様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　　１　日現在

機関番号：　３２６１２
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３００４８

研究課題名（和文）データサイエンスの基盤：クラウドを活用したDandDインスタンスライブラリの構築

研究課題名（英文）Fundamentals of Data Science: Creation of DandD Instance Library

研究代表者
柴田　里程（Shibata, Ritei）

慶應義塾大学・理工学部（矢上）・名誉教授

研究者番号：６００８９８２８

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：データサイエンスの基盤確立を目指した基礎研究を行うとともに，その成果を活用した高度なデータサイエンス実践をサポートするソフトウエア環境TRADを構築した．TRAD は TextilePlot，R，DandD を統合したデータ解析環境であるが，さまざまな形態でネットワーク上に存在するデータを自己説明的に併用するためのメディアとして， DandD (Data and Descrtion)を採用し，ニュートラルなデータ可視化を TextilePlot で，高度な解析とモデリングを R で行うシームレスな環境である．厚労省の患者調査データを DandD ライブラリ例として構築し，公開している．

研究成果の概要（英文）：Aiming at further development of data science, fundamentals of data science are investigated. The result is implemented as a software environment TRAD, which is an integrated environment of TextilePlot, R and DandD(Data and Description), which is an XML instance which supports full use of variety of data scattered over internet, TextilePlot is an advanced software to visualize data as it is, and R is well known software to execute a deep data analysis. Seamless combination of such three elements enables everyone to do efficient and advanced data analysis to find a model for underlying penomena. We also created a DandD instance library based on open source data "Patient survey" conducted by Ministry of Health and Labor, Japan.

研究分野：　Data Science

キーワード：　TextilePlot　DandD　R　Environment　TRAD

１．研究開始当初の背景

社会の至るところで，データが大量に蓄積されるようになったにもかかわらず，必ずしも有効活用されず眠ったままという，もったいない状況を打開する，ひとつの解決策として２０１０年代からデータサイエンスの実践が提唱されるようになったが，その基盤がかならずしも確立していない現状では，どう実践したら新たな価値を見出せるのかわからず右往左往するだけというのが社会における実態であった．特にデータにもとづく客観的な判断が重視される欧米と比べ，そのような歴史が浅い日本では戸惑いも大きかった．このような背景のもと，すでに２０年以上にわたってデータサイエンスという新しいパラダイムを提唱してきた本研究の代表者は，データサイエンスの基盤確立とその成果を反映したソフトウエア環境の実装を目指し，本研究を開始することにした．

２．研究の目的

(1) データ解析の基礎理論
- ① ゲノムデータ解析
- ② 汚染に強い適合度検定
- ③ 空間データ解析モデル

(2) データサイエンスの基盤確立
- ① データ変容のモデル化
- ② データ型の再定義
- ③ データの正規化と逆正規化

(3) データサイエンス実践環境の実装
- ① クラウド利用可能性の検証
- ② TRAD の実装

(4) オープンデータの活用
- ① オープンデータ特に官庁公開データの現状把握
- ② オープンデータの DandD インスタンスライブラリ構築

３．研究の方法

代表者のこれまでの研究蓄積をフル活用し，連携研究者，研究協力者の支援を得ながら，理論，実践，ソフトウエア開発の 3 側面から総合的な研究を行った．

４．研究成果

(1) データ解析の基礎理論

① ゲノムデータ解析

家系図情報を利用した疾患遺伝子の座位同定アルゴリズムを開発し，実データによる検証を行った

② 汚染に強い適合度検定

実データの解析では，さまざまな原因でデータが汚染されていることが多い．データサイエンスの基盤構築の一つとして，このような汚染に強い推測法だけでなく，多少の汚染があってもそれに対して頑健なモデルの適合性を判断する方法が必要である．そのため本研究では，すでに海洋調査データでその有用性が確かめられている Cramer-von Mises 距離にもとづく頑健な検定法の理論的な正当化を行った．

③ 空間データ解析モデル

空間上のデータの解析には，天体から不動産価格まで様々な応用分野があるが，適切なモデルを用いないかぎり，表面的な解析にとどまる．本研究では，空間自己回帰モデルの実用化に向け，積分などを含まない形でのパラメータ推定を行う効率的なアルゴリズムを発見し，その最適性の証明に成功した．ただ，空間には時間のように過去から未来へといった流れがないため，局所的には最適であるが，大域的には数多くの解が存在し，場合によってはうまく解がみつからないことがあることも明らかになったため，この問題をどうクリアーするかが新たな課題として残されている．

(2) データサイエンスの基盤確立

① データ変容のモデル化

データはその取得から解析まで，ひとつの流れを作るが，その過程で，さまざまな姿に形を変える．本研究では，変容の仮定で必要となる「関係の関係」をどう表現するか研究を進めた．リレーショナルデータベースの基本は関係形式であるが，複数の関係形式データの間には，基本的に「値の共有」，「値のマッピング」，「値制約」の 3 種類の関係があることがわかった．そこで拡張ドメインの概念を導入することで，これらを統一的に記述することに成功した．この結果は DandD ルールとしても定式化することができ，実装待ちの段階にある．

② データ型の再定義

データの型はその目的によって様々に定義されるが，データサイエンスの基盤としては，どのようなデータ型を導入する必要があるのか明らかにはされてこなかった．本研究では，さまざまな側面から研究を進めた結果，データサイエンスの実践にあたって本当に必要な

型は, Measurement と Mark だけであり, Measurement のサブタイプとして，Ordinal, Cardinal, Frequency, Date, Mark のサブタイプとして Ordered Mark と Logical があれば必要十分な型であることを示すことができた.

③ データの正規化と逆正規化

リレーショナルデータベースの正規形の概念は, データベースの運用と管理を目的としており, データサイエンスの実践に当たって, いつでも正規形が望ましいとは限らない. 必要に応じて正規化と逆正規化を行き来できるほうがデータ解析の柔軟性を確保できることを実証できた.

(3) データサイエンス実践環境の実装

① クラウド利用可能性の検証

本研究の開始に先立ってクラウド利用の可能性を様々な側面から検討した. その結果データサイエンスの実践といった重いシリアスなタスクには現在のクラウドシステムは機能の面からも，負荷の面からも, 時期尚早であることがわかり, クラウド化は将来の課題とすることになった.

② TRAD の実装

クライエントサーバシステムの形で実装されていた TextilePlot, DandD Sever を全面的に組み換え，R とのシームレスな連携機能も加えることで, データサイエンスの基盤環境 TRAD (TextilePlot, R and DandD)を実装することができた. その段階で，上述のようなデータサイエンスの基盤となる研究成果を取り入れるだけでなく, ギガ単位のデータもストレスなく扱え，TextilePlot によるビジュアルインタフェースを活用したフィルタリングや型の変更, 正規化と逆正規化なども自由に行える環境とした. さらに, CSV ファイルからだけでなく，Excel の複数テーブルからも DandD インスタンスを作成し，R との間を自由に行き来できるような設計としたことで, 可用性は高まり, 十分実用の域まで達したソフトウエア環境となった. TRAD は http://datascience.jp で公開しており, だれでも自由にダウンロードし利用できるようになっている.

(4) オープンデータの活用

① オープンデータとくに官庁公開データの現状把握

近年, e-stat という形で官公庁データも自由にダウンロードし利用できるようになったが, これまでの印刷出版の形態を踏襲せざるを得ないのか, ダウンロードしてすぐ使える形にはなっていないことが多い. 特にことなるソースの複数のデータを併せて使おうとすると様々な困難に直面し, 大きな作業量が必要となる. ダウンロードの形式として，XML ファイルも選択できるが, 内容としては CSV ファイルや Excel ファイルと同じで, 条件と値の組という形での極端な正規化がなされているため, データを理解するというデータサイエンスの基本を実践するには, 結局テーブルデータである，CSV ファイルや Excel ファイルも参照せざるを得ない. 想定されている利用法は, 目的が明確に定まっている場合にその目的に沿ったソフトウエアを作成し, 利用することのように見受けられるが, これではせっかくの宝を活用しきれないことはあきらかである.

② オープンデータの DandD インスタンスライブラリ構築

DandD インスタンスをベースとする TRAD 環境を利用すれば, 前述のようなオープンデータの問題は大半が解消できる．それを実証するため，300 以上のファイルからなる厚労省の「患者調査データ」を DandD 化する実験を行った. 結果は DandD インスタンスライブラリーとして http://datascience.jp より自由にダウンロードできる．DandD 化することのメリットは, データソースの違いを意識することなく, データの理解を進め自由に解析できることだけでなく，DandD インスタンスの再利用が可能なため, たとえば, 今年予定されている患者調査の結果は, 現在公開されている 2011 年の調査結果の DandD インスタンスのデータ本体の参照を切り替えるだ

けで済むといった大幅な省力化にもある. もちろん多言語対応の機能もこれから役立つに違いない.

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　4　件）

① Y. Rikimaru and R. Shibata (2017) Non-identifiability of simultaneous autoregressive model and singularity of Fisher Information Matrix, *International Journal of Statistics and Probability,* 査読あり（印刷中）

② Y. Rikimaru and R. Shibata (2016) A good approximation of the Gaussian likelihood of simultaneous autoregressive model which yields us an asymptotically efficient estimate of parameters, *Journal of Statistical Planning and Inference* 173, 31-46, 査読あり

③ M. Naka and R. Shibata (2016) Asymptotic Distribution of Cram´er-von Mises Statistic When Contamination Exists. *International Journal of Statistics and Probability,* 5, 90-97, doi:10.5539/ijsp.v5n1p90, 査読あり

④ Y. Sugaya and R. Shibata (2014) Probability inheritance algorithm and Probability inheritance algorithm and its implementation. *J. Statistical Computation and Simulation,* doi:10.1080/00949655.2014.915032, 査読あり

〔学会発表〕（計　11　件）

① 力丸佑紀, 柴田里程, フィッシャー情報量行列が特異になる場合のパラメータ推定　―SAR モデルの場合―, 統計関連学会連合大会, 2016 年 9 月 7 日, 金沢大学（石川県金沢市）

② 柴田里程, 企画セッション「データサイエンスの世界的潮流とその展望」の背景とねらい, 計関連学会連合大会, 2016 年 9 月 7 日, 金沢大学（石川県金沢市）

③ 横内大介, 柴田里程, データサイエンス実践の支援環境 TRAD, 2016 年 9 月 7 日, 金沢大学（石川県金沢市）

④ 仲真弓, 柴田里程, データ解析の初期段階における TextilePlot の活用, 2016 年 9 月 7 日, 金沢大学（石川県金沢市）

⑤ 柴田里程, 横内大介, データサイエンス実践の統合支援環境 TRAD, 2015 年 9 月 9 日, 岡山大学（岡山県岡山市）

⑥ 力丸佑紀, 柴田里程, 空間斉次自己回帰モデルのフィッシャー情報量行列の正則条件, 統計関連学会連合大会, 2015 年 9 月 8 日, 岡山大学（岡山県岡山市）

⑦ 仲 真 弓, 柴 田 里 程, Robustness of Cramer-von Mises statistic under contiguous type contamination, 2015 年 9 月 8 日, 岡山大学（岡山県岡山市）

⑧ 仲真弓, 柴田里程, Cramer-von Mises 距離推定量を用いたときの適合度検定のロバスト性, 統計関連学会連合大会, 2014 年 9 月 15 日, 東京大学（東京都文京区）

⑨ 力丸佑紀, 柴田里程, 空間斉次自己回帰モデルの乱数生成とそれに基づく実験, 統計関連学会連合大会, 2014 年 9 月 14 日, 東京大学（東京都文京区）

⑩ 柴田里程, データサイエンスの基礎：関係の関係, 統計関連学会連合大会, 2014 年 9 月 14 日, 東京大学（東京都文京区）

⑪ 仲真弓, 柴田里程, 高次元可視化環境 TextilePlot によるダイナミックな回帰診断, 統計関連学会連合大会, 2014 年 9 月 14 日, 東京大学（東京都文京区）

〔図書〕（計　1　件）

① 柴田里程, 近代科学社, データ分析とデータサイエンス, 2015, 260

〔産業財産権〕

○出願状況（計 0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等:
http://datascience.jp

６．研究組織
(1)研究代表者

柴田　里程　（SHIBATA, Ritei）
慶應義塾大学・理工学部・名誉教授
研究者番号：60089828

(2)研究分担者
(　　)

研究者番号：

(3)連携研究者
横内　大介（YOKOUCHI, Daisuke）
一橋大学・国際企業戦略科・准教授
研究者番号：50407144

(4)研究協力者
島津　秀康（SHIMADZU Hideyasu）
英国・ラフバラ大学・専任講師
Peter Thomson
ニュージーランド・SRA・ディレクター