を得ることができた。この結果は、学術論文 (14) で公表された。
- (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 (13) で述べるとともに、国内会議 (18), (21) 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 10) やモーションチャート (図 11) 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 (15), (16), (24) で発表するとともに、学術論文 (6), (7), (9), (15), (16) で公表した。

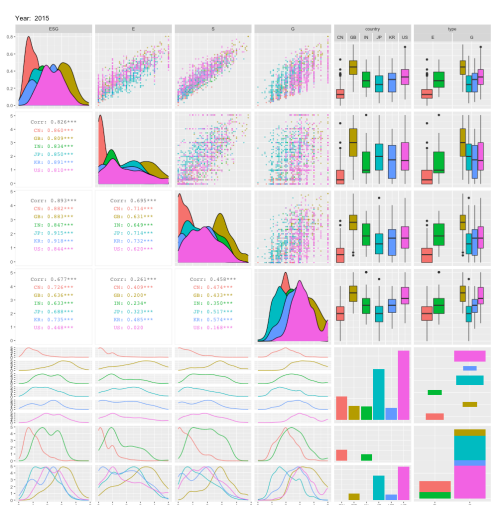


図 10 対数散布図: ESG, E, S, G, 株式時価総額の推移: アメリカ, 日本, 韓国, イギリス, 中国, インド: FTSE Russell ESG Rating の全対象企業について、対数株式時価総額, 株式時価総額, ESG, E, S, G の 2015 ~ 2020 年 (6 年間) の推移を対数散布図で可視化したもの。先進国の国別企業数上位 4 カ国であるアメリカ (桃), 日本 (青緑), 韓国 (青), イギリス (茶), 新興国の国別企業数上位 2 カ国である中国 (赤), インド (緑) の計 6 カ国を取り上げた (これらは対象企業 130 社以上の国)。右のボックスプロットと下の確率密度関数は、国別と先進国・新興国のタイプ別の 2 種類が示されている。この 6 カ国では、イギリス (茶) のスコアが最も高いこと、インド (緑) も比較的高く、中国は企業数が最も多いこともあり、スコアは低い が 6 年間で上昇してきていることなどがわかる。

- (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 (図 5) 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。これらの可視化が可能になったのは、研究目的 (1) に対する成果 (b) によるも



図 11 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: FTSE Russell ESG Rating の全対象企業について, 対数株式時価総額を y 軸, ESG スコアを x 軸として, 49 カ国の 2,401~3,945 社, 2015~2020 年 (6 年間) の推移をモーションチャートで可視化したものがこの図である. 色は, 先進国が黄緑, 新興国が青である. それぞれの円が企業を示しており, 円の面積は各年度の当該企業の売上高を示す. このモーションチャートから, 先進国と新興国の企業の ESG スコアの分布, 推移, 株式時価総額との関係などを確認することができる.

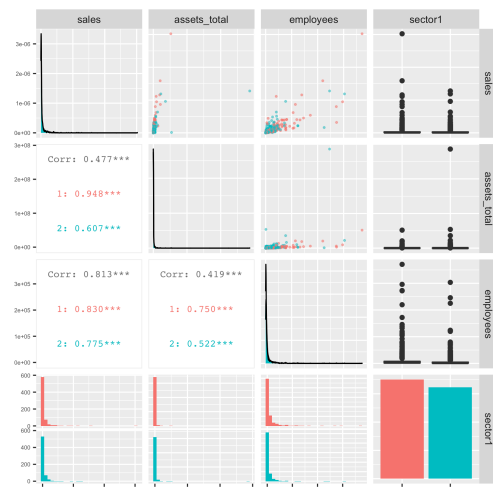


図 12 対角ブロックには売上高 (sales), 資産合計 (assets_total), 従業員数 (employees) に関する推定された密度関数と, 日経産業分類 (sector1 大分類) の頻度が描かれている. また, 上三角ブロックには, 売上高 (sales), 資産合計 (assets_total), 従業員数 (employees) の 2 つの組合せに関する散布図と最終列には, それぞれのデータのボックスプロット (1: 製造業, 2: 非製造業) が描かれている. さらに, 下三角ブロックには, 売上高 (sales), 資産合計 (assets_total), 従業員数 (employees) の 2 つの組合せに関する相関係数が与えられており, 最終行には, それぞれのデータのヒストグラム (1: 製造業, 2: 非製造業) が描かれている.

のである. この結果は, 国内会議 (17) で報告した (図 13 参照).

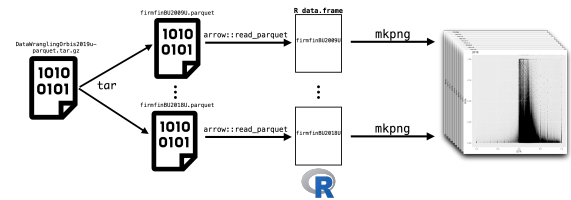


図 13 2009 年~2018 年の ETR (x 軸) と ROA (y 軸) の散布図の描画プロセス

- (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 12 参照). この結果は, 学術論文 (13) で公表するとともに, 国内会議 (23) 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 (15), (16) 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 (19) 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。
この結果を学術論文 (12) で発表した。

- (d) 2019 年度の研究目的 (3) に対する成果 (a) によって得られた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 (22) 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 14 参照)。この結果は、国内会議 (23) で発表した。

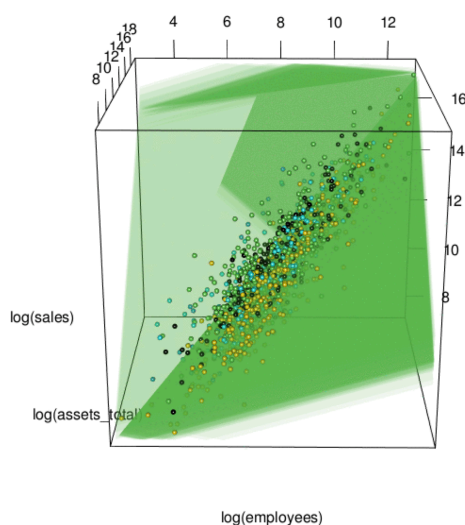


図 14 3次元散布図と対数非対称ティー分布を誤差構造にもつ両対数モデルの回帰平面群: 日経業種分類 (中分類) によるダミー変数によって切片を変化させることによって回帰平面群を生成

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、本年度に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 Parquet) を利用して行い、企業の富の蓄積と偏在、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 8 編, 学会報告 12 回) 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、以下が今後の展望として考えられる:

研究目的 (1) に関する展望

- (a) データ前処理のさらなる高速化: 2020 年度は、データの前処理を従来よりも高速に再現性を確保して実行できる成果を得ることができたが、国内会議 (14) の質疑応答の際に頂いたコメントを検討した結果、さらなる高速化・効率化の余地があることがわかったため、このことを検討する。
- (b) データラングリングのさらなる高速化: これまでの成果として、データラングリングを GPGPU 環境下で PG-Strom を利用することによって、実行可能な時間で行うことができるという結果を得ているが、さらに短時間でラングリングを行うための方策として、Apache Arrow 形式を利用することを検討する。分散処理環境である Apache Spark も最新バージョンになってから GPGPU 環境にも本格的に対応したことから、最新動向のキャッチアップを行う。
- (c) 複数のデータの結合: 2020 年度、新たに非財務データである FTSE Russell ESG レーティングデータの前処理と財務データの結合を行ったが、この処理を新しいツール (R の purrr パッケージなど) を利用して効率化することを検討する。
- (d) データ抽出環境の整備: 今年度は主に日経 NEEDS 企業財務データを抽出するシステム (SKWAD) を実験的に学内に構築したが、Osiris 及び Orbis データのデータベース化を行うことによって、より汎用性をもつように改良する。さらには、JHPCN 環境へのポーティングを検討する。

研究目的 (2) に対する展望

- (a) 財務データ・非財務データ・結合データの可視化: これまでの研究成果として、様々な財務データと非財務データ、結合データが整備されてきており、その利用も従来より容易になってきている、今後もこれらのデータを可視化することによってモデリングにつなげる予定である。
- (b) 可視化環境の整備: 可視化の環境 (ツール・手法等) も日進月歩のため、最新の動向のキャッチアップ行うとともに整備する。

研究目的 (3) に関する展望

- (a) 株式時価総額のモデリング: 2020 年度に得られ

た財務データと非財務データを結合したデータセット DS-Osiris-FTSE-2020 を利用した株式時価総額のモデリングに関する結果を、さらに詳細に検討し、再現可能研究の観点から執筆し、投稿する。

- (b) 変換による売上高データの統計モデリング: 研究目的 (3) に対する成果 (d) を再現可能研究の観点から執筆し、投稿する。
- (c) 東京証券取引所第一部上場企業に対する財務データの統計モデリング: 2020 年度に得られた 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) のデータを用いた統計モデリングの結果を再現可能研究の観点から執筆し、投稿する。
- (d) その他のモデリング: 企業サイズ分布や確率フロンティア分析に基づくモデリングを検討する。

7 研究業績一覧 (発表予定も含む)

学術論文

- (1) M. Jimichi, Miyamoto, D., Saka, C, and Nagata, S., ‘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> (査読あり)
- (2) 大鹿智基, 阪 智香, 地道正行, 『企業の租税回避行動をめぐる証拠の可視化 –グローバルデータの探索的解析–』, 産業経理, 第 79 巻, 第 2 号, pp. 118–128, 産業経理協会, 2019 年 7 月。
- (3) 地道 正行, 『変換による財務データの統計解析 –売上高の場合–』, 商学論究, 第 67 巻, 第 1 号, pp. 27 – 46, 関西学院大学商学研究会, 2019 年 10 月. <http://hdl.handle.net/10236/00028298>
- (4) C. Saka, T. Oshika, and M. Jimichi, ‘Visualization of tax avoidance and tax rate convergence: Exploratory analysis of world-scale accounting data’, *Meditari Accountancy Research*, Vol. 27, No. 5, pp. 695–724, October 2019, Emerald Publishing Limited. (査読あり)
- (5) I. A. Jumbri, S. Ikeda, M. Jimichi, C. Saka, S. Managi, ‘Inequality of health stock and the relation to national wealth’, *International Journal for Equity in Health*, Vol. 18, No. 1, pp.188–, November 2019. (査読あり)
- (6) 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。
- (7) 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。
- (8) 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>
- (9) 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
- (10) 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>
- (11) 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. (査読あり)
- (12) 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*, 10.1007/s42081-020-0. (査読あり)
- (13) 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>
- (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』, 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>
- (15) 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>
- (16) 阪 智香, 『ESG と企業価値 –将来に向けての視点』,

国際会議発表

- (1) M. Jimichi*, D. Miyamoto, C. Saka, and S. Nagata, *Exploratory Financial Big Data Analysis and Reproducible Research*, DSSV 2019, Doshisha University, Imadegawa Campus, August 14th, 2019.
- (2) C. Saka* and M. Jimichi, *Visualization of Corporate Tax Avoidance and Value Added Distribution: Exploratory Analysis of Financial Big Data*, DSSV 2019, Doshisha University, Imadegawa Campus, August 14th, 2019.

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 –前処理, データラングリング, 再現可能性–』, 日本計算機統計学会シンポジウム予稿集, 滋賀大学データサイエンス学部, 2018 年 11 月 11 日 (日).
- (2) 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と再現可能研究』, 日本経営数学会第 41 回 (通算 61 回) 研究大会, 拓殖大学茗荷谷キャンパス, 2019 年 6 月 1 日 (土). (査読あり)
- (3) 阪 智香*, *Visualization of tax avoidance and tax rate convergence: Exploratory analysis of world-scale accounting data*, 日本会計研究学会特別委員会「税制が企業会計その他の企業行動に及ぼす影響に関する研究」研究会, 慶應義塾大学日吉キャンパス, 2019 年 7 月 6 日 (土).
- (4) 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 11 回 シンポジウム, THE GRAND HALL(品川), 2019 年 7 月 11 日 (木).
- (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 –前処理の並列化–』, 国際数理科学協会, 2019 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, 関西学院大学大阪梅田キャンパス, 2019 年 8 月 24 日 (土).
- (6) 阪 智香*, 國部克彦, 地道正行, 『会計と平等 –付加価値分配率の探索的データ解析–』, 日本会計研究学会, 第 78 回大会, 神戸学院大学ポートアイランドキャンパス, 2019 年 9 月 9 日 (月).
- (7) M. Jimichi, D. Miyamoto*, C. Saka, and S. Nagata, *Exploratory Financial Big Data Analysis and Reproducible Research*, 2019 年度年会統計関連学会連合大会, 滋賀大学彦根キャンパス, 2019 年 9 月 10 日

(火).

- (8) 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と再現可能研究』, RIMS 共同研究「マクロ経済動学の非線形数理」, 京都大学数理解析研究所, 2019 年 10 月 17 日 (木).
- (9) 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 –前処理の並列化–』, 日本計算機統計学会 第 33 回シンポジウム, 青山学院大学 青山キャンパス, 2019 年 11 月 30 日 (土).
- (10) 阪 智香*, 『財務ビッグデータの探索的データ解析 –企業の租税回避と付加価値分配–』, 統計数理研究所・リスク解析戦略研究センター, 第 7 回 金融シンポジウム, フクラシア丸の内オアゾ, 2019 年 12 月 5 日 (木).
- (11) 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 –前処理の並列化–』, 2019 年度日本経営数学会 秋季研究会, 専修大学 神田キャンパス, 2019 年 12 月 7 日 (土). (査読あり)
- (12) 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 –前処理とデータラングリングの並列化–』, 統計数理研究所共同研究集会 2019 年度「データ解析環境 R の整備と利用」, 統計数理研究所, 2019 年 12 月 21 日 (土).
- (13) 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).
- (14) 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).
- (15) 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).
- (16) 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).
- (17) 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).
- (18) 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).

- (19) 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).
- (20) 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).
- (21) 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).
- (22) 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)
- (23) 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).
- (24) 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).

その他 (評論)

- (1) 阪 智香, 『世界を正しく知るために、企業活動の実態を解析』, 東洋経済 ACADEMIC、SDGs 大学特集 Vol.2, 東洋経済新報社, pp. 14–15, 2020 年 6 月.
- (2) 阪 智香, 『会計情報から企業行動を読み解く』, 旬刊経理情報, 談・論 (巻頭言), 第 1597 号, 中央経済社, p. 1, 2020 年 12 月.
- (3) 阪 智香, 『データは対話のツール』, 企業会計, 第 73 巻, 第 1 号, pp. 42–43, 中央経済社, 2021 年 1 月.