jh191002-NWJ

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, 企業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 157 カ国・9 万社超の上場企業の (最長) 30 年間・86 系列の財務データ (Osiris, 300 万行, 1.4GB), および非上場企業を含む 2,400 万社超の 10 年間・84 系列の世界最大規模の財務データ (Orbis, 2.4 億行, 124GB 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）に基づき, 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ダイナミックで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) 本研究では, 以下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, 処理速度 (ベロシティー) を改善する研究を行う:

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	0siris	上場	連結
DS-Osiris-U-2018	2018	0siris	上場	非連結
DS-Orbis-C-2018	2018	0rbis	上場・非上場	連結
DS-Orbis-U-2018	2018	0rbis	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

[*1] Bureau van Dijk（BvD） のデータベースから抽出

(2) (1) で処理された世界規模の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ダイナミックでインタラクティブなデータ可視化（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に関する新しい知見と課題を明らかにする.

(3) (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, 近年その重要性が指摘されている, 再現可能性を全体に対して確保することも, 本研究の目的である.

また, 本研究が扱う財務データは, 既存研究が扱う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視点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は, その規模が 120GB 超の複数のテキストファイルであり, 通常の計算機環境のメモリ容量を超える. これらは, 抽出してそのまま解析することは不可能であり, データ形式の変更（ヘッダー部分とデータ部分の分離）, 文字・行末コードの変換, 欠損値の置換などの前処理が必要である. 具体的には, Unix 系 OS のコマンド（`grep`, `sed`, `dos2unix` 等）とデータ解析環境 R を利用して前処理を行う必要がある. この工程を, GNU parallel 等の並列処理環境を利用することによって高速化するために, 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・2018 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 年度と 2017 年度に抽出されたデータセットにもとづくものである. また, 非上場企業を含む世界の全企業の財務データベース Orbis については, データ前処理とラングリングを行った. 2018 年度に実施した会計学的・統計学的研究成果について, 次の四つのテーマに分けて概要を述べる. なお, ここで与えられる研究成果の総合報告を, JHPCN 第 11 回シンポジウム (国内会議 (4)) にて行った.

(A) **企業の富の偏在と国際・国内格差について**

企業の富の偏在と格差の現状の証拠を示すため, (1) 国家間で企業の富が過度に集中し, (2) ストック（資産）の集中はフロー（利益）の集中より大きいこと（図 1）, (3) 企業間格差も拡大傾向にあること, (4) Piketty が示した格差のメカニズム（資本利益率 $r >$ 成長性 g）が過去 30 年間の企業データでも見られること等を明らかにした.

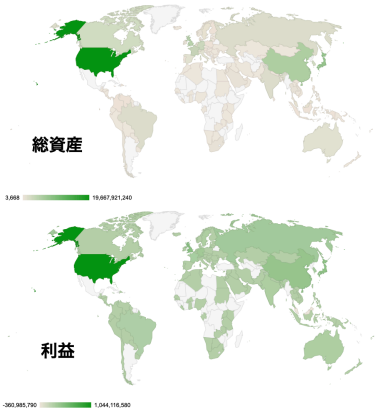

図 1　世界 157 か国, 全上場企業の資産合計・利益合計の国別分布のジオチャート

(B) **付加価値分配に見る富の移転**

世界の全上場企業の, ステークホルダーへの付加価値分配の実態を確認したところ, 過去 25 年間で, 企業は従業員への分配を減少させ, 利益を増加させていること (図 2) を明らかにした.

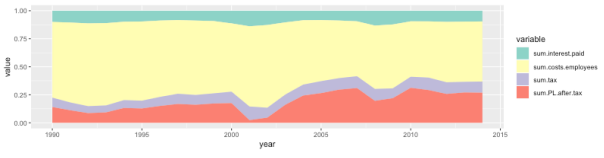

図 2　付加価値分配 (157 カ国, 25 年間)

(C) **企業の租税回避の研究**

租税回避が注目される中, 企業の付加価値分配の 1 つである税金支払に着目し, (a) 租税回避の蓋然性の証拠（支払税金ゼロのラインに企業が集中, 実効税率が法定税率を下回る実態, 実効税率-法定税率が全体で負）と, (b) 税率の引下競争により過去 15 年間に企業の実効税率と各国法定税率が世界規模で下方に収斂した実態を示した (学術論文 (7) と図 3参照).

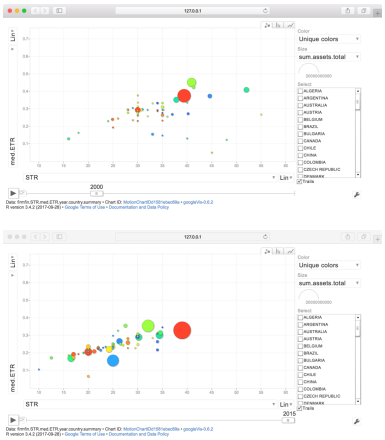

図 3　58 か国の法定税率 (x 軸), 実効税率 (y 軸), (上 2000 年, 下 2015 年)

(D) **再現可能研究の観点からのデータの前処理, ラングリング及び探索的データ解析の研究**

非対称ティー分布を誤差分布としてもつ両対数モデルが, 上場企業（Osiris）の売上高を予測するために利用できるという結果を得た (学術論文 (3) 参照). また, この結果を得るために, データの前処理・ラングリングを行い, 探索的データ解析を実行した結果を動的に文書生成するまでの全工程を, UNIX の `make` コマンドによって実行し, 再現可能研究として行った (学術論文 (1), (2) と図 4 参照).

さらに, 非上場企業を含む全企業のデータ (Orbis) について, 再現可能性を確保した上で, 前処理, とラングリングを Spark と MySQL 環境のもとで行った. なお, 結果については国内学会 (1) において報告された.

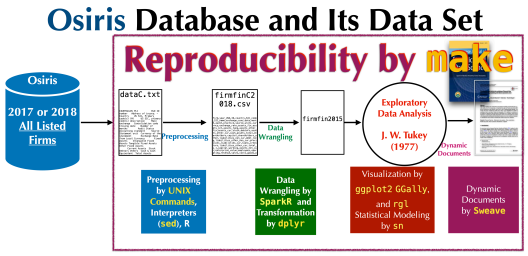

図 4　再現可能研究: Osiris データの前処理・ラングリング, 探索的データ解析, 動的文書生成を `make` コマンドによって自動実行

5　今年度の研究成果の詳細

今年度の研究成果を以下に列挙する:

(A) データセット DS-Osiris-C-2018, DS-Osiris-U-2018, DS-Orbis-C-2018, DS-Orbis-U-2018 の前処理を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, 大幅に改善する可能性があることがテスト環境で検証できた (図 5参照).

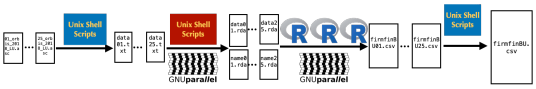

図 5　財務ビッグデータ (データセット DS-Orbis-U-2018) の並列化前処理の全工程

この結果を, 論文 (学術論文 (10)) で報告すると共に, 学会報告を行った (国内会議 (5), (8), (9), (10), (12), (13)).

さらに, この並列化によって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定された PG-Strom 環境で利用できるようにデータベース化し, データラングリングを行ったところ, それまで Spark 環境で行うよりも高速化できることが, 実験的に検証できた (図 6). なお,

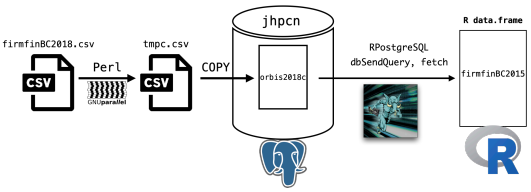

図 6　Orbis データの PG-Strom によるデータラングリング

この結果については, 学会報告 (国内会議 (13)) を行うと共に, 今後論文として発表する予定である.

(B) 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, 論文 (学術論文 (6)) として公表し, Sweave を利用し, LaTeX 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. なお, この結果を得る工程を, ダイナミックかつインタラクティブな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も, この論文で報告した (図 7).

(C) 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の実態から, 従業員賃金が

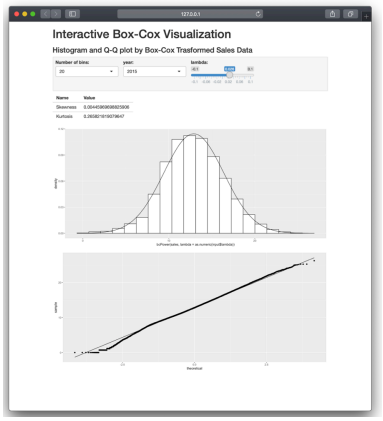

図 7　RStudio Shiny による Box-Cox 変換のインタラクティブなデータ可視化

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論文 (学術論文 (5), (7)) で報告すると共に, 学会報告を行った (国際会議 (2), 国内会議 (3), (7)).

(D) データセット DS-Osiris-C-2018, DS-Orbis-U-2018 を用いて, 以下の二つの考察を行った:

(1) 非上場企業の租税回避の蓋然性の確認
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 160 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 8 参照).

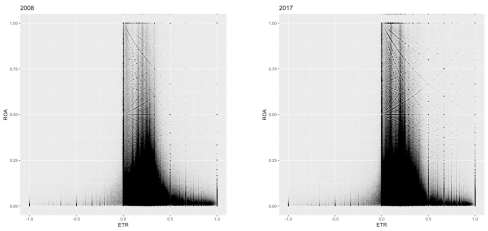

図 8　ETR (x 軸) と ROA (y 軸) の散布図: 2008 年, 2017 年

(2) 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
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 9 参照). 人工資本や人的資本 (図 9 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 9 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 10).
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

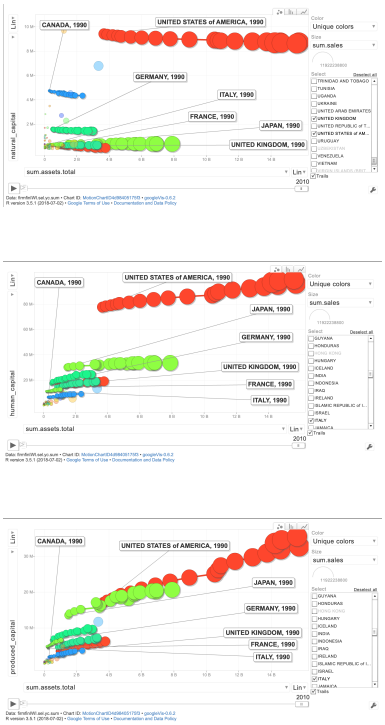

図 9　140 カ国企業資本を x 軸, 環境資本 (上段)・人的資本 (中段)・人工資本 (下段) を y 軸にとったモーションチャート: G7 の 1990 年から 2010 年までの推移の軌跡

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民主的な参加やガバナンスに不可欠である. 本研究によってその証拠を多くの人々に提示し, 企業・ステークホルダーの行動やルールの変化を促すために役立つよう, 引き続き様々な視点から分析を行う.

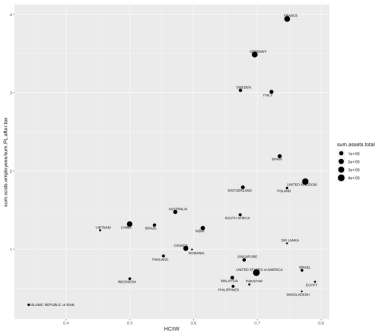

図 10　新国富に占める人的資本の割合 (x 軸) と当期純利益に対する従業員給付の割合 (y 軸) のバブルチャート: 資産合計の総和を円の大きさにマッピング

(E) 学術論文 (3) の結果をデータセット DS-Osiris-C-2018 を用いて検証し, 全工程が再現することを確認した (図 4 も参照). この結果については, 学術論文 (5) 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), (8), (9)).

6　今年度の進捗状況と今後の展望

これまでの研究成果をふまえて, 今後の展望として, 以下の研究テーマがあげられる:

(1) これまでの研究でも利用してきた財務ビッグデータを GNU parallel 等の並列処理環境を利用し, さらに 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, 可視化を行うためのデータ前処理の速度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. さらに, 2019 年度に実験的に利用した PG-Strom と GPGPU を連動させることによってデータを並列分散型で取得するフレームワークを, 2020 年度は Apache Arrow も活用することによって, さらに高速にデータの前処理, ラングリング, 解析を

実行することを目指す．なお, 現在, 新たに抽出中の 2020 年のデータは, さらに規模が大きくなっており, このデータセットは次年度の研究対象とする予定である．

(2) (1) で処理された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 (付加価値分配と人的資本, 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) (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, 次年度は企業サイズ分布 (Firm SizeDistribution: FSD) や確率フロンティア分析 (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予定である．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．

7 研究業績一覧（発表予定も含む）

学術論文

(1) 地道正行,『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–』, 商学論究, 第 66 巻第 1 号, 関西学院大学商学研究会, pp. 1–31, 2018 年 9 月.

(2) 地道正行,『探索的財務ビッグデータ解析 –データ可視化, 統計モデリング, モデル選択, モデル評価, 動的文書生成, 再現可能研究–』, 商学論究, 第 66 巻第 2 号, 関西学院大学商学研究会, pp. 1–41, 2018 年 12 月.

(3) 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

(4) 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.

(5) 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, 2019 年 8 月 13 日, 投稿中.

(6) 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.

(7) 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, 2019, pp. 695–724, Emerald Publishing Limited.

(8) 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.

(9) 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.

(10) 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.

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).

(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) 齊藤 美桜里 *, 地道正行『R による GIS データの可視化』, 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).

(7) 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).

(8) 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).

(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).

(10) 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).

(11) 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).

(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).

(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 –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).