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　　８　日現在

機関番号：　３０１０３
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2016
課題番号：　２５３３００４０

研究課題名（和文）大規模欠測を伴う空間系列的超大量非典型データの統合的モデリング

研究課題名（英文）Statistical modeling for large scale missing in spatial and time-series data

研究代表者
中村　永友（NAKAMURA, Nagatomo）

札幌学院大学・経済学部・教授

研究者番号：７０２０７９００

交付決定額（研究期間全体）：（直接経費）　　3,800,000円

研究成果の概要（和文）：本研究は人工衛星で観測される磁気圏プラズマ速度データの統合的な分析を行うための種々の問題解決が目的である．データの形式は「離散×不等間隔×方角データ×擬頻度×非対称分布×ノイズ的データの存在×複数成分×大規模×時系列的」であり，これまでの統計科学が個別に扱ってきたデータ形式が混在した複雑なデータである．いくつかの要素を組み合わせつつ，最終目標として可能な限り統合的な解析の障害を取り除いていくことである．研究期間内には基本的な欠測データに対する時空間モデリングと，本研究の根源となる命題を提示して離散確率分布を通した連続型確率分布にしたがう乱数生成法の精密化を行った．

研究成果の概要（英文）：The purpose of this research is to solve various problems for integrated analysis of magnetospheric plasma velocity data observed by artificial satellites. The format of the data is "discrete × unequal interval × direction data × pseudo frequency × asymmetric distribution × presence of noise-like data × multiple components × large scale × time series". It is complicated data in which data statements handled separately by the statistical sciences so far are mixed. While combining several elements, it is the final goal to remove obstacles in the integrated analysis as much as possible. In the research period, we performed spatial modeling on fundamental missing data and presented a proposition that is the root of this research and refined the random number generation method according to the continuous probability distribution through the discrete probability distribution.

研究分野：　統計科学

キーワード：　混合分布モデル　時空間データ　欠測データ　超大量データ　一様乱数　疑似乱数

１．研究開始当初の背景

本研究は，人工衛星で観測される超大量の磁気圏プラズマ粒子速度データの統合的な分析を行うための種々の問題解決が目的である．そのデータの形式は「大規模な欠損データ×時系列観測×異なる観測解像度×大規模データ×離散データ×不等間隔観測×擬頻度観測×方角データ×非対称分布×ノイズ的データの存在×複数成分の混合」であり，これまでの統計学が個別に扱ってきたデータ形式が混在・混合している「非正則・非典型データ」である．これらの各要素は正則化等の変換や個別のモデリングにより従来の統計手法で分析もできるが，分析手法を単に組み合わせただけでは十分有意な情報抽出はできない．本研究は，これらの個々の問題解決を順次行い，空間系列統合モデルを作り，究極的には人工衛星搭載可能なソフトウェア開発を行うことである．

２．研究の目的

研究生活を始めてから一貫したテーマは超大量データの分析である．統計的分類モデリングによる LANDSAT 画像データ分析の研究を手始めに，近年研究対象のプラズマ粒子速度データは，昨今話題のビッグデータであり，大量データ分析の重要性が社会的に認識されてきた．これらの研究を通して，データの様相をよく観察し適切な統計モデリングをすることで有意な情報が引き出すことができ，小標本と大標本の間を埋めうる理論の確立が急務であることを実感してきた．今回の研究も根底の基本テーマである．研究の目的の記述のためにも，まず，データの背景を説明する．

人工衛星で観測される磁気圏プラズマの速度分布は３次元正規分布（Maxwell 分布）で表現され，観測性能の向上により複数のプラズマ分布が観測される事例がでてきた．地球物理学の分野では，２つ以上の成分分布が部分的に重なる複雑な分布を分離することは困難な問題であった．データは日々刻々と衛星から送信され，時系列的に数多くの観測例を扱うに至っては，この問題はより一層深刻であった．この問題に対して 2001 年に混合分布モデルによりプラズマ分布の分離に成功し（Ueno & Nakamura et al., 2001, JGR, 106, 25655-25672），当該分野に大きなインパクトを与え，現在標準手法となった．これによる現象解明の研究は連携研究者のグループによって推進されている（Annales Geophysicae, 25, 769-777 & 1417-1432 & 2069-2086 & 2229-2245; Geophysical Research Letters, 32, L06101, doi: 10.1029/2004GL021891）．一方このデータの欠損領域の存在に対してモデリングを行い（中村・上野他，2005，応用統計学，34, 57-75:優秀論文賞受賞），高い評価を得た．しかし 2001・2005 年のデータ解析は複雑な様相を呈するデータを簡素化して混合分布モデルをあてはめたものであり，時系列的な成分の動きを正確にとらえる必要性が出てきた．その後，速度０付近に大規模な欠測値の存在と，飛来するプラズマ粒子の状況に応じて観測精度（解像度）に差異があるデータが蓄積されていて，まったく分析されていないことが明らかになった．

以上の背景の下，これまでの成果としては，時系列的な動きをする分離された成分分布が本来滑らかな動きをしていることを前提とした平滑化手法の提案を行った．同時に欠測に関するデータの処理とモデリング，混合分布モデルの成分数推定を行った．いずれもこの段階ではラフなものであった．方角データに対して数式的に複雑な理論構築をしたが，これに関しては途上段階である．

更に分析を進めるために「統合的モデル構築」と「データの正則化」に関して解決すべき様々な問題が存在している．分析全体の整合性を考慮し，様々なアプローチを検討し，最も効果的かつ物理的解釈が妥当なものを見つけ出し，最終的に「空間系列モデル」として統合し，ラフなモデル（リアルタイム分析）から正確なパラメータ推定までの一連の分析の流れを確立する．このとき「現実問題の解決」と「理論」の間のバランスが重要で，これを調整しながら統合モデルを構築し，人工衛星搭載可能な「実用ソフトウェア」開発が究極の目標である．さらに，小標本と大標本（ミクロとマクロ）の間を埋める理論的考察を行ってきており，「超大量データ≠統計理論での無限大」という視点から，この間を埋める研究を行うことも目的である．

３．研究の方法

3.1 データの取り扱いの検討

データの正則化（分析しやすい形式への変換・加工）とモデリングは相補的である．対象のデータは一言では非正則・非典型データ，詳しくは大規模な欠損データ×時系列観測×異なる観測解像度×大規模データ×離散データ×不等間隔観測×擬頻度観測×方角データ×非対称分布×ノイズ的データの存在×複数成分の混合という特徴がある．個々に対しては既成統計手法の適用も可能であるが，このような多数の特徴が混在したデータの分析は大きな困難を伴うため，統計科学的に意義がある．さらに物理解釈が可能で，最終的なモデルを意識した正則化が大きな課題である．またこれはビッグデータであるが，モデル選択の観点からは複雑なモデルが選ばれることを意味し，系列データの各時点では中規模データであり，有限修正等の統計理論の構築を行う．

3.2 モデル構築

次に示すデータに内在する「分析の困難さ」を常に考慮しながらモデル構築する．(1)

データはプラズマが飛来する方向（角度）と径の長さを速度とする極座標形式で記録され，さらに極座標空間が多数のセルで分割（離散化）され，セル内のプラズマの頻度が観測される．分析する際には物理的変換を伴うので，頻度のような実数（擬頻度）となる．⑵データ空間内に非ノイズデータが存在し，現実的かつ実用的な方法（モデリング，データ正則化）の判断が必要である．⑶成分分布数の推定において複雑かつ多変量の混合分布モデルではブートストラップ情報量規準EICが有効(中村他，応用統計学，27，165-180，1998)であるが，離散×擬頻度データに対してEICは有効でないので，この問題解決が必要である．⑷大規模な欠測データが存在し，なおかつ混合分布を想定しなければいけないこと．次に系列的推移のモデリングの問題点として，⑸観測状況に応じて観測解像度に違いがあること，⑹系列的データのモデル化の多様性（混合分布モデルの系列的パラメータ推定のモデル化として確率過程，ベイズモデル等）があり，物理的解釈の容易性と実用化のバランスが必要となる．さらに，⑺出現・消失する成分分布のモデル化の困難さがある．これは「識別可能性の問題」などの困難な問題を伴うため，アルゴリズムによるアプローチで解決を図る．とくに厄介なのが大規模欠測処理のモデリングとその混合分布モデルとその成分数推定である．また離散擬頻度データに対する対数尤度のバイアス補正方法の確立とその漸近化も課題である．

3.3 中間領域の探索

小標本と大標本の間，ミクロとマクロの間，あるいは帰納と演繹の間を埋めるための研究の方法として考えたのが，連続型確率分布にしたがう乱数の利用である．連続型確率分布を離散型確率分布で近似し，その離散分布にしたがう乱数と，元の連続型確率分布にしたがう乱数の確率分布としての近さを見ることである．具体的には微小区間における前者と後者の違いは，どの段階から起こるのかを理論的にあるいは数値的に検証する．

４．研究成果

初年度に以下を行った．欠測データに対する統計的接近として，典型的なパラメトリックな統計的モデリングである値打ち切り（censored）や切断（truncated）を考慮した分析方法があるが，ある特定の領域で分析上無視できない欠測データがあり，さらにその領域では正常に測定されたデータも存在している，という問題を扱った．例えば機器の調子が悪く，ある特定の測定範囲で記録されたデータもあれば，うまく記録されなかったデータもあるといった状況である．これは部分的に一部の領域でランダムな欠測がある（partially missing at random）という視点でモデリングが可能である．このようなデータを分析するには，従来の値打ち切りや切断の統計モデルをそのまま適用できない．そこで，このような問題に対する統計モデルを示し，さらに基礎となる確率分布を正規混合分布への拡張の検討を行った．

2 年目には，1 年目の成果を踏まえて，基礎となる確率分布を正規混合分布に拡張することを行い，さらに混合分布の成分数を推定するための検討を行った．

3 年目は，大規模欠測に関する理論的考察を行った．欠測に対する処置として，データ増大法と積分による埋め込み法の比較検討を行った．その結果，確率分布が比較的単純な場合には積分による埋め込み法が有意であることが確認された．この検討にあたって離散確率分布にしたがう乱数の生成を通して，連続型確率分布にしたがう乱数生成法を提案した．大量の乱数が必要な場合は，ボックス・ミューラー法などの正規乱数生成法より，生成時間や正規性の特徴において優れた生成法であることが確認された．また欠測がある場合の混合分布の成分数推定の理論的考察を通して，潜在変数を含む統計モデルにおける効果的なブートストラップ標本を通したパラメータ推定法を提案した．

最終年は，前年度提案した離散確率分布にしたがう乱数の生成を通して連続型確率分布にしたがう乱数生成法の精密化を行った．「連続型確率分布にしたがうどんな乱数も，微小な区間においては，ある一定条件下で一様分布と見なせる」という命題とその証明のための枠組みを提示した．証明の枠組みとしては数学的，統計的，工学的の 3 つを示す必要があり，後段の 2 つは数値実験を通してある許容範囲内で一様乱数性を示すことができた．数学的証明については，確率分布の密度関数，区間幅，データ数を使って，何らかの関数的な条件を構築しているところである．また，任意の確率分布にしたがう乱数を微小区間で生成したときに，その確率分布にしたがうか否かも同時に検討し，保守的～好意的な条件を示すことが出来ると予想し，これにより工学的証明の補強が可能となる．この研究に関しては未だに最終的な結論を得ていない．全体の中間的な報告は大学の紀要論文などで公表し，最終的な研究成果は数年以内にまとめる予定である．

５．主な発表論文等

〔雑誌論文〕（計 10 件）

[1] 中村永友・土屋高宏 (2017). 正規分布の裾の確率評価と似乱数生成，札幌学院大学総合研究所紀要 （情報科学），Vol.4，1-7，2017.3.
http://hdl.handle.net/10742/00003040（査読無し）

[2] Genta UENO and Nagatomo NAKAMURA (2016). Bayesian estimation of observation error covariance matrix in ensemble-based filters, Quarterly Journal of the Royal Meteorological Society , Vol.142, Issue698, 2055-2080, 2016.6.1, DOI: 10.1002/qj.2803. **（査読有り）**

[3] **中村永友・石川千温・渡辺慎哉・小池英勝** (2016). **情報教育課題合格ログデータによる受講生の類型化**, **札幌学院大学 総合研究所紀要（情報科学）**, Vol.3, 1-6, 2016.3. http://hdl.handle.net/10742/2046**（査読無し）**

[4] **中村永友** (2016). **混合正規分布の成分数推定に関する数値的検証**, **札幌学院大学総合研究所紀要 （情報科学）**, Vol.3, 7-15, 2016.3. http://hdl.handle.net/10742/2045 **（査読無し）**

[5] **中村永友・土屋高宏** (2016). **潜在変数を含む統計モデルにおけるパラメータ推定法**, **札幌学院大学 総合研究所紀要（情報科学）**, Vol.3, 17-22, 2016.3. http://hdl.handle.net/10742/2056**（査読無し）**

[6] **中村永友・石川千温・渡辺慎哉・小池英勝** (2015). **ロジスティック回帰による課題提出ログデータの解析**, **札幌学院大学 総合研究所紀要 （情報科学）**, Vol.2, 1-6, 2015.3. http://hdl.handle.net/10742/1884 **（査読無し）**

[7] Nagatomo NAKAMURA (2015). Pseudo-normal random number generation via the Eulerian numbers, Josai Mathematical Monographs**,** Vol.8, 85-95, 2015.3. **（査読有り）**

[8]**土屋高宏・中村永友** (2014). **層別された平均系列データに対する線形回帰モデル**(A linear regression model for stratified data), **行動計量学,** Vol.41(1), 3-15, 2014.3. **（査読有り）**

[9] **中村永友・花岡重利・土屋高宏** (2014). **統計学テキストにおける「分散」の表記について**,**札幌学院大学 総合研究所紀要 （情報科学）**, Vol.1, 1-10, 2014.3. http://hdl.handle.net/10742/1875**（査読無し）**

[10] Genta UENO and Nagatomo NAKAMURA (2014). Iterative algorithm for maximum likelihood estimation of observation error covariance matrix for ensemble-based filters, Quarterly Journal of the Royal Meteorological Society, Vol.140, 295-315, 2014.1. DOI: 10.1002/qj.2134. **（査読有り）**

〔学会発表〕（計 12 件）

[1] **中村永友・土屋高宏** (2016). **疑似乱数における局所一様性に関する統計的性質**, **日本計算機統計学会 第 30 回シンポジウム**, **プラサ ヴェルデ**, **沼津市**, 2016.11.24-25.

[2] **下田 妙子・斎藤さな恵・吉村香子・南純一・柳澤直武・清水金忠・中村永友** (2016). **ビフィズス菌 BB536 株の摂取による貧血改善に関する単施設ランダム化二重盲検プラセボ対照平行群間比較試験**, **第 19 回 日本病態栄養学会 年次学術集会**, **パシフィコ横浜**, **横浜市**, 2016.1.9-10.

[3] **中村永友・土屋高宏** (2015). **離散型確率分布を通した連続型確率分布にしたがう乱数の生成**, **日本計算機統計学会第 29 回シンポジウム**, **まなぼっと幣舞**, **釧路市**, 2015.11.27-28.

[4] **中村永友**(2015). **オープンデータを活用した政策提言とそのための基礎システムの構築**, **江別市大学連携調査研究事業補助金採択事業報告会**, **江別市**, 2015.07.07.

[5] **松井祐介・島村徹平・水田正弘・中村永友** (2015). **関数データクラスタリングにおける部分区間の同定法と学習ログデータへの適用**, **日本分類学会**, **第 33 回大会**, **帝京大学**, **東京**, 2015.03.02-03.

[6] **中 村 永 友** (2014). Pseudo-normal random number generation via the Eulerian numbers, **城西大学理学研究科数学専攻ワークショップ「統計科学とその周辺」**, **城西大学**, **東京**, 2014.12.7.

[7] **中村永友・土屋高宏・小西貞則** (2014). **潜在変数を含む統計モデルにおけるブートストラップ分散減少法**, 2014 **年度統計関連学会連合大会（日本統計学会第 83 回大会**, **応用統計学会年次大会**, **日本計量生物学会年次大会）**, **東京大学**, **東京**, 2014.09.14-16.

[8] **土屋高宏・中村永友**(2014). **ソーティング過程に現れる離散確率分布とその精密化**, 2014 **年度統計関連学会連合大会 （日本統計学会第 83 回大会**, **応用統計学会年次大会**, **日本計量生物学会年次大会）**, **東京大学**, **東京**, 2014.09.14-16.

[9] **中村永友・石川千温・渡辺慎哉** (2014). **課題完成時間からの習熟度別分布の推定**, 2014 PC **カンファレンス**, CIEC **（コンピュ**

ータ利用教育学会）研究大会，札幌学院大学，北海道江別市，2014.8.8-10.

[10] 中村永友・土屋高宏・上野玄太（2013）. 一部の観測領域でランダムな欠測のあるデータへの混合分布モデルの適用，科研費シンポジウム，『一般化線形モデルの最新の展開とその周辺』，(科学研究費・基盤（A）によるシンポジウム，研究代表者：谷口正信，「非対称・非線形統計理論と経済・生体科学への応用」，課題番号：23244011），千葉大学，千葉市，2013.11.08-10.

[11] 中村永友・土屋高宏・上野玄太（2013）. 一部の観測領域でランダムな欠測のあるデータに対する混合分布モデルのあてはめ，2013 年度統計関連学会連合大会 （日本統計学会第 82 回大会，応用統計学会年次大会，日本計量生物学会年次大会），大阪大学，豊中市，2013.09.08-11.

[12] 中村永友・花岡重利・土屋高宏（2013）. 統計学テキストにおける分散の記号に関する調査報告，2013 年度統計関連学会連合大会 （日本統計学会第 82 回大会，応用統計学会年次大会，日本計量生物学会年次大会），大阪大学，豊中市，2013.09.08-11.

〔図書〕（計　0　件）

〔産業財産権〕

○出願状況（計　0　件）

○取得状況（計　0　件）

６．研究組織

(1)研究代表者
中村 永友（NAKAMURA, Nagatomo）
札幌学院大学・経済学部・教授
研究者番号： 70207900

(2)研究分担者
なし

(3)連携研究者
土屋 高宏（TSUCHIYA, Takahiro）
城西大学・理学部・准教授
研究者番号： 60316677

上野 玄太（UENO, Genta）
情報・システム研究機構 統計数理研究所・モデリング研究系・准教授
研究者番号： 40370093

(4)研究協力者
なし