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

令和　　元 年　　５ 月 ２１ 日現在

機関番号：１５３０１
研究種目：若手研究(B)
研究期間：2016～2018
課題番号：１６Ｋ１６０１９

研究課題名（和文）大規模時空間データのホットスポット検出に関する研究

研究課題名（英文）Study on hotspot detection for large-scale spatiotemporal data

研究代表者
石岡　文生（Ishioka, Fumio）

岡山大学・環境生命科学研究科・准教授

研究者番号：２０５１０７７０

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：どこでデータが観測されたかという「位置」情報が付与されたデータ（空間データ）に対し、有意に高い場所の集合（ホットスポット）を検出するための新たな手法を確立した。本研究代表者は、空間データを位相的な階層構造で表現する「エシェロン解析法」を応用したホットスポット検出のためのプログラムを独自に開発しているが、本研究期間においてはこれをさらに発展させ、大規模空間データ、多次元空間データなどを対象にする新たな手法を提案し、各種の実データへの応用、ならびに数値実験によりその有効性を示した。加えて、これら一連の解析を行うためのソフトウェアの開発も進めた。

研究成果の学術的意義や社会的意義
本研究では、「エシェロンスキャンに基づくホットスポット検出法」をさらに改良し、大規模空間データや多次元空間データへの適用を可能にするための一手法を提案した。これを実現させることには、近年のビッグデータ解析で力を発揮することが期待される。今後は、従来の地理情報データへの適用のみに限らず、ゲノムデータなどといった他分野への応用も可能と考える。加えて、本研究ではこれら一連の解析を行うためのソフトウェア開発にも力を注いでおり、そのソフトウェア公開による波及効果は極めて高いものと考える。

研究成果の概要（英文）：In this study, I have established a new method for detecting spatial clusters with significantly high or low risk, called a hotspot. I have originally developed a program for hotspot detection by using “echelon analysis” that represents spatial data in a topological hierarchical structure. By further extending this method, I proposed a new method of hotspot detection to cover large-scale data or multi-dimensional data, and showed its effectiveness by application to some real data and numerical experiments. In addition, I developed software for detecting hotspot based on the echelon analysis.

研究分野：空間統計学、計算機統計学

キーワード：空間スキャン統計量　エシェロン解析　空間集積性　ホットスポット　クラスター検出　shinyアプリ　ZDD

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

位置情報などの空間的な情報が付与された観測データ（空間データ）において、有意に高い、または低いある特定の領域を「ホットスポット」と呼ぶ（クラスター、集積性などと同義）。例えば、「インフルエンザのような感染症の患者数」や「ダイオキシンのような有害物質の測定値」等の空間的な分布に対して、「どこかにホットスポットはあるのか？それとも全体的にばらついているのか？」「ホットスポットが存在しているとしたら、どの範囲までがそうなのか？」といった事を統計的根拠に基づいて決定することは、環境状況の把握や、将来の環境や健康への影響を早期に発見するためにも大変重要である。近年、コンピュータ技術やインターネットの発達などに伴い、ビッグデータのような膨大な量のデータを個人単位でも取り扱えるようになっており、その解析手法について盛んに議論されている。しかし、大規模空間データに対するホットスポット検出手法については、まだ十分に確立されていないというのが現状である。

２．研究の目的

先行研究等で提案されている各種のホットスポット検出手法は、その検出されるホットスポットの形状に制約がかけられていたり、数千～数百万領域にも及ぶ大規模データを対象にする場合は、計算コストの面から解析が困難とされてきた。このような問題に対処するために、本研究代表者は、エシェロン解析法[1]を利用したホットスポット検出（エシェロンスキャン法[2]）について研究を行っている。本研究の目的は、エシェロンスキャン法をベースにすることにより、近年様々な分野で蓄積され続けている「ビッグデータ」のような、さらなる大規模データを解析対象として扱うための新たなアルゴリズムを構築することにある。また、地理的な二次元空間に、時間・高度などの軸を想定した「多次元空間データ」に対するホットスポット検出法についても検討する。さらには、これら一連の解析を実行するソフトの開発も目標にする。

３．研究の方法

本研究代表者は、空間データを位相的な階層構造で表現する「エシェロンデンドログラム」の上位階層から順に領域を探索していき、ホットスポットを同定するエシェロンスキャン法のプログラムを世界で初めて構築した。このアルゴリズムは、(Step1)空間データをエシェロン解析によって複数の階層に分割し、(Step2)その上位に位置する階層に含まれる領域を順に走査（スキャン）しながら、統計指標の一つである「尤度」が高くなる様な領域の集合を探索していき、(Step3)最大尤度となったところまでの集合領域をホットスポットと同定する（図１）。

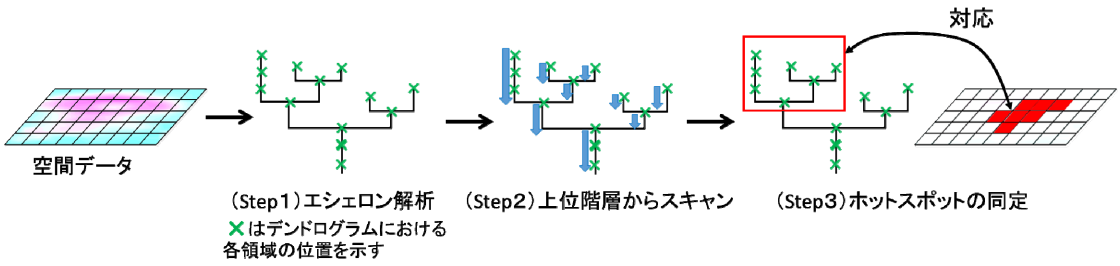

図１．エシェロンスキャン法の概要図

エシェロンスキャン法は、尤度が高くなるような領域をあらかじめデンドログラムの上位階層に配置させておき、デンドログラムの形状をそのままスキャンの道筋として定義することにより、従来法に比べてスキャンする領域の数を大幅に軽減、かつ、高尤度のホットスポットを検出できるよう工夫がなされている。しかしながら、数十～数百万領域にも及ぶような大規模データを扱う場合は、デンドログラムの形状は大変複雑になり、スキャン行程に膨大な計算時間を要する。そこで、複雑なデンドログラムに対してカットオフを行うことで単純化し、その単純化された構造に基づいてスキャンする方法を新たに提案する（図２）。

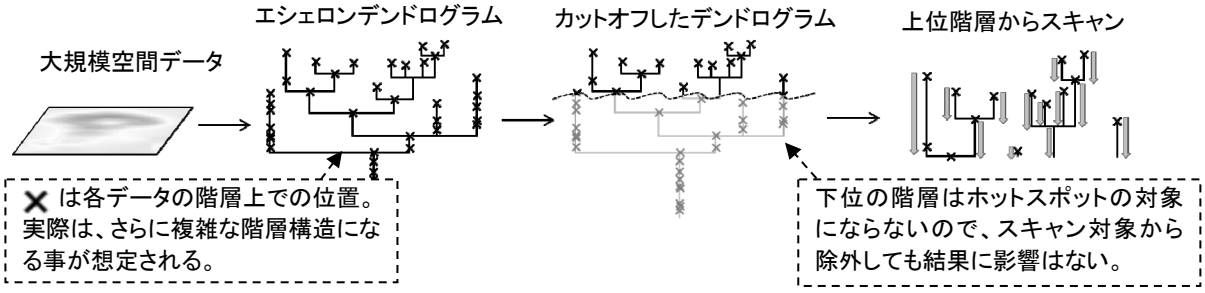

図２．大規模空間データの新しいスキャンアルゴリズムのイメージ図

加えて、従来のホットスポット検出法は、XY 平面で表される地理的な二次元空間を対象にしたものが圧倒的に多い。しかしながら、エシェロン解析は、高度などを加えた「三次元空間」や、時間軸を加えた「時空間」といった、三次元以上の構造であっても、各領域に binary な連結情報を与えさえすれば二次元上のデンドログラムで表現でき、エシェロンスキャンを実行できる。空間・時間の両方を同時に捉えたデンドログラムを描き、エシェロンスキャンを適用することにより、空間だけではなく、ホットスポットの時系列的な推移も同時に表現することが可能であると考える。

エシェロンスキャン法の有効性を世に広め社会に還元するためには、そのソフトウェア化の実現が不可欠である。ソフトウェア開発において、ホットスポットの検出結果を地図上で拡大や縮小をしながら確認できたり、空間的な階層構造のグラフ（エシェロンデンドログラム）の効果的な視覚表現、エシェロン解析を行う際の「属性値」「連結情報」を複数のタイプからマウスで選択できる、といったユーザーが扱い易いインタラクティブに操作可能な GUI の構築にも重点を置く。これらを実現するために、開発環境として統計ソフトＲ上で web アプリケーションを作成できる shiny パッケージを利用する。

４．研究成果

（大規模空間データの新しいスキャンアルゴリズム）

ここでは、Tango(2008)が提案した制限付スキャン統計量のアイデアを利用することによる「カットオフされたエシェロンデンドログラム」に基づきホットスポットを検出した例を示す。図３（左）は、1969 年から 1971 年のアメリカの 3085 郡の殺人発生件数と人口を基に、殺人発生の相対危険度を地図上で表現している。

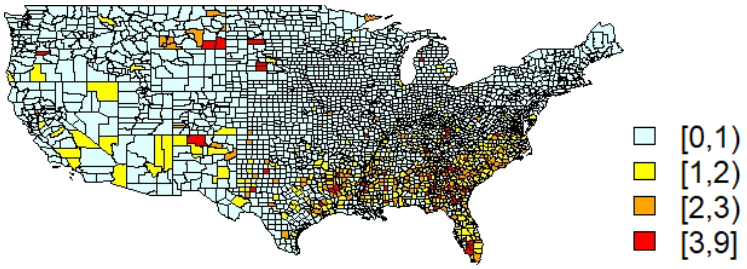

図３.アメリカ郡別殺人発生の相対危険度

このデータに対し、通常のエシェロンスキャン法を適用すると、図４（左）のデンドログラムとともに、図５（左）としてホットスポットが検出される。一方で、提案手法では図４（右）のようにデンドログラムが作成され、そのホットスポットは図５（右）のように与えられる。

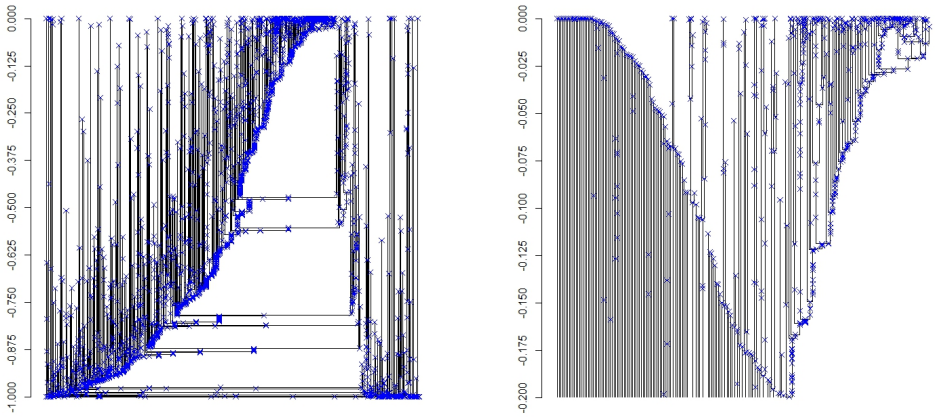

図４.エシェロンデンドログラム（左）、カットオフ後のエシェロンデンドログラム（右）

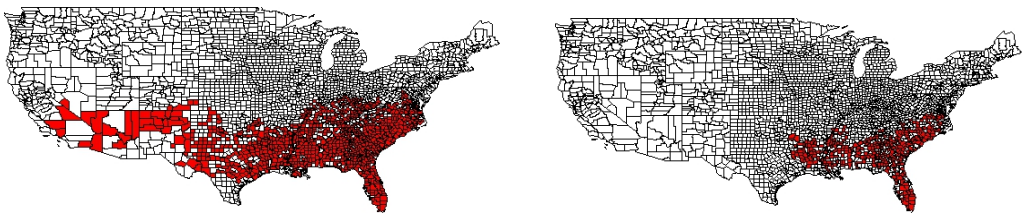

図５.エシェロンスキャン法によるホットスポット(左)、提案手法によるホットスポット(右)

このとき、従来のエシェロンスキャン法によるホットスポットの方が高い尤度を示したが、ホットスポットとして同定された領域の中には、相対危険度の低い（つまりは領域単位でみたときホットスポットとしては相応しくない）領域が多数含まれている。一方で、提案手法は相対危険度の低い領域は取り込まないようデンドログラムに工夫がなされている。さらに、予め様々な形状のホットスポットを設定したシミュレーションデータを生成し、想定したホットスポットがどれだけ正確に検出できているかを検証する数値実験を行った結果、提案手法の検出力は良好な値を示し、その計算コストについても従来の手法と比べ約 3 分の 1 から 5 分の 1 程度まで短縮された。今後の課題として、バラツキを考慮したシミュレーションデータによる数値実験、別の統計モデルに基づいた尤度によるホットスポット検出、複数のホットスポットの存在が想定される場合など、より実データに即した状況下での検出精度について検証を行う必要がある。これら成果の一部を、国内学会（日本計算機統計学会）などで発表した。

（多次元空間データに対するホットスポット検出）

二次元空間上に高度の情報を加えた「三次元空間」や、時間軸を加えた「時空間」のような多次元空間データに対してエシェロンスキャン法を応用することで、ホットスポットを多角的に捉え表現する方法について研究を進めた。その例として、ここでは日本の 1988 年から 2007 年までの二次医療圏における男性自殺者数のデータに応用したものを紹介する。各領域・期間にまたがる時系列的な連結情報に加え、年齢階級に対しても連結情報を与えることにより、領域（2 次元）×期間×年齢の 4 次元空間データに対してエシェロンスキャン法を適用し、そのホットスポットを表現した（図６）。その他の応用例として，福島第一原子力発電所事故に起因する空間線量率データに適用した際には、時間の経過とともにホットスポット領域が縮小していく様子が見て取れた。これら成果の一部を国際学会（日中国際会議，日独分類会議，COMPSTAT2016,COMPSTAT2018）や国内の研究集会などで発表した。

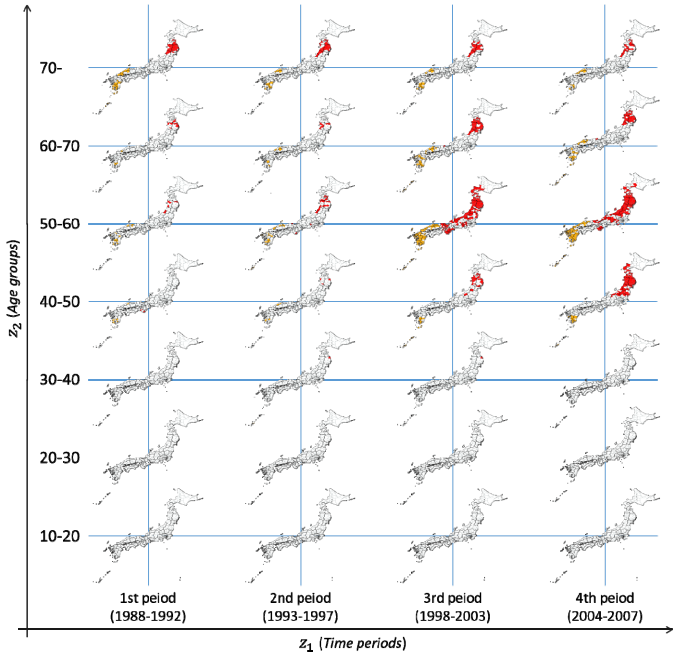

図６．領域×期間×年齢で表現した男性自殺者数のホットスポット

（ZDD 技法を応用した新しいホットスポット検出手法の提案）

組み合わせ集合を高速に数え上げ列挙する ZDD(Zero-suppressed Binary Decision Diagram)と呼ばれるアルゴリズム[3]を、ホットスポットとなりうる全ての隣接ブロック領域の探索に応用することにより、真に最大尤度を有するホットスポットを現実的な計算コストで検出しうることを見い出した。実際に、アメリカノースカロライナ州の乳幼児突然死症候群のデータに対して適用した際に、これまでの先行研究では検出できないような形状かつ高尤度となるホットスポットを得た（図７）。この新規性・有効性をまとめたものを雑誌論文（JJSD）に発表し、その成果は日本計算機統計学会において 2018 年度学会賞（論文賞）を受賞した。

図７.ZDD 技法を利用したホットスポット検出例

さらに、従来の地理的な連結情報を用いず、エシェロンデンドログラムの階層内に付置された領域間に連結を与えることにより、「エシェロン法による計算コストの軽減化」と「ZDD による総当たりスキャン」を組み合わせた新たなホットスポット検出法を提案した。これら成果の一部を国際学会（IFCS2017, IASC-ARS2017）および国内の学会・研究集会などで発表した。

（エシェロンスキャン法によるホットスポット検出を行うためのソフトウェア開発）
本研究課題を世に広め社会に還元するためには、そのシステムを構築し公開することが不可欠である。本研究代表者は、これまで実際にエシェロンスキャン法を利用したいとの要望を国内外で受けたという経緯がある。本報告時点において、システムはほぼ実装できており（図８）、公開用のサーバーの構築のための準備を行う段階にある。R-shiny をはじめとした R に関連する GIS パッケージ等を利用して開発しているため、当初の予定通りユーザーが利用し易いインタラクティブに操作可能な GUI の構築を備えたものになっている。これら成果の一部を国際学会（ECDA2018）および国内の研究集会などで発表した。

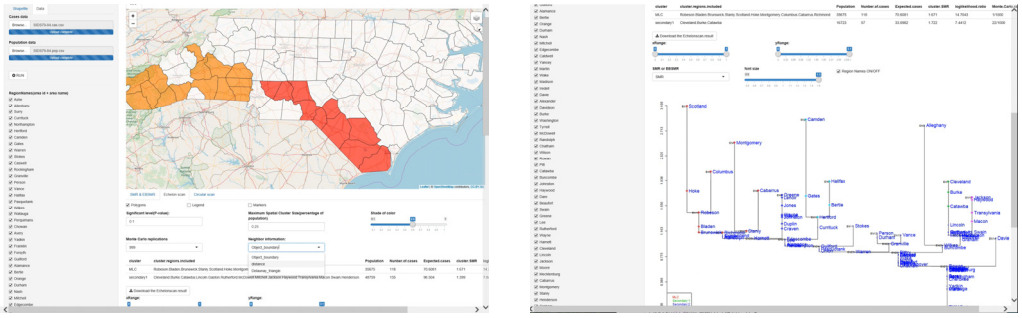

図８. 開発中のエシェロンスキャン法のアプリケーション

＜引用文献＞

[1] Myers, W. L., Patil, G. P. & Joly, K. (1997). Echelon approach to areas of concern in synoptic regional monitoring. Environmental and Ecological Statistics, 4(2), 131-152.

[2] 栗原考次. (2003). 階層的空間構造を利用したホットスポット検出. 計算機統計学, 15(2), 171-183.

[3] Minato, S. (1993). Zero-suppressed BDDs for set manipulation in combinatorial problems. Proceedings of the 30th ACM/IEEE Design Automation Conference, 272-277.

５．主な発表論文等

〔雑誌論文〕（計　3 件）

[1] Ishioka, F., Kawahara, J., Mizuta, M., Minato, S. & Kurihara, K. (2019). Evaluation of hotspot cluster detection using spatial scan statistic based on exact counting. Japanese Journal of Statistics and Data Science, 2(1), 241-262.査読有.
DOI: 10.1007/s42081-018-0030-6

[2] Minato, S., Kawahara, J., Ishioka, F., Mizuta, M. & Kurihara, K. (2019). A Fast Algorithm for Combinatorial Hotspot Mining Based on Spatial Scan Statistic. In Proceedings of the SIAM International Conference on Data Mining (SDM 2019), 91-99. 査読有. DOI: 10.1137/1.9781611975673.11

[3] Ishioka, F. & Kurihara, K. (2016). Detection of space-time clusters for radiation data using spatial interpolation and scan statistics. In Proceedings of the 22nd International Conference on Computational Statistics (COMPSTAT2016), 85-97. 査読有.

〔学会発表〕（計 23 件）

[1] 石岡文生, 川原純, 水田正弘, 湊真一, 栗原考次. Exact counting に基づいたホットスポットクラスターの検出について. 北海道大学情報基盤センター萌芽型共同研究「大規模・複雑化データに対する解析手法の多面的研究」. 2019.

[2] 梶西将司, 石岡文生, 栗原考次. Echelon 構造を利用した空間複雑性の評価. 科研費シンポジウム「空間データと災害の統計モデル」. 2019.

[3] 竹村祐亮, 石岡文生, 栗原考次. 制限付エシェロンスキャン法によるクラスター検出手法の提案. 日本計算機統計学会第 32 回シンポジウム. 2018.

[4] Ishioka, F. & Kurihara, K. Cluster detection for multi-dimensional spatial data based on hierarchical structure. The 23rd International Conference on Computational Statistic (COMPSTAT2018). 2018.

[5] Ishioka, F., Kajinishi, S. & Kurihara, K. Visualization of cluster detection based on hierarchical structure for geospatial data and its application. European Conference on Data Analysis (ECDA2018). 2018.

[6] Kurihara, K. & Ishioka, F. Echelon clustering and its applications. The 7th German-Japanese Symposium on Classification. 2018.

[7] 湊真一，川原純，水田正弘，石岡文生，栗原考次．スキャン統計量に基づく組合せホットスポット抽出を行う高速アルゴリズム．第 169 回アルゴリズム研究発表会．2018.

[8] 竹村祐亮，石岡文生，栗原考次．制限付スキャン統計量に基づく空間集積性の新たな検出手法について．日本計算機統計学会第 32 回大会．2018.

[9] 梶西将司，石岡文生，栗原考次．R-Shiny による空間データの集積性検出．北海道大学情報基盤センター萌芽型共同研究「大規模データ科学に関する多面的研究」．2018.

[10] Ishioka, F., Kawahara, J. & Kurihara, K. Evaluation of spatial cluster detection method based on all geographical linkage patterns. The Conference of the Asian Regional Section of the International Association of Statistical Computing (IASC-ARS2017). 2017.

[11] 石岡文生，梶西将司．R を用いた空間データの構造分析と集積性の検出．統計数理研究所共同研究集会「データ解析環境 R の整備と利用」．2017.

[12] Ishioka, F. & Kurihara, K. Visualization for radiation monitoring post data using spatial interpolation. Hangzhou International Statistical Symposium. 2017.

[13] 石岡文生，川原純，栗原考次．空間データに対するエシェロン解析の新たな展開について．統計数理研究所共同研究集会「環境・生態データと統計解析」．2017.

[14] Ishioka, F. & Kurihara, K. Cluster detection using spatial scan statistic and its new development in large-scale scanning. Conference of the International Federation of Classification Societies (IFCS2017). 2017.

[15] Kurihara, K., Kajinishi, S. & Ishioka, F. Statistical evaluation for spatial complexity based on echelon trees. Conference of the International Federation of Classification Societies (IFCS2017). 2017.

[16] 石岡文生．エシェロン解析を利用した空間データ分析．第 8 回多摩データサイエンス研究会．2017.

[17] 石岡文生，栗原考次．ZDD によるグラフ列挙技法を利用したホットスポット検出と空間スキャン法の性質評価．第 38 回大規模データ科学に関する研究会「複雑データ解析法に関する研究」．2017.

[18] 梶西将司，石岡文生，栗原考次．空間疫学における集積性の検出と shiny への実装．日本計算機統計学会第 30 回シンポジウム．2016.

[19] Ishioka, F. & Kurihara, K. Detection of space-time clusters for radiation monitoring post data based on echelon scan statistics. 2016 International Conference for JSCS 30th Anniversary in Seattle. 2016.

[20] 石岡文生，栗原考次．放射線モニタリングデータの時空間集積性に対する空間補間法の応用．統計数理研究所共同研究集会「環境・生態データと統計解析」．2016.

[21] 石岡文生，栗原考次．空間スキャン統計量による集積性検出の新たなアプローチ．第 35 回大規模データ科学に関する研究会「大規模医療情報に対する高度統計技法の開発と可視化」．2016.

[22] 石岡文生，栗原考次，水田正弘．空間データに対するホットスポット検出手法の性質評価について．2016 年度統計関連学会連合大会．2016.

[23] 石岡文生，栗原考次．空間情報を利用したホットスポットの検出について．基盤(S) 離散構造処理系プロジェクト 「2016 年度 初夏のワークショップ」．2016.

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。