jh171002-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　ビューロー・ヴァン・ダイク(BvD)社のデータベース Osiris から 8 万社を超える全世界(148 カ国)の上場企業を対象として抽出された 30 年間の売上高, 利益, 資産などの 86 系列の財務指標に関するデータ(財務ビッグデータ)を, GPGPU 環境で Apache Spark+Hadoop と R (RStudio, SparkR, sparklyr)を連動させて利用することによって, 財務データの構造を解明し,企業行動の実態を明らかにする. その際, 探索的データ解析(Exploratory Data Analysis: EDA) の考えに基づいて時空間の観点から可視化を行い, その結果として得られた知見に基づき財務データの統計モデリングを行う. さらに BvD 社のデータベース Orbis から, 非上場企業を含む 2,000 万社を超える企業の財務指標データを入手することができた. これは財務データとしては最大規模であり, 企業行動に関する全く新しい知見が得られることが期待できる.

1. 共同研究に関する情報

(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

(2) 共同研究分野

- □ 超大規模数値計算系応用分野
- □ 超大規模データ処理系応用分野
- ■ 超大容量ネットワーク技術分野
- □ 超大規模情報システム関連研究分野

(3) 参加研究者の役割分担

地道 正行 (関西学院大学 商学部):

- データ操作・データ整形
- 探索的データ解析にもとづく統計モデリング

宮本 大輔 (奈良先端科学技術大学院大学 情報科学研究科):

- データ解析環境の構築

阪 智香 (関西学院大学 商学部)

- 企業財務データの会計学的考察

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

目的: ビューロー・ヴァン・ダイク(BvD)社のデータベース Osiris から 8 万社を超える全世界 (148 カ国) の上場企業 (含:上場廃止企業) を対象として抽出された 30 年間の売上高, 利益, 資産などの 86 系列の財務指標に関するデータ(財務ビッグデータ)を利用する.探索的データ解析(Exploratory Data Analysis: EDA)に基づいて可視化を行い, その結果として得られた知見に基づいて企業行動の実態の解明と統計モデリングを行う. さらに BvD 社のデータベース Orbis を利用し, 非上場企業を含む 2,000 万社を超える企業の財務指標データを入手することができた. これは最大規模の財務データであり, 企業行動の新しい知見が得られることが期待できる.

意義: 日本にとどまらず全世界の上場・非上場企業の各種財務指標の構造と企業行動の実態を大規模な財務データにもとづいて時空間の観点から可視化すると共に, 複雑な統計モデリングを行うことは, これまで計算機環境の制約から困難な課題であった. 本研究では, GPGPU 環境で Apache Spark+Hadoop と R (RStudio, SparkR, sparklyr)を連動させることによって, 世界で初めて財務ビッグデータを用いて財務データの構造と企業行動の解明が可能となる.

3. 当拠点公募型共同研究として実施した意義

この研究で扱うデータは, その規模(テキストベースで 2GB～100GB 程度)から, 通常

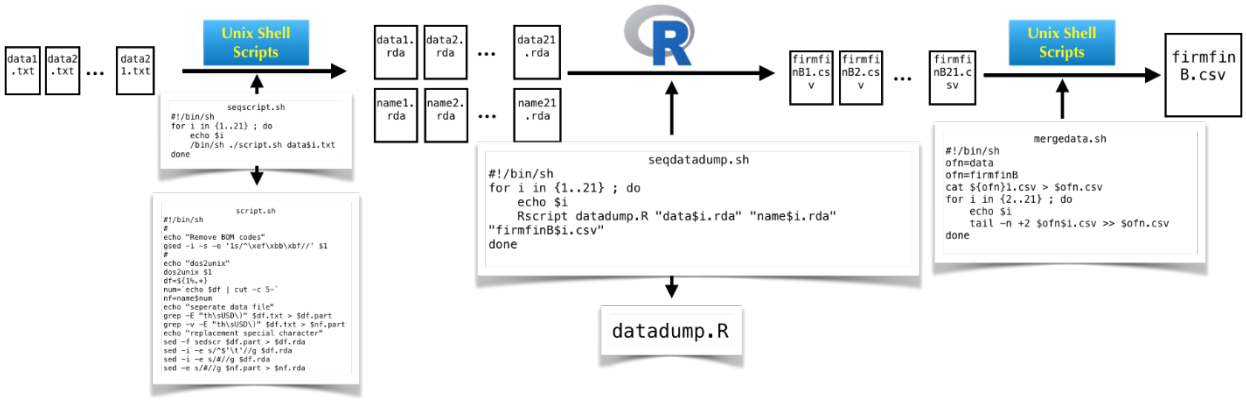

図 1　(DS-Orbis) データ処理の流れの概略

のコンピュータ環境では扱うことが難しい. 例えば, データを拠点から転送することや, 拠点内でネットワークを介した分散環境下で利用することが難しく, 研究を進める上で GPGPU 等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であった. これにより, コンピュータ・シミュレーションをベースとした評価法も利用可能となった.

4.　前年度までに得られた研究成果の概要

(今年度から採用された研究課題)

5.　今年度の研究成果の詳細

(1) 本研究に用いるデータセットを JHPCN 環境で利用するためのデータ整形作業

本研究で利用するデータセットは 2 種類である. 1 つめは, データベース Osiris から抽出した世界 148 カ国、約 83,000 万社の上場・上場廃止企業の 30 年間・86 系列の財務指標に関するデータセット(以下(DS-Osiris)と略す), 2 つめは, データベース Orbis から抽出した非上場企業を含む 2,000 万社の連結・単独決算の 10 年間・83 系列のデータセット(以下(DS-Orbis)と略す)である.

これらのデータセットは抽出段階で, 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, 欠損値の置換などが必要であり, そのままデータ解析環境を用いて解析できないため, Unix 系 OS のコマンド(grep, sed, dos2unix 等)を利用し, 次に R を利用して再整形を行い, データ解析環境で利用できるようなデータセットを用意した. これらの工程にはそれ相応の処理能力を有する計算機環境が必要である. なお, (DS-Osiris) を整形した後のデータセットは 1.3GB 程度のテキストファイルであるが, (DS-Orbis) では, 21 個に分割した各 5GB のテキストファイルとして抽出・処理する関係上, 100GB を超えるファイルの処理が必要となり, 東京大学情報基盤センターの高速・高機能な計算機環境が必要不可欠であった（図 1）.

このような巨大なデータセットを実際に解析する際に, データ処理過程を効率的に行うツールをどのように適切に選択するかという問題は, ビッグデータを扱うデータサイエンスの分野で論点となっている. 今回の研究では, 近年注目されている並列分散処理フレームワークである Spark と R を, SparkR を連携させて使用しており, この点からも東京大学情報基盤センターの環境が利用できることは有益であった（図 2）.

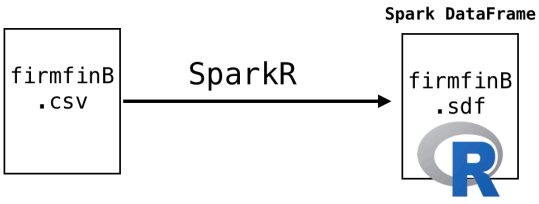

図 2　SparkR を利用したデータ処理

以上のデータ処理に関連する話題については地道ら(会議発表[3])で報告した.

(2) データセットを用いた研究の概要

これらのデータセットを利用して行った4つの研究の概要を述べる.

① 企業の租税回避の研究

現在国際的に注目されている企業の租税回避問題に焦点をあて, 租税回避は短期的には企業の業績を高めても, 長期的にはステークホルダーからの信頼を失い, 企業のサステナビリティを阻害するという仮説に基づき検証を行った. 58 カ国の全上場企業(75,732社)の 20 年間(1995～2015 年)のデータを用い, 企業の租税回避の実態の可視化を行い(図 3), その知見をもとに企業の実効税率の水準と財務的サステナビリティの関連について実証分析を行った. 結果は, (1)企業の租税回避は世界のほとんどの国でみられること, (2)租税回避の水準は企業間で異なることを確認し, (3)租税回避は企業のサステナビリティを阻害すること, が明らかとなった. この研究結果は, Saka et al. (学術論文[3]) として発表し, 先行研究結果を世界的な大規模データで大幅に強化した上で, 租税回避とサステナビリティの関連に関する証拠を初めて提示した.

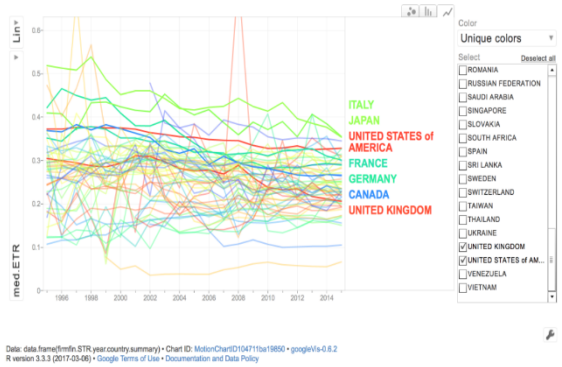

図 3　実効税率の推移（58 カ国、1995-2015 年）

さらに, Saka et al. (学術論文[8])では、別の方法で①租税回避の証拠を示し（図 4[1])、②企業の実効税率の低下と各国法定税率の引き下げ競争によって、過去 20 年間に双方の税率が世界規模で下方に収斂してきた実態を可視化によって示した（図 5）.

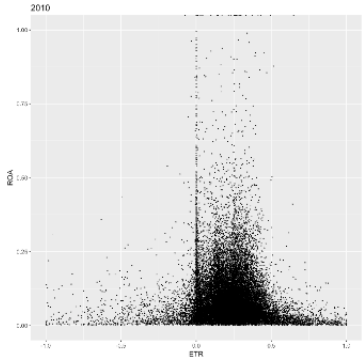

図 4　利益率（ROA、縦軸）、実効税率（横軸）の散布図（2010 年）

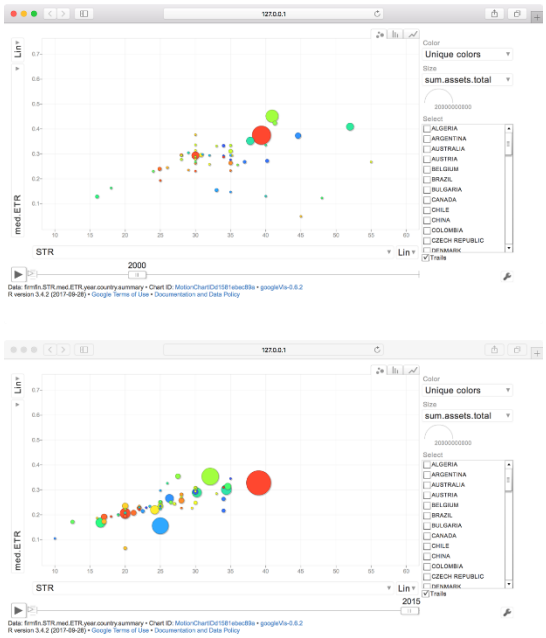

図 5　58 か国の実効税率（縦軸）、法定税率（横軸）のモーションチャート（2000 年, 2015 年）

② 付加価値分配に見る企業の共有価値創造

Oshika and Saka (学術論文[4]) では、100 年以上存続する企業では、株主よりも、従業員等のステークホルダーへの付加価値分配率が高いこと等を示し、長寿の要因を示した。

③企業財務データの可視化と格差の研究

(DS-Orbis)の世界 140 カ国の 1985 年～2013 年の上場企業の財務データを,可視化の手法（R パッケージ dplyr, ggplot2, googleVis 等）を用いて国・企業間の格差について複数の観

1 実効税率 0（税金ゼロ）付近に租税回避の作為的行動が（利益率の高い企業にも）みられる.

点から証拠を提示した. Saka and Jimichi (学術論文[5]) では, (1) Piketty(2014)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 年の企業データでも見られること, (2) 国家間で企業の富が過度に集中し, (3) 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 6), (4) 企業間格差も拡大傾向にあり、世界全体ではトップ 1%企業の売上占有率は 48%、トップ 10%では 86%にのぼること等を明らかにした.

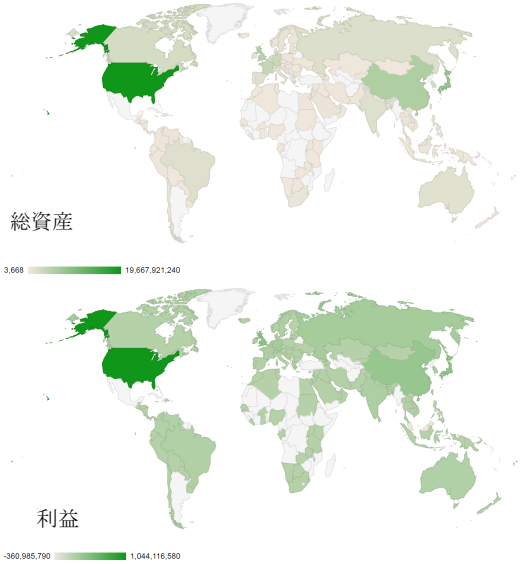

図 6　世界 140 か国、全上場企業の総資産合計・利益合計の分布のジオチャート（2013 年）

④財務データの構造の解明に関する研究

経済学における「生産関数」の推定に則して、全世界の 2015 年度の売上高, 従業員数, 総資産を用いて, 探索的データ解析に基づき可視化を行った.その結果, 1 変量, 2 変量, 3 変量全ての場合において原点付近が高密度であり, 原点を離れるに従つて粗になる (右に歪みがある) 現象を確認した. (図 7[2])

なお, 財務データは対数変換することによって正規分布で近似できるケース(対数正規分布)があること, また, 分布の「ボディー」はほぼ正規分布に従うものの「裾」の部分に歪みを持つという特性を反映できる非対称正規分布で近似できることがわかっている (地道 (学術論文[1])).

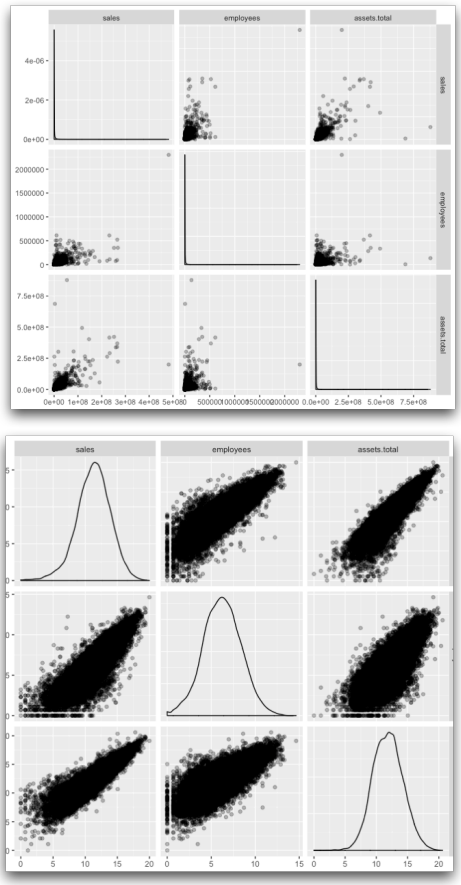

図 7　売上高, 従業員数, 総資産データの構造（148 カ国の全上場企業,上は実数,下は対数変換後の分布と推定された密度関数）

今回の研究では, さらに裾の部分で当てはまりに問題があることが新たな事実として判明し, この結果を反映させた統計モデリングが必要であることがわかった. これを解決するために, 非対称ティー分布を想定し売上高 (対数スケール)に当てはめると, 可視化の結果,当てはまりの程度が改善されることがわかった. (図 8[3])

さらにモデル評価基準である赤池情報量規準と一部のものに関しては竹内情報量規準を用いて比較したところ, 正規分布, 非対称正規分布に比べて非対称ティー分布の当

2　全世界の上場・上場廃止企業 26,682 社の（売上高, 従業員数, 総資産)の対数布図とそれぞれの変数の対数スケールの対数布図. 各散布図とも通常スケールでは, 原点付近に集中しており, 歪みがあるが, 対数スケールでは歪みが解消されていることがわかる.

3　全世界の上場・上場廃止企業 26,682 社の売上高(対数スケール)のヒストグラムに最尤法によって推定された母数をもつ非対称ティー分布の密度関数(統計モデル)を当てはめたもの.

てはまりが良いことが確認できた.

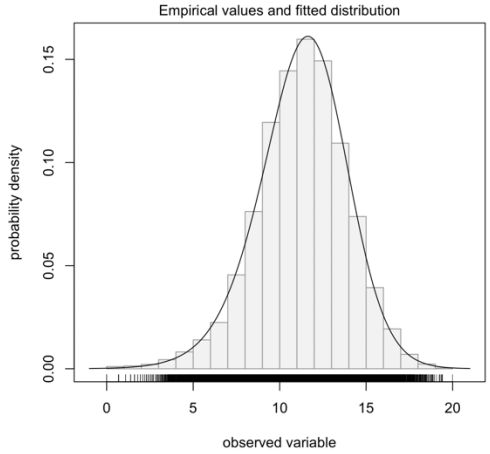

図 8　売上高（対数スケール）のヒストグラムと非対称ティー分布の推定された密度関数（統計モデル）

以上の結果を使い売上高を従業員数と総資産で説明するために誤差項に正規分布, 非対称正規分布, 非対称ティー分布をそれぞれ仮定した対数線形モデルによるモデリングを行ったところ, 非対称ティー分布を仮定したものが良いことがわかり, さらに赤池情報量規準による比較からもこの結果を肯定する結論を得た. (表 1[4])

表 1　正規分布, 非対称正規分布, 非対称ティー分布の当てはまりに関する比較

	df	AIC
lm.log.firmfin2015	4.00	74980.13
selm.log.firmfin2015	5.00	71972.08
selm.ST.log.firmfin2015	6.00	67897.56

以上の結果は, 地道(学術論文[2])で発表した. さらに, 上記の考察で構築されたモデルの予測精度が問題となり, それを *K*-分割交差確認(*K*-fold Cross-Validation) 法を実行することによって検証した. (図 9)

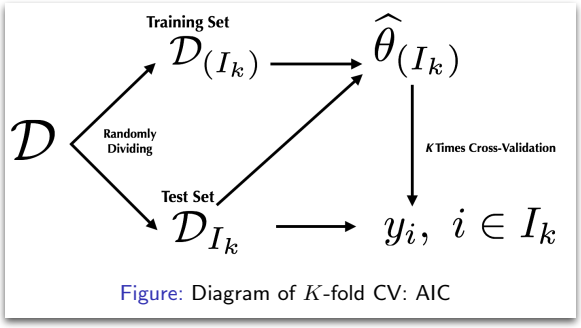

Figure: Diagram of K-fold CV: AIC

図 9　*K*-分割交差検証法（AIC）の概略

その結果, 予測自乗誤差基準に基づくものでは上記の 3 つのモデルの差は明確にでなかったが, 赤池情報量規準に基づくものでは誤差項に非対称ティー分布を仮定した対数線形モデルが最も予測精度が良くなることがわかった. (図 10[5])

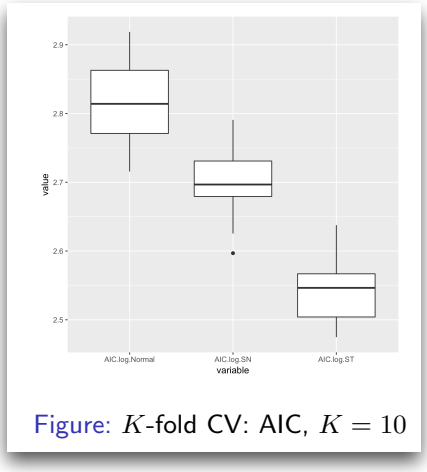

Figure: K-fold CV: AIC, $K = 10$

図 10　*K*-分割交差確認法（赤池情報量規準）によるモデル評価

以上の結果は, Jimichi et al. (会議発表[6]) で報告し, Jimichi et al. (学術論文[7])に公刊した. さらに, 上記は単年度 (2015 年) の結果であるが, 2006〜2015 年の 10 年間においても単年度毎(クロスセクション)と同様の結果が成り立つこともわかった. この結果は, 今後報告・論文発表する予定である. (図 11[6])

---

4 非対称ティー分布を誤差にもつ対数線形モデル (selm.ST.log.firmfin2015) の AIC の値が最小であり, この結果として, このモデルが選択される.

5 データ 1 個あたりの AIC の値を乖離関数として採用し, データセットを 10 分割した場合の *K*-分割検証の実行結果をボックスプロットで可視化したもの. 非対称ティー分布を誤差にもつ対数線形モデル (AIC.log.ST) が予測誤差が最も小さいという結果となった.

6 この図は, (売上高, 従業員数, 総資産) の 3 次元散布図（対数スケール）に予測平面(標本回帰平面)を重ね書きしたものであり, 2006 年から 2015 年の 10 年間の推移をアニメーション形式にしたものである. 紙面では伝わりにくいが, この期間の予測式が安定的であることが結果としてわかる.

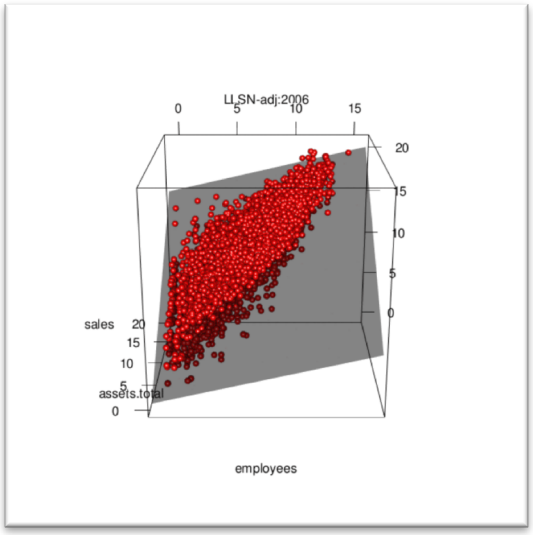

図 11　2006 年~2015 年の売上高(対数)の予測平面

なお, 以上の結果を公表する際には, データ処理の過程も含めて, R に付属の Sweave 関数と Unix コマンドの make を利用した動的文書生成環境を利用し, 再現可能研究を行った. (図 12) これらの詳細は, 地道, 豊原(その他[2])として公刊した.

6. 今年度の進捗状況と今後の展望

本研究は 2017 年 3 月に採択され, 2017 年 6 月に本格的にプロジェクトを開始した. 今年度(2017 年度)の主な進捗を時系列的に列挙する. (ただし, 研究発表・論文執筆は割愛した.)

2017 年 6 月

・東京大学情報基盤センターの専有利用型リアルタイムデータ解析ノード(FENNEL)への接続テスト

・FENNEL 上で Spark, R, SparkR, Hadoop 等の環境整備

・対数非対称分布族による財務データの統計モデリングの基礎研究

2017 年 7 月

・ (DS-Osiris) のサーバーへのアップロード

・Spark と R 環境による DataFrame 化とデータ検証

・ (DS-Orbis) の検証

2017 年 8 月

・Spark と R 環境を用いた (DS-Osiris) のデータ可視化, および, 誤差項が非対称分布族に従う対数線形モデルによる統計モデリングの検証

・統計的機械学習によるモデル評価の検討

・ (DS-Orbis) の検証

2017 年 9 月

・ (DS-Orbis) の検証

・K-分割交互検証法による対数線形モデルの予測誤差の検証

・関西学院大学から FENNEL 環境へ IPsec, IKE, L2TP を用いた接続環境整備

2017 年 10 月

・K-分割交互検証法による対数線形モデルの予測誤差の検証

・FENNEL 環境の GPGPU を R から利用するためのパッケージ gputools と gpuR の整備とチェック

・ (DS-Orbis) を FENNEL 環境に転送・整備

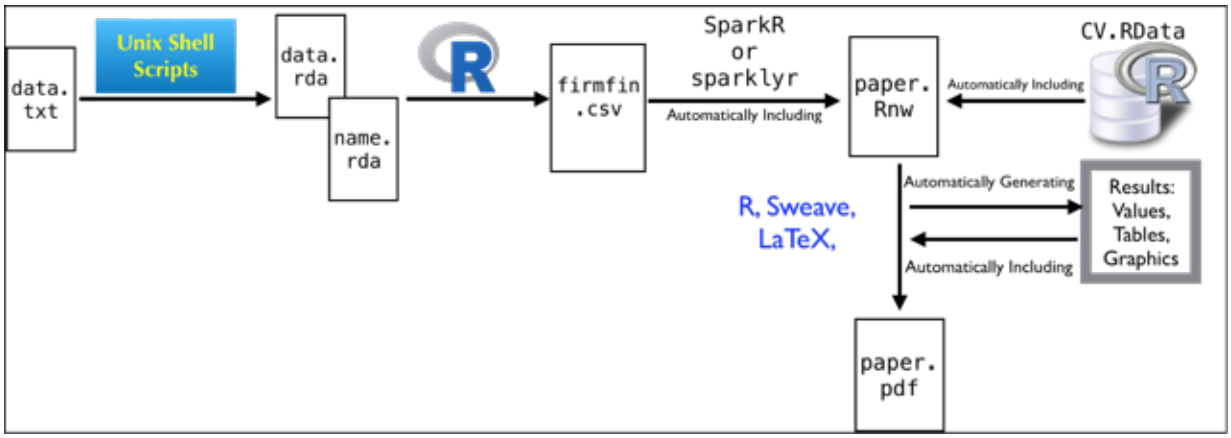

図 12　動的文書生成環境

・ (DS-Orbis) の SparkR を利用した読み込みとデータ加工

2017 年 11 月~12 月

・ (DS-Orbis) の要約と可視化

・gputools と gpuR パッケージを利用した R パッケージ sn の関数を GPGPU 環境へ実装

2018 年 1 月~3 月

・ (DS-Orbis) の要約と可視化

・gputools と gpuR パッケージを利用した R パッケージ sn の関数を GPGPU 環境へ実装

・ (DS-Orbis) の分散環境化とチューニング

・誤差項が非対称分布族に従う対数線形モデルの時間的推移に伴う安定性のチェック

データセット (DS-Osiris) に関する今後の方向性としては, FENNEL 環境における GPGPU を R における gputools と gpuR パッケージを利用してチューニングし, 対数線形モデルに対して「一個抜き交差確認」(LOOCV) 法を実行することにより, さらに高精度の予測誤差の検証を行うことが可能となる. また, 今年度のモデリングでは時間を固定した (クロスセクショナル) 観点からのモデリングを主に考察したが, 時間的な推移を考慮した観点からのモデリングを行うことによって, 時間的な予測が可能となる.

一方, データセット (DS-Orbis) の検証には, そのサイズ (100GB 超) に起因して, 要約と可視化を行うまで整理できた. (図 13[7])しかしながら, Hadoop と Hive 環境を利用して分散環境下で処理する試みが, 必ずしも「熟れた」段階 (スピード, 安定性など) に達しているとはいいがたく, さらに(DS-Osiris)とは異なった分布構造をもつことも可視化からわかった. これらの点についてチューニングを行うと共により深く検討する.

また, (DS-Orbis) は抽出段階で連結決算と単独決算の企業を分離した方法で別々に再抽出する方が研究の目的上, 得策であることがわかったため再度抽出することとした.

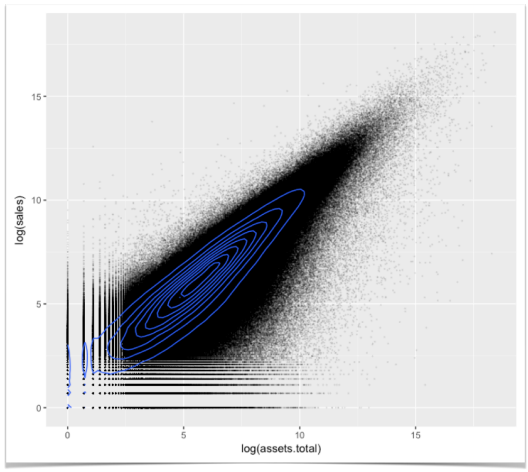

図 13　(DS-Orbis)による全世界の上場・非上場企業の売上, 総資産の可視化

以上をふまえて, (DS-Osiris)と再抽出する (DS-Orbis) を利用し, 2018 年度は次のテーマに沿って研究を続ける予定である:

(1) (DS-Orbis) の分散環境下でチューニングと R 関数の GPGPU 環境下でのチューニング

(2) 非対称誤差をもつ対数線形モデルのフィッティングと交差確認法(LOOCV 法)による検証

(3) 時間的な推移を考慮した観点からのモデリング

(4) 全世界の上場企業の租税回避行動の実態の可視化とサステナビリティとの関連の時系列分析

(5) 全世界の上場・非上場企業の付加価値分配等の企業行動と, 企業活動のグローバルがもたらす格差等の実態の証拠を示し,社会的課題の解決に向けての方策を提示しその経済的帰結を予測

(6) データセット (DS-Orbis) と国連の Inclusive Wealth（新国富）データから, 企業活動が人工資本・人的資本・自然資本に与える影響と格差等の分析

7 全世界の上場・非上場企業のうち, 単独決算企業 3,437,796 社の総資産と売上高の散布図. 軸は, 対数スケールであり, 推定された 2 次元密度関数がプロットされている.

7. 研究成果リスト

(1) 学術論文

[1] 地道正行 (2017)『R による対数非対称正規線形モデルによる財務データの統計モデリング』, 商学論究, 第 64 巻, 第 5 号, pp. 159-185, 2017 年 3 月, 関西学院大学商学研究会.

[2] 地道正行 (2017)『R を利用した対数非対称分布族にもとづく財務データの統計モデリング』, 経済学論究, 第 71 巻, 第 2 号, pp. 141-174, 2017 年 9 月, 関西学院大学経済学部研究会.

[3] Saka, C., Oshika, T. and Jimichi, M. (2017) "Does Tax Avoidance Diminish Sustainability?", SSRN, http://ssrn.com/abstract=3061565.

[4] Oshika, T. and Saka, C. (2017) "Sustainability KPIs for Integrated Reporting", *Social Responsibility Journal*, Vol. 13, No. 3. pp. 625-642.

[5] Saka, C. and Jimichi, M. (2017) "Evidence of Inequality from Accounting Data Visualisation", *Taiwan Accounting Review*, Vol. 13, No. 2, pp. 193-234, December 2017.

[6] Saka, C., Noda, A. and Jimichi, M. (2018) "Cultural Influence on Corporate Social Responsibility Disclosure in East Asia", *International Review of Business*, No. 18, pp. 1-28.

[7] Jimichi, M., Miyamoto, D., Saka, C. and Nagata, S. (2018) "Visualization and Statistical Modeling of Financial Big Data: Log-Linear Modeling with Skew Error", submitted.

[8] Saka, C., Oshika, T. and Jimichi, M. (2018), "Visualization of World-Scale Evidence on Tax Avoidance and Tax Rate Convergence", SSRN, http://ssrn.com/abstract=3115022, submitted.

(2) 国際会議プロシーディングス

(該当なし)

(3) 会議発表

[1] 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』, JHPCN: 学際大規模情報基盤共同利用・共同研究拠点 第 9 回 シンポジウム（於：THE GRAND HALL（品川）2017 年 7 月.

[2] Jimichi, M. "Applied Feasible Generalized Ridge Regression Estimation to Linear Basis Function Models", 2017 Conference of the International Federation of Classification Societies, August 8, 2017. (招待講演)

[3] 地道正行, 宮本大輔, 阪智香, 永田修一『Spark + R 環境を利用した財務ビッグデータ解析』, 国際数理科学協会年次大会「統計的推測と統計ファイナンス」分科会研究集会（於：大阪府立大学）2017 年 8 月.

[4] Saka, C., Oshika, T. and Jimichi, M. "Does Tax Avoidance Diminish Sustainability?", Meditari Accounting Research Conference 2017 - Global Perspectives in Accounting Research, September, 2017.

[5] 阪智香, 大鹿智基, 地道正行『租税回避とサステナビリティ』, 日本社会関連会計学会第 30 回全国大会（於：法政大学）2017 年 10 月.

[6] Jimichi, M., Miyamoto, D., Saka, C. and Nagata, S. "Visualization and Statistical Modeling of Financial Big Data", Joint Meeting of 10th Asian Regional Section of the International Association for Statistical Computing and the NZ Statistical Association, December, 2017

[7] 阪智香, 大鹿智基, 地道正行『サステナビリティと税務行動の関係について』, 日本ディスクロージャー研究学会第 16 回研究大会（於：法政大学）2017 年 12 月.

(4) その他（特許，プレス発表，著書等）

[1] 阪智香「会計ビッグデータの可視化」,『企業会計』第 70 巻第 4 号, 中央経済社, pp. 4-5, 2018 年 3 月.

[2] 地道正行, 豊原法彦『景気先行指数の動的文書生成にもとづく再現可能研究』, 豊原法彦編著『関西経済の構造分析』, 第 5 章, pp. 77-111, 中央経済社, 2018 年 3 月.