

多項空間スキャン統計量を用いた 新型コロナウイルス型別構成の空間クラスター検出

－ 全球スケールでの流行推移と遺伝子変異の検討 －

松村玲央*・中谷友樹**

Detection of Spatial Clusters of Virus Clade Composition of the 2019 Novel Coronavirus Using Spatial Scan Statistics for Multinomial Data: Epidemic Trends and Genetic Variations at a Global Scale

Reo Matsumura* and Tomoki Nakaya**

Abstract: The pandemic of coronavirus disease 2019 (COVID-19) has made a huge impact on an enormous number of people in the world. Detecting spatial clusters of COVID-19 would provide critical information for better allocation of resources and assuring effective decision-making of public health actions as the pandemic continues to grow. Spatial scan statistics for multinomial data is a variant of spatial scan statistics, which is designed for detecting geographical clusters in terms of composition of multinomial data. In this article, we applied this method to the data of 7 genetic clades of the coronavirus by country/region obtained from Global Initiative on Sharing All Influenza Data (GISAID) and revealed the geographic patterns of the different composition of the virus genetic clades at the global scale by mainly focusing on the early period of the pandemic from January to April 2020.

Keywords: 空間疫学 (spatial epidemiology), 新型コロナウイルス感染症 (COVID-19), 多項空間スキャン統計量 (spatial scan statistics for multinomial data)

1. はじめに

近年, 新型コロナウイルス感染症 (COVID-19) の世界的な流行に対し, 世界各地の国・地域での時空間的な感染者発生状況が調査されている (Desjardins et al., 2020; Mo et al., 2020) .

こうした時空間的な疫学的解析 (空間疫学) では空間データの分布特性を把握するために, 疾病地図とよばれる疾病の発生分布を表す地図が基礎的な分析手段として用いられる. 疾病地図からは, ある特定の領域に疾病の高リスク値がまとまる空間クラスターがしばしば見いだされ, 感染拡大対策において優先的に対処すべき対象を知る手がかりが得られる (丹後ほか, 2007; 中谷ほか, 2004).

これまでも疾病空間クラスターの統計学的な検出手段として様々なものが提案されてきた. 特に

多様な位置・大きさを持つ空間クラスターを自動的に検出する手法として Openshaw et al. (1987), Besag and Newell (1991) 等は全地理空間領域の中を一定の規則に基づいて移動する小領域により走査 (スキャン) する空間クラスター検出法を提唱した. その系譜の中で現在最も広く利用されている手法が, 尤度比をベースとする Kulldorff (1997) の空間スキャン統計量である. この手法は, 疫学的イベントの発生を記述する確率分布モデルに応じて尤度比統計量を変更することで, 様々な拡張が容易である. その中でも, 疾病リスクの高低ではなく複数の種類の疾病患者の構成比について, 局所的な地域差を検出するために開発された手法が, 多項分布を利用した多項空間スキャン統計量である (Jung et al., 2010) .

* 学生会員 東北大学理学部 (Tohoku University)

〒980-0845 宮城県仙台市青葉区荒巻字青葉 6 番 3 号

E-mail : reo.matsumura.q1@dc.tohoku.ac.jp

** 正会員 東北大学大学院環境科学研究科 (Tohoku University)

本研究では、この多項空間スキャン統計量を用いて、新型コロナウイルスの遺伝子型別構成比の空間的特徴を複数の時点について全球スケールで解析し、ウイルスの遺伝子変異からみた COVID-19 の空間的な流行パターンについて考察した。

2. 使用するモデルとデータ

2.1. 多項空間スキャン統計量

一般的な空間スキャン統計量は、ポアソン分布やベルヌーイ分布などの確率分布に基づく尤度比検定統計量を用いて、有病率や死亡率で示される疾病リスクの集積領域を空間クラスターとして検出するために開発された。他方、多項空間スキャン統計量では、対象となるデータが複数の種類の疾病別の患者数のように複数のカテゴリ別カウント変数である場合に、そのカテゴリ別の患者数構成比に着目し、地図上の他の地域とカテゴリの構成比が統計的に有意に異なっている小領域を空間クラスターとして検出する。

いま、ある疾病が K 個のカテゴリに分かれており、対象地域は I 個の地域に分けられるとし、 c_{ik} を地域 i におけるカテゴリ k の患者数 ($k = 1, \dots, K, i = 1, \dots, I$) とする。ここで、小領域 Z を想定し p_{Zk}, q_{Zk} をそれぞれ Z 内と Z 外においてある患者がカテゴリ k の疾病である確率とした時、領域 Z を空間クラスターと判断すべきかの次のような局所的な仮説検定を考える。

帰無仮説 $H_{0Z} : p_{Z1} = q_{Z1}, \dots, p_{ZK} = q_{ZK}$

対立仮説 $H_{1Z} : \text{少なくとも1つのカテゴリ } k \text{ において } p_{Zk} \neq q_{Zk} \text{ が成立する}$

ただし、小地域 Z は通常は円領域を想定し、その中心位置や半径として多数の取り得る組み合わせが生じる。そのため、小領域 Z それぞれに対して、上記の仮説検定を繰り返すと、多重比較の問題が発生してしまう。そこで、大局的には、帰無仮説をすべての Z に対して H_{0Z} が成立すること、対立仮説としては、 H_{1Z} が成立するような領域 Z が少なくとも1つある仮説検定問題を考える。

これを達成するために、空間スキャン統計量では、尤度比に基づく統計量 $L_1(Z)/L_0$ の値が最大となる

「もっとも疑わしいクラスター」 (Most Likely Cluster ; MLC) 領域 Z を探し、この尤度比統計量の最大値 $\lambda(Z)$ を検定統計量とする。

$$\lambda(Z) = \max_Z L_1(Z)/L_0 \quad (1)$$

ここで多項空間スキャン統計量では、尤度関数は帰無仮説 H_{0Z} に従う場合、

$$L_0 = \prod_k \prod_i p_{0k}^{c_{ik}} \quad (p_k = q_k) \quad (2)$$

であり、対立仮説 H_{1Z} に従う場合は

$$L_1(Z) = \prod_k (\prod_{i \in Z} p_{Zk}^{c_{ik}} \prod_{i \notin Z} q_{Zk}^{c_{ik}}) \quad (3)$$

と表せる。ただし、 $c_i (= \sum_k c_{ik})$ を対象地域中の領域 i における患者総数、 $C_k (= \sum_i c_{ik})$ を対象地域全体におけるカテゴリ k の患者総数、そして $C (= \sum_k \sum_i c_{ik})$ を対象地域全体の患者総数とする。帰無仮説のもとで尤度関数 L_0 が最大となるような最尤推定量は $\hat{p}_{0k} (= \hat{q}_{0k} = C_k/C)$ 、対立仮説のもとで $L(Z)$ が最大となるような最尤推定量は $\hat{p}_{Zk} (= C_k(Z)/C(Z))$ 、 $\hat{q}_{Zk} (= (C_k - C_k(Z))/(C - C(Z)))$ であり、これを式 (2)、(3) にそれぞれ代入すると、局所的な尤度比検定統計量 $L_1(Z)/L_0$ は次のようになる。

$$L_1(Z)/L_0 = \begin{cases} \frac{\prod_k \hat{p}_{Zk}^{C_k(Z)} \hat{q}_{Zk}^{C_k - C_k(Z)}}{\prod_k \hat{p}_{0k}^{C_k}} & , p_{Z1} \neq q_{Z1} \vee \dots \vee p_{ZK} \neq q_{ZK} \\ 1, & \text{その他} \end{cases} \quad (4)$$

この最大値で定義される大域的な検定統計量 $\lambda(Z)$ の有意性の評価は、モンテカルロ検定により行う。モンテカルロ検定では、各地域の観測事象を帰無仮説の基でランダムに発生させ、そのランダムなデータに基づく検定統計量の分布を求めて p 値を計算する。多項スキャン統計量を用いる場合、患者の空間分布は所与とし、それぞれのカテゴリ別の患者総数は対象地域全体での観察値に一致するとして、それぞれの患者のカテゴリをランダムに割り当てる。つまり、ここで想定している帰無仮説は患者のカテゴリ別の出現比率は、空間的な位置に依らず一様であると考え (Jung et al., 2010)。

本研究では、多項空間スキャン統計量を用いた分析を行うために、SaTScan™ version 9.6

(<https://www.satscan.org/>) を使用した。全球スケールで検出されるクラスター半径が大きすぎると

地域ごとの時空間的なウイルス型別構成比の推移を解釈するのが困難であるため、クラスターの検索半径の最大値は 2000 km として分析を行い、MLC 以外の 2 次クラスターは p 値が 0.05 以下の空間クラスターのみを地理的な重複が起きない条件で検出した。なお検索半径の最大値は、これを半径とする円領域が欧州、東南アジア等の世界的地域区分に対応する空間クラスターをおおよそ検出する値として設定した。

2.2. データ

分析には、インフルエンザ情報共有の国際推進機構 (Global Initiative on Sharing All Influenza Data, GISAID) (<https://platform.gisaid.org/>) の提供する世界の国・地域における COVID-19 患者から得られた新型コロナウイルスのゲノムデータ (2020 年 7 月 25 日) を使用した。なお、検体が得られた詳細な地域名が記録されていないことも多いため、各ゲノムデータの地理的な参照位置は主に各国の首都の位置として分析を行った。ただし、中国については香港、台湾を独立した単位とした。

GISAID は、新型コロナウイルスの系統発生学的な遺伝距離に基づくウイルスの系統分類 (Rambaut et al., 2020; Tang et al., 2020) の情報を提供しており、本研究では、6 つの類似するグループ (G, GH, GR, L, S, V) とその他 (O) の 7 カテゴリ分類を利用した (以下、これを「(ウイルス)型」と呼ぶ)。図 1 にこの型に関する系統発生上の位置を示した。

この分類の 6 つの型 (G, GH, GR, L, S, V) は 9 つの遺伝子マーカーの組み合わせにより定義され、GISAID に登録されたゲノムデータの 95% がこの 6 つの型にいずれかに含まれる (2020 年 7 月 4 日現在)。この分類では、ウイルス発生初期は S 型と L 型が主な型であり、その後 L 型が変異して G 型と V 型に分かれ、G 型はさらに GH 型と GR 型へと変異していったと解釈できる。

また、世界の国・地域における新型コロナウイルス感染者数のデータは Our World in Data の Coronavirus Source Data (2020 年 8 月 22 日)

(<https://ourworldindata.org/coronavirus-source-data>) を用いた。

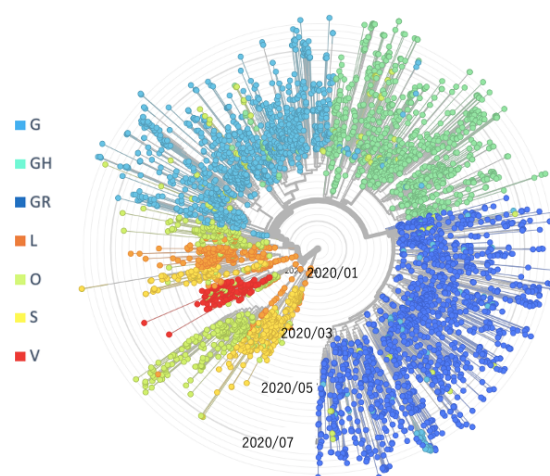


図 1：新型コロナウイルスの系統発生図
(GISAID により提供されている COVID-19 の系統発生図画像を改変，2020 年 8 月 23 日の情報による)

3. 分析結果と考察

3.1. 多項空間スキャン統計量を用いた分析結果

まず、7 つのタイプのウイルス型別構成比は 2020 年 1 月から 6 月にかけて図 2 のように推移してきた。

図 2 より、2 月と 3 月の間で登録されているウイルス型別構成比の大きな変化が確認できる。2 月までは L, O, S 型のウイルスが構成比の大部分を占めているのに対し、3 月以降は G, GH, GR 型の割合が高くなっている。そこで、この 2 月、3 月のウイルス型の空間分布を調べるために多項空間スキャン統計量による分析を行い、その結果を図 3-5 に示した。

図 3-5 中に示された円領域 (ただし投影法によって歪んでいるため真円にはなっていない) が多項空間スキャン統計量によって検出された空間クラスターであり、その付近に表示された円グラフは、検出された空間クラスターにおけるウイルス型別構成比を示している。また、図の下に全てのクラスター外に位置する国・地域のウイルス型別構成比も示した。

図 3 より、1・2 月時点での感染拡大の中心は中国であり、中国、東南アジア、オーストラリアで検出された空間クラスターでは L, O, S, V 型の占める割合が高いことがわかる。一方、欧州で検出された

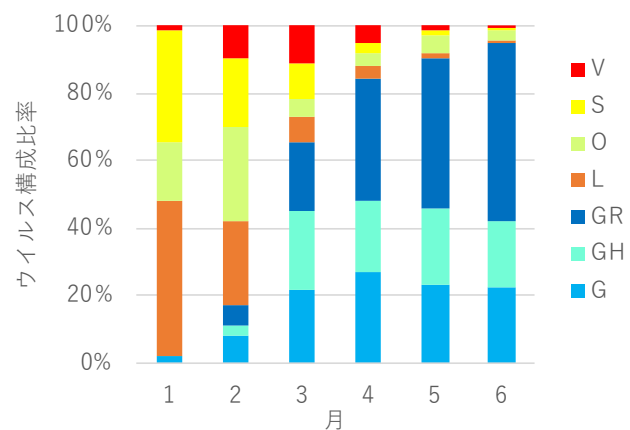


図2：新型コロナウイルス型別構成比の月毎の推移

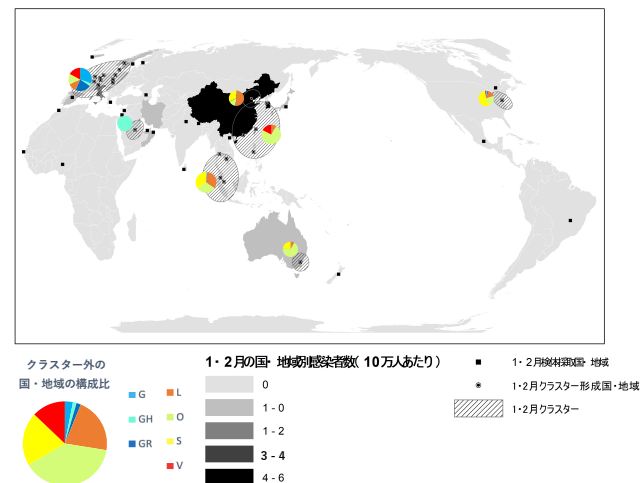


図3：1・2月の多項空間スキャン統計量によって検出された新型コロナウイルス型別構成比の空間クラスターおよび国・地域別感染者数（10万人あたり）

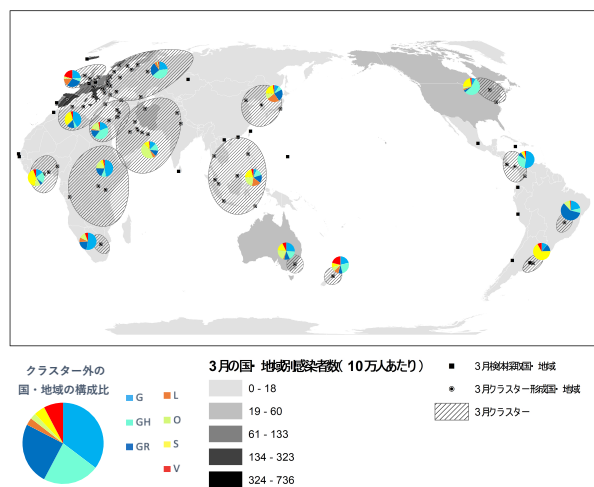


図4：3月の多項空間スキャン統計量によって検出された新型コロナウイルス型別構成比の空間クラスターおよび国・地域別感染者数（10万人あたり）

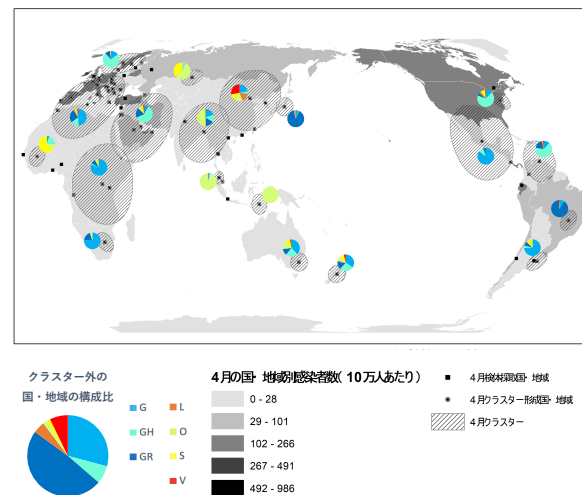


図5：4月の多項空間スキャン統計量によって検出された新型コロナウイルス型別構成比の空間クラスターおよび国・地域別感染者数（10万人あたり）

空間クラスターでは G, GR 型が半分以上の割合を占めており, アジアや他の国・地域とは大きく異なる構成比になっていたことが分かる. また, サウジアラビアでは GH 型のみが発生しており単独で空間クラスターを形成している.

これらの結果と 3 月以降の国・地域別のウイルス型別構成比を併せて考えると, 2 月から 3 月にかけて G, GH, GR 型のウイルスが世界的に急増した背景には, 1・2 月の欧州や中東起源のウイルスが高い感染力をもって (Korber et al., 2020) 世界的に拡大したことを示唆している.

また, 図 3-5 を見てみると, 近接した国・地域でもウイルス型別構成比が大きく異なるケースが観察される. 例えば, 図 5 の 4 月のウイルス型別構成比を見てみると, 日本で検出されたクラスターと中国・韓国を含むクラスターの構成比は大きく異なっている. 日本のクラスターでは欧州型の GR が大半を占めているのに対して, 中国・韓国を含むクラスターでは GR 型がほとんど見られず, G や GH 型も相対的に低い構成比になっている. これらの結果から日本では欧州からの人の流動によってもたらされたウイルス型によって 4 月にみられた流行拡大が導かれたことが示唆される (国立感染研究所, 2020). そのさらなる理解には, 旅客流動など各国間の人の移動を考慮した空間クラスターの解析が必要になると考えられる.

4. おわりに

本研究では, 多項空間スキャン統計量を用いた新型コロナウイルスの型別構成比の時空間的な分析を行い, 主要なウイルス型の全球的な感染経路について考察を行った. その結果, ウイルス型別構成比の時系列変化と多項空間スキャン統計量によるウイルス型別の特異な構成比を持った地域の検出によって, ウイルス型別と時空間的な感染拡大の関連を把握することができた.

ただし今回の分析では, Kulldorff (1997) による Circular scan 法をベースに空間クラスター領域の検出を行ったが, この手法を全球スケールで適用することには課題もある. この手法ではクラスター

一を円領域で検出するため, 線状や他の形状をした領域の空間クラスターの検出力は低い. また, 地理的には遠くても, 旅客流動量などが多ければ感染を多く生み出す社会的な距離は近いと考えることもできる. そのため, 全球スケールでの感染拡大をウイルスの移動に関連した人の流動に即した形で分析するためには, 人の流動に基づく社会的な距離を考慮する必要があると考えられる. 今後は全球スケールにおけるより適切な空間スキャン統計量の分析方法について検討を行い, さらなる分析を進める予定である.

参考文献

- 国立感染症研究所 (2020) 新型コロナウイルス SARS-CoV-2 のゲノム分子疫学調査 2 (2020/7/16 現在) <
<https://www.niid.go.jp/niid/ja/basic-science/467-genome/9787-genome-2020-2.html>>
- 丹後俊郎・横山徹爾・高橋邦彦 (2007). 『空間疫学への招待』, 医学統計学シリーズ 7, 朝倉書店, 東京.
- 中谷友樹・谷村晋・二瓶直子・堀越洋一 (2004) 『保健医療のための GIS』, 古今書院.
- Besag, J., & Newell, J. (1991). The Detection of Clusters in Rare Diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(1), 143.
- Desjardins, M. R., Hohl, A., & Delmelle, E. M. (2020). Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: Detecting and evaluating emerging clusters. *Applied Geography*, 118, 102202.
- Jung, I., Kulldorff, M., & Richard, O. J. (2010). A spatial scan statistic for multinomial data. *Statistics in Medicine*, 29(18), 1910-1918.
- Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W.,

- Hengartner, N., Giorgi, E. E.,
 Bhattacharya, T., Foley, B., Hastie, K.
 M., Parker, M. D., Partridge, D. G.,
 Evans, C. M., Freeman, T. M., de Silva,
 T. I., Angyal, A., Brown, R. L.,
 Carrilero, L., ... Montefiori, D. C.
 (2020). Tracking Changes in SARS-CoV-2
 Spike: Evidence that D614G Increases
 Infectivity of the COVID-19 Virus. *Cell*,
 182(4), 812-827.e19.
- Kulldorff, M. (1997). A spatial scan
 statistic. *Communications in Statistics -
 Theory and Methods*, 26(6), 1481-1496.
- Mo, C., Tan, D., Mai, T., Bei, C., Qin, J.,
 Pang, W., & Zhang, Z. (2020). An analysis
 of spatiotemporal pattern for COVID-19
 in China based on space-time cube.
Journal of Medical Virology, 92(9), 1587-
 1595.
- Openshaw, S., Charlton, M., Wymer, C., &
 Craft, A. (1987). A mark 1 geographical
 analysis machine for the automated
 analysis of point data sets.
*International Journal of Geographical
 Information Systems*, 1(4), 335-358.
- Rambaut, A., Holmes, E. C., O' Toole, Á.,
 Hill, V., McCrone, J. T., Ruis, C., du
 Plessis, L., & Pybus, O. G. (2020). A
 dynamic nomenclature proposal for SARS-
 CoV-2 lineages to assist genomic
 epidemiology. *Nature Microbiology*, 1-5.
- Tang, X., Wu, C., Li, X., Song, Y., Yao, X.,
 Wu, X., Duan, Y., Zhang, H., Wang, Y.,
 Qian, Z., Cui, J., & Lu, J. (2020). On
 the origin and continuing evolution of
 SARS-CoV-2. *National Science Review*,
 7(6), 1012-1023.