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　１３　日現在

機関番号：　６２６０３
研究種目：　基盤研究(C)（一般）
研究期間：　2012～2016
課題番号：　２４５００３５６

研究課題名（和文）時空間構造を持ったスキャン統計量の同時確率計算の実用化

研究課題名（英文）Computation of a joint distribution for scan statistics with temporal-spatial structures

研究代表者

栗木　哲（Kuriki, Satoshi）

統計数理研究所・数理・推論研究系・教授

研究者番号：９０１９５５４５

交付決定額（研究期間全体）：（直接経費）　　4,000,000円

研究成果の概要（和文）：時間的，空間的相関構造を持った確率変数の組からマルコフ性を抽出することによって，時空間スキャン統計量の多重性調整p値の高速計算法を確立した．また空間疫学の疾病集積クラスターの同定問題へ応用した．いくつかのデータ解析を通して手法の有効性を確認した．提案手法の考え方は，(i) 空間情報，すなわち地理上の隣接関係を点と無向辺からなるグラフで記述し，(ii) そのコーダル化によってマルコフ性を抽出し，(iii) マルコフ構造を用いた階層的逐次数値積分を用いて多重性調整p値を計算する，というものである．また2016年10月29日（土）統計数理研究所にて研究集会『応用統計学のひろがり』を主催した．

研究成果の概要（英文）：By extracting the Markov structure from a set of random variables with temporal and spatial structures, we have established a fast computation method to evaluate the multiplicity-adjusted p-value of temporal-spatial scan statistics. In addition, we applied the proposed method to spatial epidemiology. We demonstrated the usefulness of the proposed method by analyzing several data. The basic idea is (i) to describe the geographic information (adjacent relationship) by a non-directed graph, and (ii) to extract the Markov structure via the chordal extension of the graph, and (iii) to evaluate the multiplicity-adjusted p-value by successive numerical integrations based on the Markov property. On Saturday 29th October 2016, at the Institute of Statistical Mathematics (Tachikawa, Tokyo), we organized the workshop "New Horizons of Applied Statistics".

研究分野：　統計科学

キーワード：　空間疫学　多重性調整　マルコフ性　コーダルグラフ

１．研究開始当初の背景

本研究の動機の一つは，空間疫学のホットススポット検出問題である．空間疫学においては，疾病や事故の地点・時点毎の件数データ（サーベイランスデータ）を観測し，「ホットスポット」の候補である地域・時点の組み合わせに対して，その観測値が有意に大きいかどうかを判定する．これはいわゆるスキャン統計量に対する典型的な多重比較・多重検定問題である．空間疫学では，その有意性判定はモンテカルロシミュレーションで行われている．ホットスポット検出で重要なのは非常に稀な事象の検出である．これは p 値が非常に小さい場合，つまり最大値分布の裾の部分に対応する．一般に裾領域を乱数シミュレーションで精度良く推定することは大量の計算時間をかけても難しく，この精度の向上が空間疫学における一つの重要な課題となっていた．

２．研究の目的

この種の相関を持った多次元スキャン統計量の同時分布の計算は，相関構造が時間差によって引き起こされる場合には，いろいろな文脈で過去に研究されてきた．それらの問題において最大値分布（同時分布）を求めるための一つの指導原理は，マルコフ性の抽出とその利用である．本研究では，地理情報から抽出したマルコフ性を利用し，多数の時空間スキャン統計量の同時分布の高速計算法の確立を目的とする．特に空間疫学の疾病集積クラスター (ホットスポット) 検出問題で懸案となっている多重性調整 p 値と検出力の計算の実用化を最終的な目標とした．

３．研究の方法

空間的位置関係を点と無向辺からなるグラフで表現する．そのグラフの「コーダル化」を通してマルコフ性を抽出し，それを用いて p 値の計算量を大幅に低減する．ここで用いる概念は，グラフィカルモデルの統計推測の分野ではポピュラーなものであり，得られるマルコフ性はジャンクション木とよばれる．基本的にはこの方針でアルゴリズムの構築とインプリメント，それらの性能評価ならびに実データへの適用を行う．

４．研究成果

本研究では，空間的相関構造からマルコフ性を抽出することにより，多数の時空間スキャン統計量の同時分布の高速計算法を確立した．特に項目１で述べた，空間疫学の疾病集積クラスター検出問題を念頭におき，多重性調整 p 値の正確計算法を開発した．

(i) 基本的な考え方は項目 3 で述べた方法，すなわちスキャン統計量の空間情報（地理上の隣接関係）から定義されるグラフのコーダル化によってマルコフ性を抽出し，階層的逐次数値積分を用いて多重性調整 p 値を数値計算するものである．この考えに基づき，多重性調整 p 値計算のためのアルゴリズムを構築した．その際に，与えられたコーダルグラフからそのランニングインターセクション性 (RIP) をもつクリーク列（完全列）を生成する新しい方法を見いだした．

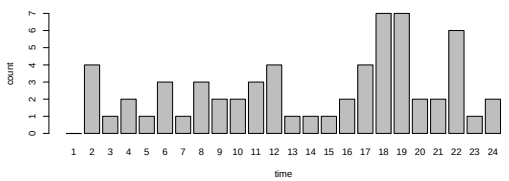
(ii) アルゴリズムの計算量と記憶容量を評価した．インシデント数の総数を N，コーダル化したグラフの i 番目のクリークの頂点の数を c(i)，i 番目のクリークを親とする頂点の数を pa(i) とするとき，計算量は N の max(c(i)+pa(i))乗に比例するものであった．また，提案アルゴリズムは逐次型であり，メモリは随時解放され再利用されるため，記憶容量の制限は問題にならない．

(iii) 上述のアルゴリズムを実現する C プログラムを作成した．作成したプログラムにおいては max(c(i)+pa(i))=6 が実際的な計算時間で計算できる最大の値であった．

(iv) コーダル化は一意ではないので，pa(i)が大きくならないようなコーダルグラフ拡張アルゴリズムを工夫した．さらに c(i)が大きくならないような新たなスキャンウィンドウを構成した．

(v)空間疫学のスキャン統計量は，ポアソンサンプリングの条件付分布としての多項分布に従う．この分布だけではなく，観測値がガウス分布の場合の条件付分布である縮退ガウス分布，ガンマ分布の条件付分布であるディリクレ分布，2 多変数超幾何分布，ディリクレ多項分布でも本手法は有効であることを確認した．

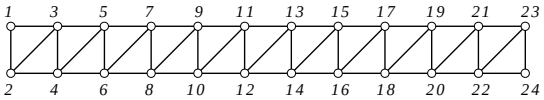
(vi) Wallenstein (1980) の流産件数データを解析し，Naus の結果 (Wallenstein and Naus, 1973; Neff and Naus, 1980) との比較を行った．下図は 1975 年 7 月から 1977 年 6 月の 24 ヶ月の間に，ニューヨークの 3 つの病院で自然流産におけるトリソミー件数である．全件数は N = 62 である．Wallenstein (1980) は，連続する 2 ヶ月での最大頻度 として時点 18,19 の 14 件に着目し， Naus の方法に基づき，その条件付分布での多重性調整 p 値を 0.038 と評価した．



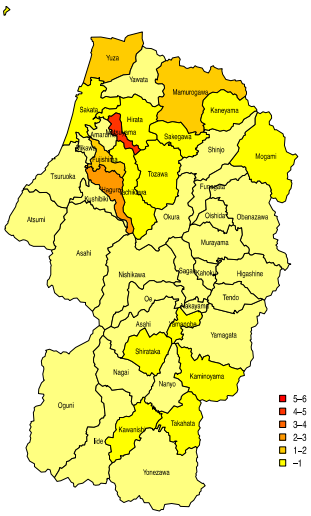
我々は Kulldorff (1997) のスキャン統計量を用いて，クラスター候補に対する統計量と多重性調整 p 値を計算した．以下の表では，大きな値をとる統計量（上位 5 つ） と，その値の，他の連続する L 時点の統計量との比較における多重性調整 p 値を与える．時点 17，18，19 が有意なクラスターをなしていると判断することができる．

statistic	period	L = 5	L = 4	L = 3	L = 2
5.954	17,18,19	0.0175	0.0151	0.0135	NA
5.847	18,19	0.0217	0.0194	0.0180	0.0140
5.143	18,19,20,21,22	0.0453	NA	NA	NA
4.507	17,18,19,20	0.0716	0.0695	NA	NA
4.507	16,17,18,19	0.0716	0.0695	NA	NA

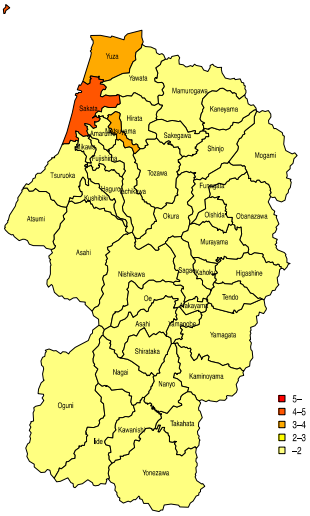
なお本データのマルコフ構造を表すグラフは以下のものである．



(vii) 1996-2000 年，山形県市町村別の胆嚢ガン件数データ（丹後・高橋・横山，2007; Tango，2010）を解析した．下図は SMR（standardized mortality ratio）の値が大きな地域を図示したものである．



また次の図は Kulldorff 統計量の値が大きい値をとる地域を図示したものである．統計量はサンプル数も考慮されるため， SMR とは結果が異なっている．



p 値の計算においては，2 種類のウィンドウを仮定した．一つは，各市町村単独をスキャンウィンドウとするもの（solo）で，市町村数 N=44．もう一方は，各市町村単独と隣接する 2 市町村のペアをスキャンするもの（solo+pair）で，ウィンドウ数は 154．（ここではそれをランダムに約半分 76+78 に分け，それぞれの p 値を計算し，それを足し合わせた）．結果は下図の通りであり，酒田，遊佐の組合せに大きな有意性が観測されている．

statistic	window	solo+pair	solo
7.651	{酒田, 遊佐}	0.00953	–
4.578	{酒田}	0.1847	0.0433
4.356	{酒田, 平田}	0.2247	–
4.247	{酒田, 三川}	*	–
3.924	{酒田, 余目}	*	–
3.570	{酒田, 八幡}	*	–
3.364	{松山}	*	0.1739
3.205	{藤島, 羽黒}	0.6458	–
3.071	{遊佐}	*	0.2065

(viii) 2016 年 10 月 29 日（土）統計数理研究所にて研究集会『応用統計学のひろがり』を主催した．講演数は 11 件であった．

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 1 件）

(1) Kuriki, Satoshi; Takahashi, Kunihiko; Hara, Hisayuki, "Recursive computation for evaluating the exact p-values of temporal and spatial scan statistics", arXiv, 査読無, arXiv:1511.00108 [stat.CO], 2015.

〔学会発表〕（計 3 件）

(1) 栗木哲，「地理情報のマルコフ化による空間疫学スキャン統計量の p 値の正確計算」，2012 年度統計関連学会連合大会，2012 年 9

月 12 日, 北海道大学.

(2) S. Kuriki, K. Takahashi, H. Hara, "Exact calculation of multiplicity-adjusted p-values of scan statistics in spatial epidemiology", 7th International Conference of the ERCIM WG on Computational and Methodological Statistics (ERCIM 2014), 2014 年 12 月 7 日, Pisa, Italy.

(3) Kuriki, Satoshi; Takahashi, Kunihiko; Hara, Hisayuki, "Recursive computation for evaluating the exact p-values of temporal and spatial scan statistics", 2015 IMS China (招待講演), 2015 年 7 月 1 日, Kunming, China.

６．研究組織

(1)研究代表者

栗木 哲（KURIKI, Satoshi）
統計数理研究所・数理・推論研究系・教授
研究者番号：90195545

(2)連携研究者

高橋　邦彦（TAKAHASHI, Kunihiko）
名古屋大学大学院・医学系研究科・准教授
研究者番号：　50323259

原　尚幸（HARA, Hisayuki）
同志社大学・文化情報学部・准教授
研究者番号：　40312988