様式 C-19

科学研究費補助金研究成果報告書

平成２４年 ３月３１日現在

研究種目：基盤研究（Ｂ）
研究期間：2007～2010
課題番号：19300097
研究課題名（和文） 高次元大規模データのモデル化を助けるデータヴィジュアリゼーションの理論と実際

研究課題名（英文）Theory and Practice of Data Visualization for Modeling Complex Large Scale Data

研究代表者
柴田 里程（SHIBATA RITEI）
慶應義塾大学・理工学部・教授
研究者番号：60089828

1. 研究成果の概要（和文）：大規模で高次元なデータからのモデル化を助けるための効果的なデータヴィジュアリゼーション環境の確立を目指して，Textile Plot を中心に研究を進めた．具体的には，ゲノム解析，金融データ解析，海洋調査データの解析といった実際問題へ適用することによってその有効性を確かめるとともに問題点と改良点を明らかにする形で進めた．結果として超高次元でも適用可能なアルゴリズムの開発に成功し，数々の理論的な成果も得ることができた．

研究成果の概要（英文）：The aim of this project is to validate the effectiveness of data visualization technique like Textile Plot for analysis of large complex data to find out new aspects of underlying phenomena and result in a better modeling. The project has been conducted through real application to real problems. Three application fields are taken into account in this project, genome analysis, financial data analysis and marine survey analysis. It has been shown that Textile Plot is a quite powerful tool in finding new genomes compared with a traditional way of visualization LD Display. We have developed a new algorithm to be able to deal with super high, for example several million dimensional data, which is common in genome analysis. In the field of finance, it is shown that Textile Plot is a powerful tool in describing difference of hedge fund operations as well as FX rates. We have conducted cooperative research works with Commonwealth Science and Industry Research Organisation in the field of marine survey data analysis. We could find several interesting aspects of the effect of environmental changes to seabed or estuary fauna by using Textile Plot together with PP plots. It led us to develop a new distributional modeling of the weight and number of such animals, as a stationary distribution of a stochastic difference equation for the growth.

交付決定額

（金額単位：円）

	直接経費	間接経費	合 計
2007 年度	2,300,000	690,000	2,990,000
2008 年度	2,000,000	600,000	2,600,000
2009 年度	2,000,000	600,000	2,600,000
2010 年度	2,000,000	600,000	2,600,000
2011 年度			
総 計	8,300,000	2,490,000	10,790,000

研究分野：総合領域
科研費の分科・細目：統計科学
キーワード：データヴィジュアリゼーション、Textile Plot、大規模複雑データ、モデリング

１．研究開始当初の背景

ネットワークと計算機利用技術の高度化に伴い大規模で複雑なデータの取得が容易になってきたにもかかわらず，それを高度に利用する技術が追いつかないためそのまま蓄積されるだけで有効利用されていない状態が社会の至るところに見受けられる．そこで，大規模で複雑なデータの解析を容易にするためのユーザフレンドリーな視覚インタフェースの開発を目的として本研究を開始した．

２．研究の目的

データが有効利用されない大きな理由の一つが，大規模で複雑になると，その全体像をつかむことが困難になり，どう扱ったらよいかその方針も立てづらいことにある．全体像をつかむにはどうしても人間の直感に訴える視覚表現が欠かせないが，大規模で高次元になるとかなり念入りな設計と実験の積み重ねが必要となる．本研究では，代表者らが開発した，次元数に制約のない平行座標プロットを高度化した Textile Plot を研究の中心に据え，その有効性の検証とさらなる改良，補助的に必要となる他の視覚表現の開発，さらにはそれを裏付ける理論の発展を目的とした．

３．研究の方法

このような研究では机上の空論に終始する危険をもっとも警戒しなければならないので，具体的な検証現場として，ゲノム解析，ファイナンスデータ解析，海洋調査データ解析といった大量で高次元のデータの高度な利用を必要としている３分野を柱として設定した．そのうえで Textile Plot などの視覚表現を駆使してどこまでデータの全体像をつかむことができるか検証することとした．どこまで全体像をつかむことができたかは，最終的にはどこまで現象の本質をとらえられたか，それを適切なモデルの構築に結びつけられたかで判断することにした．つまり，データ取得からモデルの検証までの流れにそってトータルに視覚化の効果を検証することにした．

４．研究成果

Textile Plot の有効性がさまざまな形で検証できただけでなく，多くの必要な改善点も判明し，ソフトウエアの改善につなげた．また独立同分布を仮定できる場合の分布検証ツールである QQ (Quantile Quantile) プロットは独立ではあるものの同分布性は仮定できない b 場合には使えないが．そのような場合には PP(Probability Probability)プロットがきわめて有効な視覚表現であることが判明した．Textile Plot と PP プロットを併用することでデータの全体像をつかみ，それを確率微分方程式などで構成された数理モデルに結びつける道筋は極めて有効であることが検証できた．また，独立異分布の場合の推測理論や，PP プロットの数値版である Cramer-Von Mises 適合度統計量の漸近分布の再評価，これを最小にする最小距離推定量の頑健性の証明とその評価など数々の理論的な成果も得られた．以下，各適応分野別に成果をより詳しく述べる．

(ア) ゲノムデータ
DNA シークエンスの解析に Textile Plot が極めて有効であることが判明し，数十万次元でも処理し表示できるアルゴリズムの開発に成功した．その有効性の検証とともに，これまで標準的な方法として広く用いられてきた LD (Linkage Disequilibrium) Display との比較検討を行い，LD Display では発見困難な SNP の差異が Textile Plot ならより幅広く明確に認識できることも判明した．

(イ) ファイナンスデータ
ヘッジファンドの収益率を主な対象として，時間を変量に取る Textile Plot と時間ごとに変量を導入する Textile Plot の併用により，その姿を総合的に理解でき金融データベース利用システムを ICS FinAnalyzer として完成させた．ファイナンス時系列を Textile Plot で眺めるときには，時間軸を軸の一つとしてとるよりは，カラーで区別したほうが有効であることが確かめられた．複数のデータベースからの異なる形式のデータを併合し眺めることを可能とするシステムが FinAnalyzer である．

(ウ)海洋調査データ
オーストラリアの CSIRO(Commonwealth Science and Industry Research Organisation)との共同プロジェクトの一環としておこなった．対象としたデータはトロール漁の海底生物の生態系に対する影響を調べるための大規模な実験調査データ，グレートバリアリーフにおける海洋生物の豊富さと環境要因との関係を探るための長期間にわたる大規模調査データである．Textile Plot 以外の PP プロットなど様々なデータヴィジュアリゼーション手法を併用することで，海底生物のドレッジデータについては，種ごとの，個体数については Thomas 分布が，重量に関しては確率微分方程式の定常解として現れるガンマ分布が適切であることが判明し，その確証実験を行った．

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 6 件）

① Y. Sugaya and R. Shibata, Exploration of the disease locus by a careful evaluation of the likelihood polynomial for pedigree data. *Journal of Human Genetics*, 56, 383-389, 2011, 査読有
② R. Miura, Y. Aoki and D. Yokouchi, A Note on Statistical Models for Individual Hedge Fund Returns, Mathematical Methods of Operations Research, 69, 553-557, 2009, 査読有
③ N. Kumasaka and R. Shibata, High Dimensional Data Visualisation: the Textile Plot, Computational Statistics and Data Analysis, 52, 3636-3644, 2008, 査読有
④ H. Shimadzu, R. Shibata and Y. Ohgi, Modelling Swimmer’Speeds Over the Course of a Race, J. Biomechanics, 41, 549-555, 2008, 査読有
⑤ 熊坂夏彦，柴田里程，Textile Plot 環境，統計数理，55，47-98，2007, 査読有
⑥ M. Tajima, C. Hamada, T. Arai, M. Miyazawa, R. Shibata and A. I. Shino, Characteristic Features of Japanese Women’s Hair With Aging and With Progressing Hair Loss, J. DermatologicalScience, 45, 93-101, 2007, 査読有

〔学会発表〕（計 17 件）

① M. Naka and R. Shibata, Approximation of Cramer-von Mises test statistics by weighted sum of chi-square variables. Workshop on Asymptotic Methods in Data Science, 2011 年 12 月 8 日，慶應義塾大学矢上キャンパス．
② Y. Takei and R. Shibata, Asymptotics of spatial AR parameter estimation. Workshop on Asymptotic Methods in Data Science, 2011 年 12 月 6 日，慶應義塾大学矢上キャンパス．
③ 柴田里程，　離散分布モデルのモデル選択 ―ポアッソンとトーマスを中心に―, 統計関連学会連合大会, 2011 年 9 月 6 日，九州大学
④ 仲真弓，柴田里程，Cramer-von Mises 統計量の漸近表現と MLE・MDE によるパラメータ推定の影響評価, 統計関連学会連合大会，2011 年 9 月 5 日，九州大学
⑤ 仲真弓，柴田里程，パラメータ推定を含む Cramer-von Mise　タイプ適合度検定統計量の漸近分布のパラメータ依存性, 統計関連学会連合大会，2010 年 9 月 7 日，早稲田大学
⑥ M. Naka, R. Shibata, Goodness of fit of Gamma distribution to sea fauna weights, *The International Biometric Society Australasian Region (New Zealand),* 2009 年 11 月 30 日，Lake Taupo, New Zealand
⑦ 柴田里程，菅谷勇樹，独立異分布標本の PP プロットによる分布適合性の検証, *統計関連学会連合大会,* 2009 年 9 月 8 日，同志社大学
⑧ 仲真弓，柴田里程，オーストラリア北部における大規模海底生物調査データの解析，*統計関連学会連合大会,* 2009 年 9 月 7 日，同志社大学
⑨ 菅谷勇樹，柴田里程，確率継承アルゴリズムによるマルコフ樹の尤度評価，*統計関連学会連合大会,* 2009 年 9 月 7 日，同志社大学
⑩ R. Shibata and Y. Tanizawa, Smile Curve and Local Volatility, Australia-Japan Workshop on Data Science, 2008 月 3 月 27 日，慶應義塾大学
⑪ R. Shibata and Y. Sugaya, Modelling Counts in Trawling Data, Australia-Japan Workshop on Data Science, 2008 年 3 月 27 日，慶應義塾大学
⑫ D. Yokouchi and R. Miura, What can we do for hedge fund return data under the DandD environment? Workshop on Data Science, 2008 年 3 月 24 日，慶應義塾大学

⑬ Y.Sugaya and R.Shibata, Likelihood-based method for estimating penetrance and disease susceptibility allele frequency, Workshop on Data Science, 2008年3月25日, 慶應義塾大学
⑭ M.Naka and R.Shibata, Weight distribution in trawling data., Wrokshop on Data Science, 2008年3月24日, 慶應義塾大学
⑮ 島津秀康, 柴田里程, ニューロン活動電位の3ステージダイナミックモデル, 統計関連学会連合大会, 2007年9月8日, 神戸大学
⑯ 菅谷勇樹, 柴田里程, アイスランド家系データにもとづく染色体上の交差確率の変化の推定, 統計関連学会連合大会, 2007年9月7日, 神戸大学
⑰ 熊坂夏彦, 柴田里程, Textile Plotによる高次元非線形関係の発見, 統計関連学会連合大会, 2007年9月7日, 神戸大学

〔その他〕
ホームページ等
http://www.stat.math.keio.ac.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・教授
研究者番号：60089828

(2)研究分担者
該当なし

(3)連携研究者
横内大介 (Daisuke Yokouchi)
一橋大学・国際企業戦略研究科・講師
研究者番号：50407144