北海道科学大学研究紀要
第50号（2022年）
研究報告

Bulletin of Hokkaido University
of Science, No.50(2022)
Research Reports

教師なし学習を用いたＣＯＶＩＤ－１９新規感染者数のクラスター分析

Cluster Analysis on the Number of COVID-19 Newly Infected People using Unsupervised Learning

小松 隆行*

Takayuki Komatsu

概要

本稿では，インターネット上に公開されている COVID-19（新型コロナウィルス）に関するオープンデータに関して，相関分析とクラスター分析を行う．対象とするデータは，日本の全都道府県，日本を含むアジア諸国，およびヨーロッパ諸国の新規感染者数の日次データである．これらを各々時系列データと捉え，地域間の時系列データを統計学的に相関分析し，さらに機械学習の教師なし学習を用いてデータサイエンス的な視点から似たものを同一グループに分類するクラスター分析を試みる．

1. はじめに

2020 年 1 月頃から 2 年を経た現在も継続している COVID-19（新型コロナウィルス）のパンデミックは，人類の健康や社会的危機，経済的危機を生じさせている．ワクチン接種が進んでいるにもかかわらず，感染者の急拡大と収束を繰り返している状況にある．これに関連する情報は，数値データのオープンデータとして，全世界の多くの機関がインターネット上に公開しており，それらをダウンロードし分析や解析を行うことが可能となっている．これらのサイトから新規感染者や現在患者数，死者数，病床占有率など多種類のデータがオープンデータとして CSV 形式や Excel 形式でダウンロード可能である．その種類によっては，データ取得間隔の長さや欠損などがあるが，これらのデータの推移をグラフ化するなどして可視化したり，基本的な統計分析したりした結果や知見は，日本では行政やメディアのインターネットサイト[(1)]などにおいて，世界の国と地域のデータは世界保健機関（WHO）のサイト[(2)]で公開されている．北海道においても道内の大都市や振興局単位での感染状況のデータが公開されており，ダウンロードし解析することを可能にしている．

本報告の目的は，COVID-19 の新規感染者数のデータをある非定常の情報源から得られる実測値データと考え，これらの時系列データのみから統計学的手法とデータサイエンスの機械学習の手法を用いることによって何らかの知見を得ることである．具体的には，相関分析や機械学習の教師なし学習と呼ばれている手法の Time Series k-means 法[(3)]によるクラスタリング（自動グループ分け）を試みて考察を行うことにする．

2. 全国４７都道府県の新規感染者数に関する分析

まず，日本全国の４７都道府県毎のデータ[(1)]について分析してみる．ここでは，全国的に同時に急拡大した第５波よりも前の，都道府県毎の新規感染者数のデータの増減の傾向に差異が見られた第１波から第４波までの期間である 2020 年 1 月 16 日から 2021 年 6 月 28 日までのデータ使用することにする．

2.1 全国４７都道府県間の相関分析

まず，相関分析を行う．第１波から第４波の期間，2020/1/16～2021/6/28 のデータを用いて，都道府県間の相関係数を算出し，その値のヒートマップとして図 1 に示した．縦軸と横軸には，４７都道府県名が配置され，座標の交差した場所に算出された相関係数が記載されている．都道府県名は，その所在地が概ね北から南の順に地方毎にまとめられて記載されている．この図 1 では，相関係数の大小に依存して該当するマス目の色が水色（相関係数 0.0 ）から赤色（相関係数 1.0）まで変化して表示されている．相関係数が 0.9 以上場合は，2 つの変数間に

＊北海道科学大学未来デザイン学部メディアデザイン学科

非常に強い相関があると言われている．図１のヒートマップでは，濃い赤色のマス目の部分が非常に強い相関があると言える．また，相関係数が 0.7 以上 0.9 未満の場合は強い相関があると言われており，図１のヒートマップでは橙色の部分である．図１において，赤色のマス目が集中しそれらの周囲にも橙色のマス目が集中している場所としては，関東地方（東京，神奈川，埼玉，千葉，栃木，群馬），東海地方（愛知，岐阜，三重，静岡），関西地方（大阪，兵庫，京都，和歌山，奈良）が挙げられ，これらの地域間の相関係数の値も高いことが分かる．また東北地方では，宮城県と山形県に強い相関があることが分かる．一方，鳥取，島根，高知，山梨，福井などは他都道府県との相関が低いことが分かる．さらに，近い地方の組合せの分析として，図２には関東と中部，及び関西地方の都府県間，図３には北海道と東北，及び関東の都道県間，図４には中国と九州の県間のヒートマップとデンドログラムをそれぞれ示した．これらのデンドログラムからも，所在地が近い関東，関西，九州から中国西部の各地方の都道府県が概ね同一のクラスターとして分類されていることが分かる．一方で，沖縄，山陰地方，山梨，福井は他県との相関が低いクラスターに分類されている．

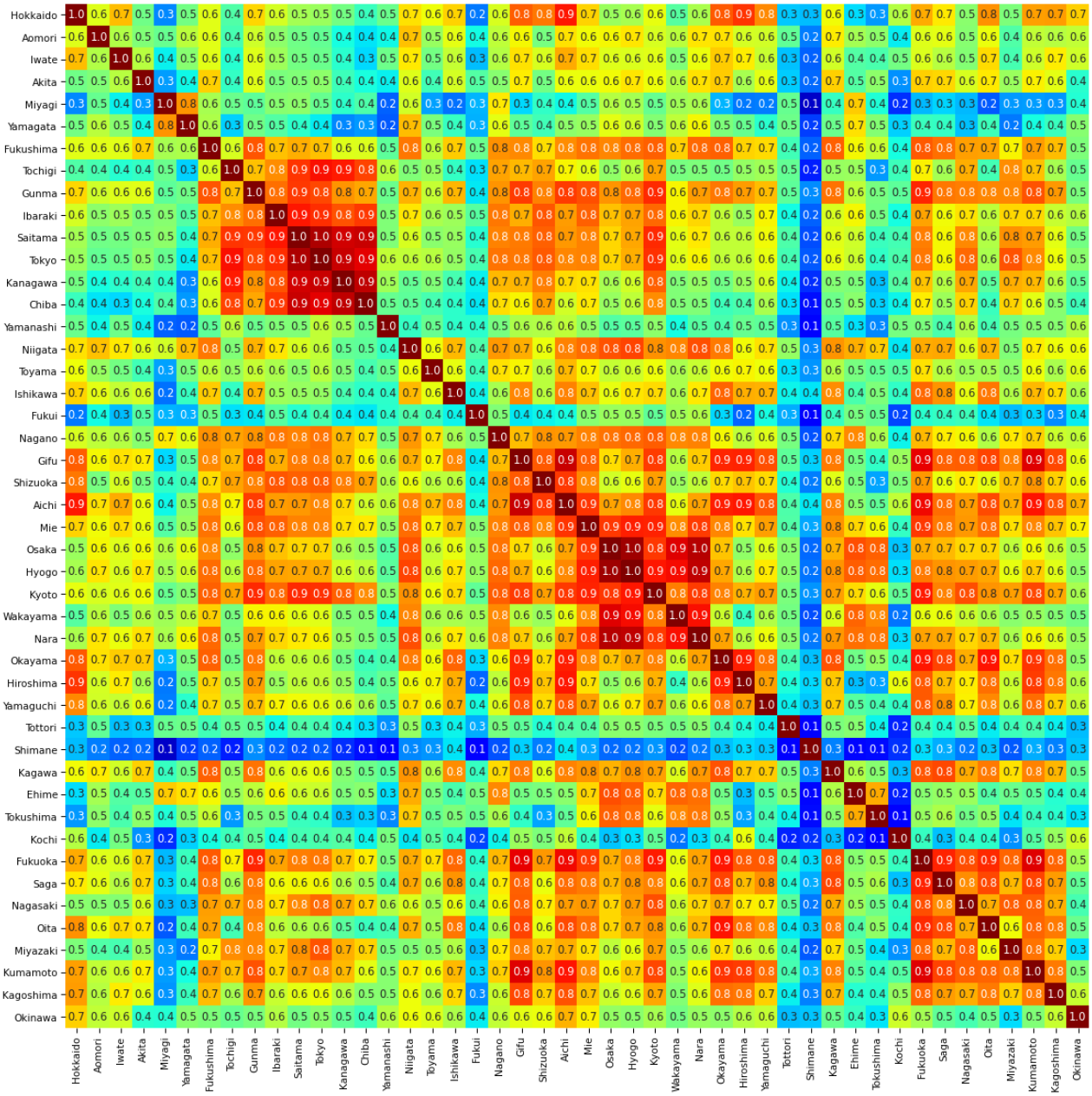

図１　全国４７都道府県間の相関係数のヒートマップ（2020/1/16～2021/6/28）

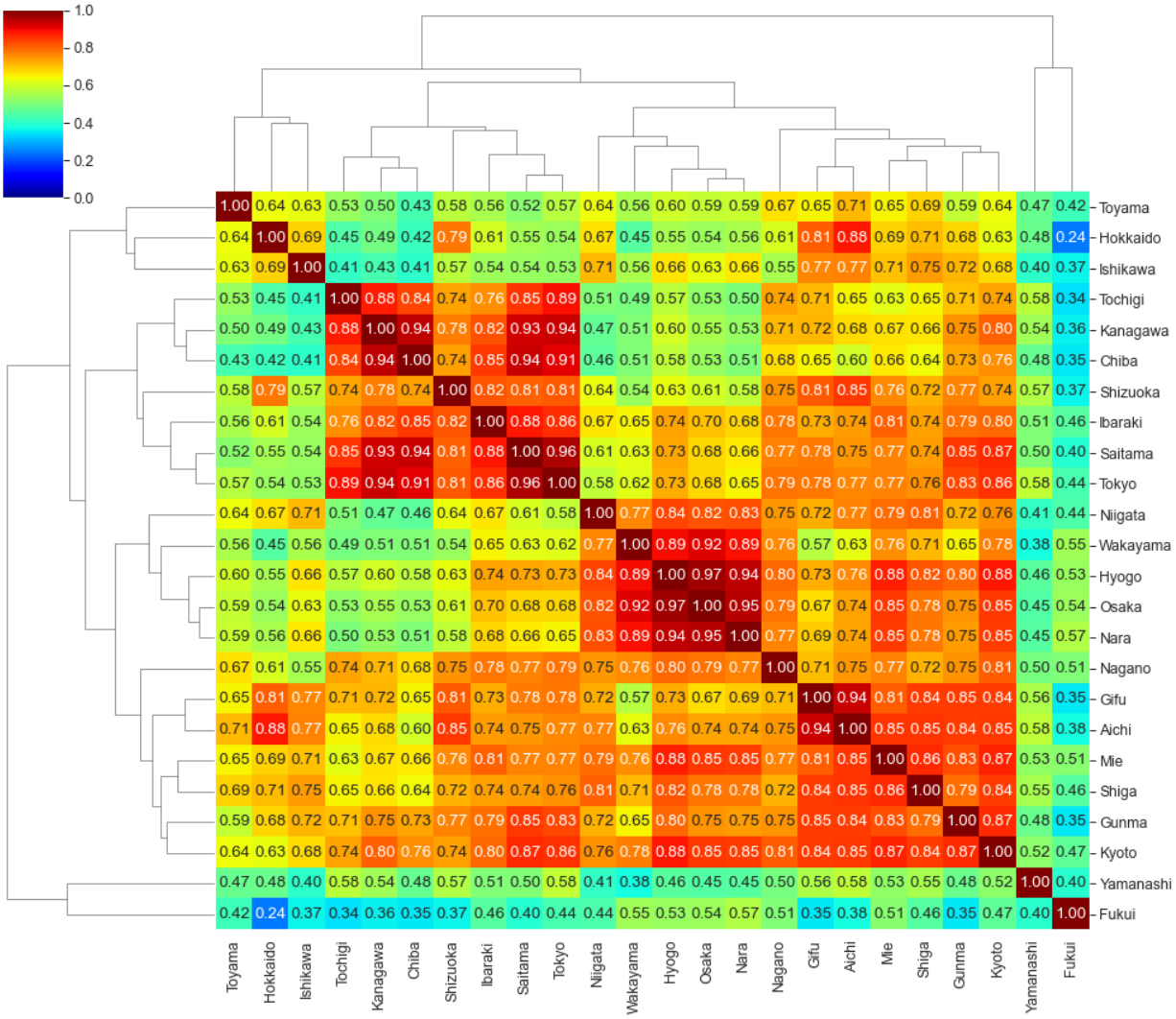

図 2　関東/中部/関西の相関係数のヒートマップとデンドログラム（2020/1/16～2021/6/28）

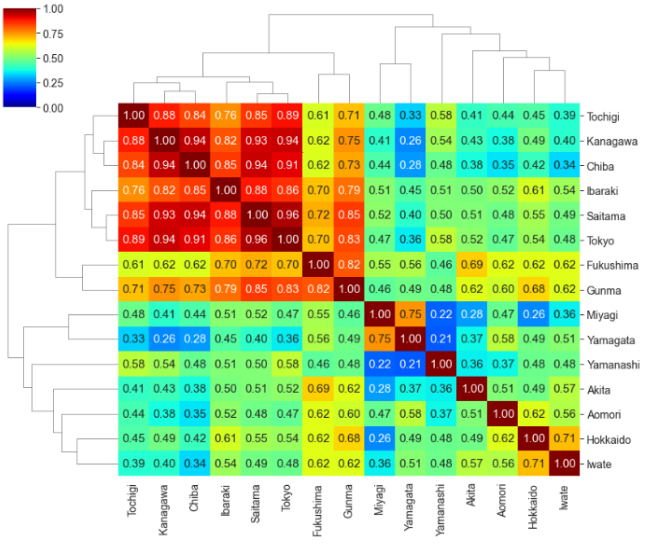

図 3　北海道/東北/関東の相関係数のヒートマップとデンドログラム（2020/1/16～2021/6/28）

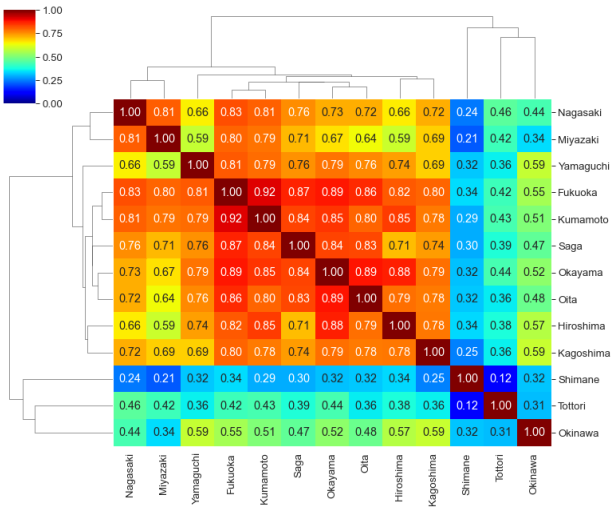

図 4　中国/九州の相関係数のヒートマップとデンドログラム（2020/1/16～2021/6/28）

2.2 全国４７都道府県のクラスタリング

ここではまず，全国４７都道府県毎に平均と標準偏差で標準化された新規感染者数の 14 日平均の推移を Time series k-means 法[3]（Euclid 計量）を用いてクラスタリングを行った．クラスター数はエルボー法により 10 とした．図 5 は，同じクラスターに分類された都道府県毎の推移を，そのクラスター毎にまとめてプロットしたグラフである．横軸は日付ですべて同一であり，縦軸は標準化による標準偏差の何倍かを表している．それに対応してスケールと最大値がクラスター毎に異なることに留意する．

図 5 から，関東，東北と北陸，関西と隣接県，などがそれぞれ大きなクラスターを形成し，隣接県同士の鳥取と島根，宮城と山形がそれぞれ同じクラスターとなっていることが分かる．

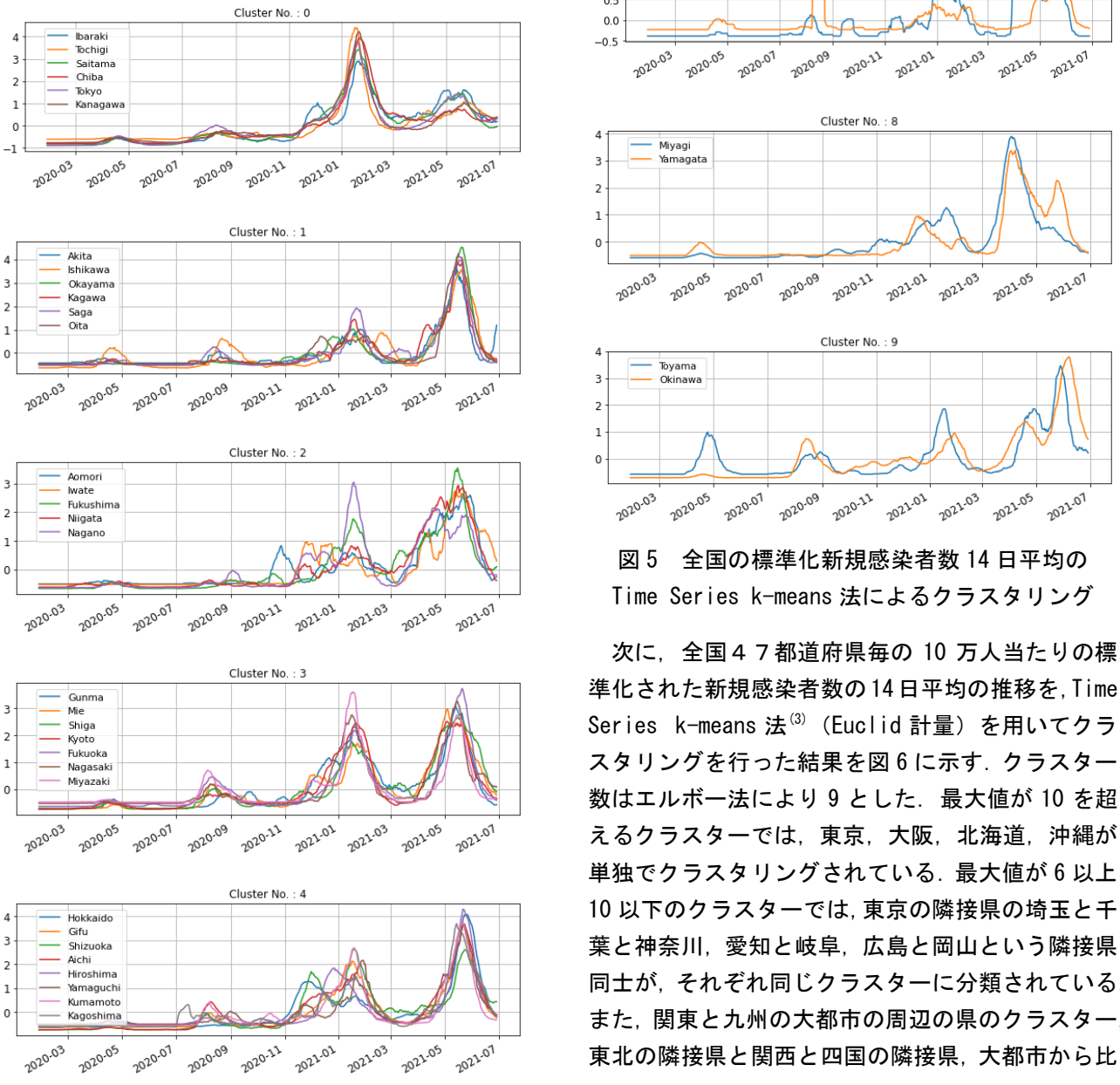

図 5　全国の標準化新規感染者数 14 日平均の Time Series k-means 法によるクラスタリング

次に，全国４７都道府県毎の 10 万人当たりの標準化された新規感染者数の 14 日平均の推移を，Time Series k-means 法[3]（Euclid 計量）を用いてクラスタリングを行った結果を図 6 に示す．クラスター数はエルボー法により 9 とした．最大値が 10 を超えるクラスターでは，東京，大阪，北海道，沖縄が単独でクラスタリングされている．最大値が 6 以上 10 以下のクラスターでは，東京の隣接県の埼玉と千葉と神奈川，愛知と岐阜，広島と岡山という隣接県同士が，それぞれ同じクラスターに分類されている．また，関東と九州の大都市の周辺の県のクラスター，東北の隣接県と関西と四国の隣接県，大都市から比

較的遠い県のクラスターも形成されている.

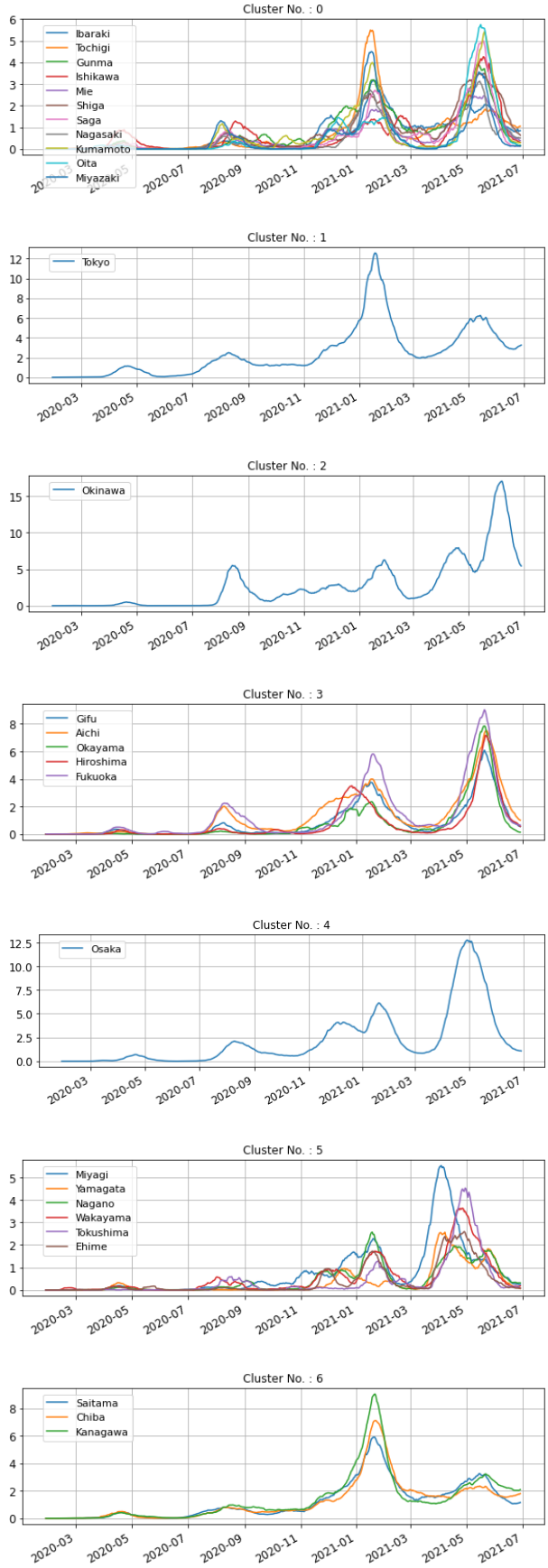

図 6　10 万人当たりの標準化された新規感染者数の 14 日平均の Time Series k-means 法によるクラスタリング（2020/1/16～2021/6/28）

3. アジア諸国とヨーロッパ諸国についての分析

ここでは，2 章と同様にアジア諸国とヨーロッパ諸国について分析を試みる．データは世界保健機関（WHO）のサイト[(2)]から取得した．使用データの期間は，2020 年 1 月 16 日から 2021 年 6 月 28 日までである．図 7 と図 10 は，それぞれアジア諸国間，及びヨーロッパ諸国間の相関係数のヒートマップとデンドログラムである．また図 8 と図 9，及び図 11 は，それぞれアジア諸国とヨーロッパ諸国の標準化された新規感染者数 7 日平均の Time Series k-Means 法[(3)]によるクラスタリングの結果である．クラスター数は，エルボー法により 8 及び 10 とした．

アジアでは，単独国のクラスターを除いて，ヒートマップと共に考察すると，おおよそインドその周辺国，タイとカンボジアとマレーシアとその周辺国，中国の周辺国（モンゴルやベトナムなど），日本とインドネシアと韓国などの少し離れた島国のような国々で，それぞれクラスターが形成されている．

ヨーロッパでは，ドイツとオランダとデンマークとスウェーデン，イタリアとスイスとオーストリア，ブルガリアとセルビアとハンガリー，フランスとベルギー，スペインとポルトガル，フィンランドとノルウェイ，などの隣接国が各々同じクラスターに分類されている．島国のアイルランドやアイスランドは，単独クラスターとなっている．ヒートマップと共に考察すると，ドイツとその北の隣接国，イタリアとオーストリア周辺，旧東欧諸国周辺などで，それぞれ強い相関を示している．

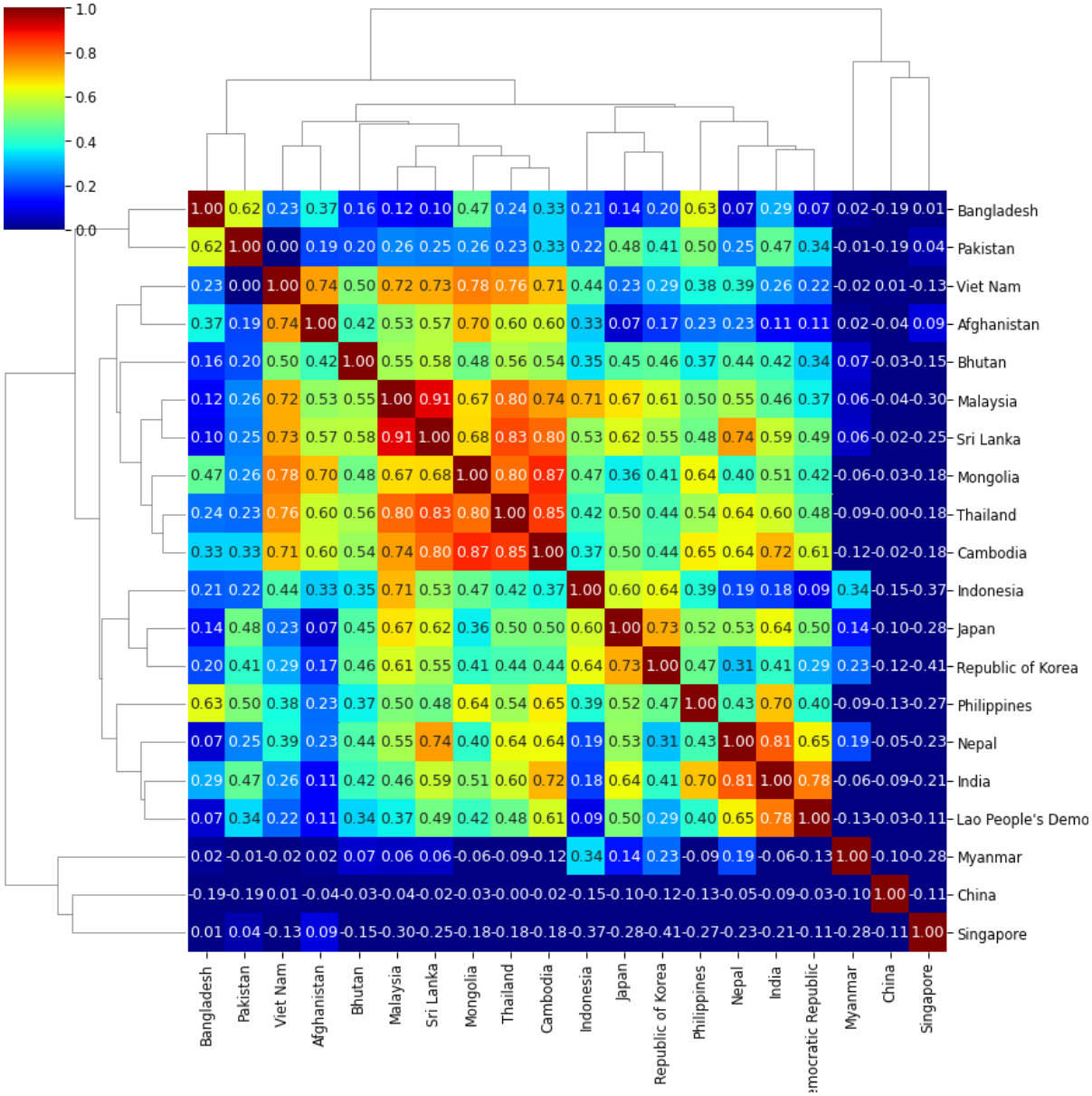

図 7　アジア諸国間の相関係数のヒートマップとデンドログラム（2020/1/16～2021/6/28）

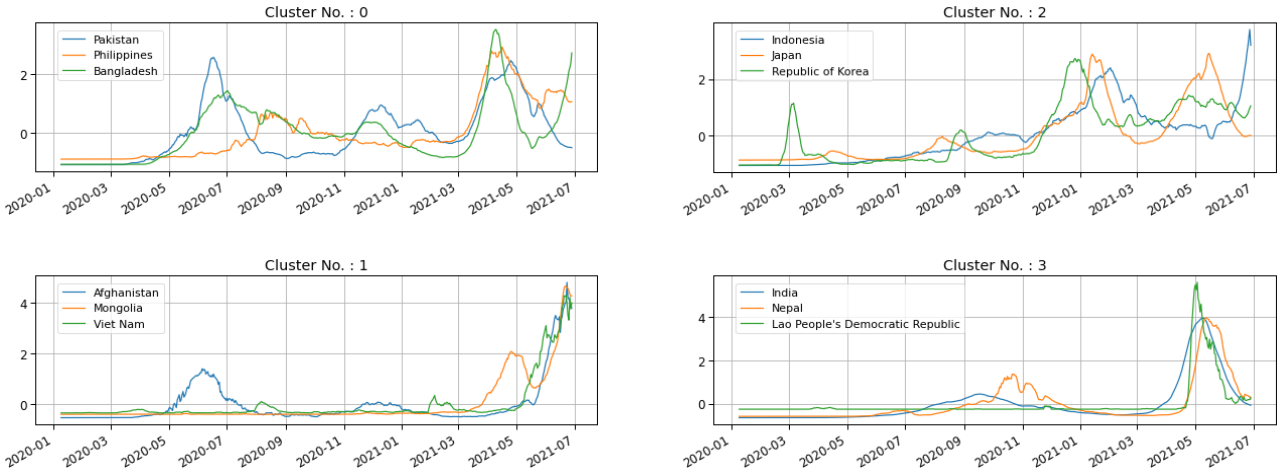

図 8　アジア諸国の標準化新規感染者数 7 日平均の Time Series k-Means 法によるクラスタリング（続く）

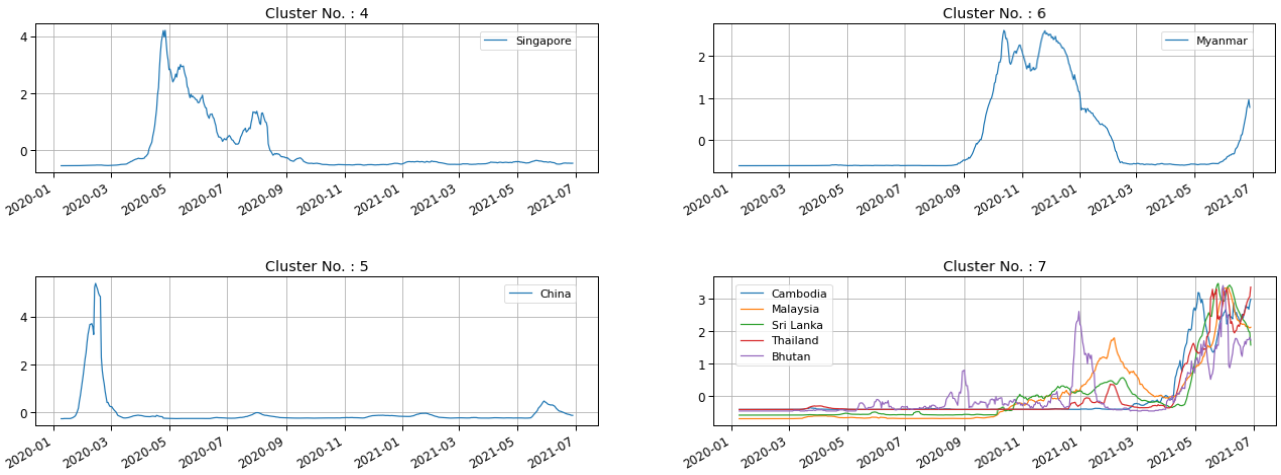

図 9　アジアの標準化された新規感染者数 7 日平均の Time Series k-Means 法によるクラスタリング

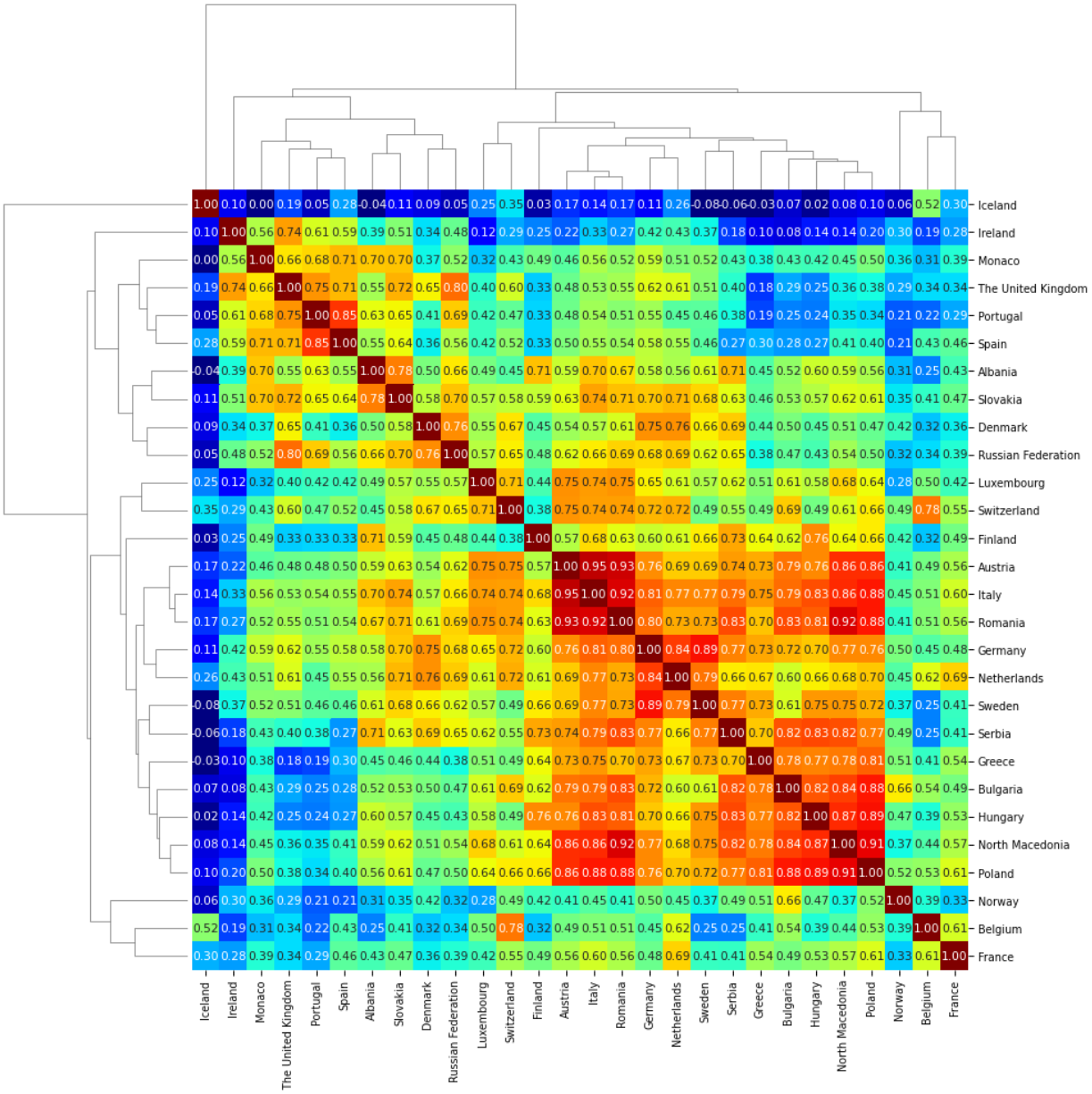

図 10　ヨーロッパ諸国間の相関係数のヒートマップとデンドログラム（2020/1/16～2021/6/28）

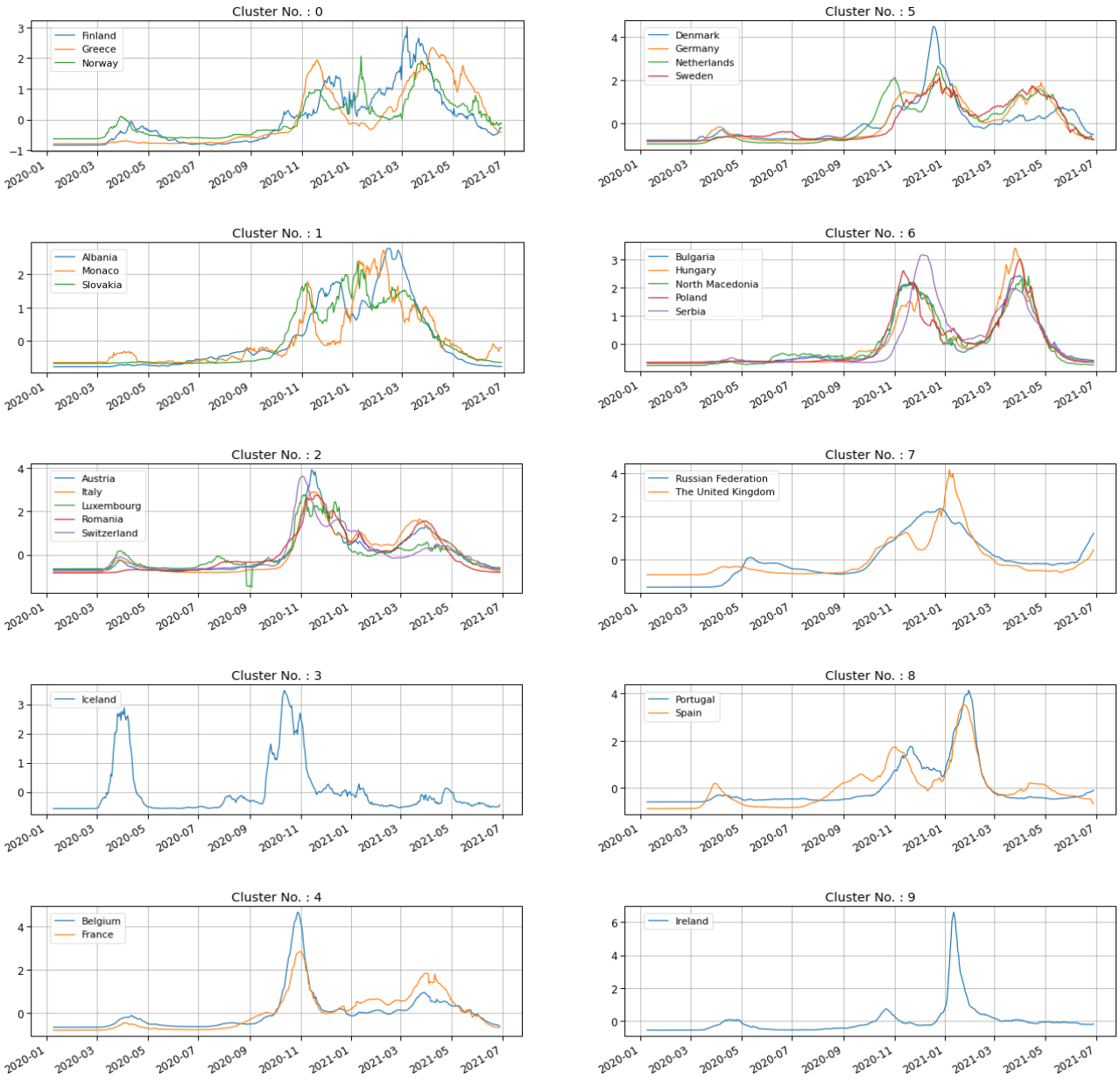

図 11　ヨーロッパの標準化された新規感染者数 7 日平均の Time Series k-Means 法によるクラスタリング

4. まとめ

COVID-19 の新規感染者数のオープンデータを時系列データと捉えて，相関分析と機械学習の教師なし学習によるクラスタリングを試みた．これらの分析だけでは，因果関係の有無を論じることはできないが，データサイエンスの視点からの分析におけるいくつかの知見は得られたと考えられる．今後は，新たなデータを加えた分析や，他のデータサイエンスや機械学習の分析手法との比較，それらを組み合わせた分析や応用が必要であると考えられる．

参考文献

(1) ＮＨＫ特設サイト新型コロナウィルス:新型コロナデータ一覧：都道府県ごとの感染状況, 2021 年 6 月 30 日, https://www3.nhk.or.jp/news/special/coronavirus/data-widget/.

(2) World Health Organization: WHO Coronavirus (COVID-19) Dashboard, 2022 年 2 月 27 日, https://covid19.who.int/info/.

(3) Xiaohui Huang, Yunming Ye, Liyan Xiong, Raymond Y.K. Lau, Nan Jiang, Shaokai Wang: Time series k-means: A new k-means type smooth subspace clustering for time series data, Information Sciences, Volumes 367-368, 1 November, pp. 1-13, 2016.