コミュニティ抽出法による産業クラスタの検出

佐藤　加斐[1]・福本　潤也[2]

[1]学生会員　東北大学大学院　情報科学研究科（〒980-8577 仙台市青葉区片平2-1-1）
E-mail:kai_sato@plan.civil.tohoku.ac.jp
[2]正会員　東北大学大学院准教授　情報科学研究科（〒980-8577 仙台市青葉区片平2-1-1）
E-mail:fukumoto@plan.civil.tohoku.ac.jp

産業集積の働きを強化する産業政策に対する関心が世界中で高まっている．産業政策のあり方を議論する前段において，産業が集積する地域（産業クラスタ）を把握したり，産業集積を構成する産業群を把握することは有益である．本研究では，ネットワーク科学で発展したコミュニティ抽出法に基づく産業クラスタの検出手法を提案する．具体的には，地理的単位毎の事業所数と地理的単位間の空間距離を用いて仮想的なネットワークを定義した上で，仮想ネットワークに対してコミュニティ抽出法を適用して産業クラスタを検出する手法を提案する．さらに，仮想ネットワークの定義やコミュニティ抽出法の適用時に用いるパラメータの設定法についても検討する．市区町村別の事業所数データを用いたケーススタディを行い，提案手法の有効性を検証する．

Key Words :industrial cluster, community finding, modularity, BIC

1.　はじめに

産業集積を強化する産業政策への関心が世界中で高まっている．日本でも，既存産業の国際競争力の強化や新産業の育成を目的として，産業クラスタ政策を始めとする様々な政策を推進している．産業集積のあり方について議論する前段において，単一あるいは複数の産業が集積する地理的単位の集合（産業クラスタ）を把握したり，集積しやすい産業群を把握することが有益である．本研究では，産業クラスタの検出問題に着目する．

産業クラスタの検出手法を提案した先行研究に，Mori and Smith (2010)がある．彼らは，離散化された地理空間上で事業所が多項分布に従って立地するという仮定を置き，BIC 最大化基準により産業クラスタを検出する手法を提案した．産業クラスタの生成プロセスを確率モデルとして定式化しており，理論的整合性が非常に高い手法である．

ただし，彼らの手法は詳細地理情報を用いた産業クラスタの検出問題への拡張は難しいと考えられる．何故なら，産業クラスタの検出にあたり，BIC が増大するように地理的に近接する地理的単位を逐次的に結合していく計算を行っており，さらに，検出される産業クラスタが擬凸性の条件を満たさなければならないという制約条件を置いているからである．詳細地理情報を利用する場合，計算量が必然的に多くなるため，逐次計算の計算負荷を小さくする必要がある．擬凸性の条件も厳しすぎる可能性があり，擬凸性を満たさずに地理的に隣接した地理単位を結合する手法が求められる．

また，彼らの手法は複数の産業で構成される産業クラスタの検出問題への拡張も容易ではないと考えられる．複数産業で構成される産業クラスタを検出する場合，地理的単位間の地理的近接関係に加えて，産業間の関係性も考慮する必要があり，産業数が増加すると，逐次計算の計算負荷の問題が生じると考えられる．

これに対し，本研究では，ネットワーク科学分野で発展したコミュニティ抽出法に基づく産業クラスタの検出方法を提案する．具体的には，地理的単位毎に得られる事業所数と経済面積のデータと，地理的単位間の空間距離を用いて仮想的なネットワークを定義した上で，仮想ネットワークに対してコミュニティ抽出法を適用して産業クラスタを検出する手法を提案する．さらに，仮想ネットワークの定義や，コミュニティ抽出法の適用時に用いるパラメータの設定法についても検討する．具体的には，Mori and Smith が産業クラスタの検出基準として用いた BIC(Bayesian Information Criterion)を用いてパラメータを設定する方法を検討する．提案手法（産業クラスタの検出法とパラメータの設定法）をレース・繊維雑品製造業のクラスタ検出問題に適用し，提案手法の有効性を検討する．

なお，本稿では，Mori and Smith と同じく市区町村を地

理的単位とする単一産業の産業クラスタ検出問題を取り上げる．今後の研究では，コミュニティ抽出分野における研究成果を援用して，詳細地理情報を用いた産業クラスタの検出法や複数産業で構成される産業クラスタの検出法へと拡張していく計画である．

2. 産業クラスタ検出手法

(1) 地理空間の想定と利用可能な情報

離散化された地理空間を想定する．地理的単位毎に産業別事業所数と経済面積（湖沼や山間地を除いた面積）が記録されているとする．また，地理的単位間の空間距離が定義されているとする．以下では，地理的単位毎に記録された産業別事業所数と経済面積，地理的単位のペア毎に記録された空間距離を用いて，産業クラスタを検出する問題について考える．

(2) モジュラリティ最大化法

本研究では，代表的なコミュニティ抽出法であるモジュラリティ最大化法に基づく産業クラスタの検出手法を提案する．モジュラリティ最大化法は，観測されたネットワークと，分析者によって仮説的に設けられた帰無ネットワークを比較することにより，観測ネットワーク内でリンクが高密度に張り巡らされたノードの部分集合（コミュニティ）を抽出する手法である．具体的には，以下の最大化問題として定式化される．

$$\max_{\{C_i\}_{i\in N}} Q=\sum_{i,j}\left(A_{ij}-\gamma P_{ij}\right)\delta\left(C_i,C_j\right) \qquad (1)$$

ただし，A_{ij} は観測ネットワークにおけるノード i,j 間のリンクの本数（重みなしネットワークの場合はノードの有無），P_{ij} は帰無ネットワークにおけるノード i,j 間のリンクの本数の期待値である．C_i はノード $i\in N$ が帰属するコミュニティを表す変数，δ はデルタ関数である．γ は解像度パラメータと呼ばれ，検出されるクラスタの大きさや数を調節する働きを持つ．通常，観測ネットワークのリンク総数と帰無ネットワークのリンク総数の期待値が等しくなるように帰無ネットワークを定義して，$\gamma\geq 1$ を満たす解像度パラメータを用いる．

モジュラリティ最大化法を産業クラスタ検出問題に援用するにあたり，観測ネットワークと帰無ネットワークを以下の通り定義する．

(3) 観測ネットワークの定義

新たに定義する観測ネットワークのノードは，地理的単位に一対一対応するとする．リンクについては，任意のノード間に重み付きリンクが張られているとする．地理的単位 i に立地する（ある産業の）事業所数を x_i，地理的単位 i と地理的単位 j の空間距離を d_{ij} で表し，リンク ij の重みを

$$A_{ij}=W_{ij}x_ix_j \qquad (2)$$

と定義する．ただし，W_{ij} は空間重み行列の ij 成分である．W_{ij} の定式化として，距離減衰パラメータ α を用いる次の2つの定式化を以下では考える．

$$W_{ij}=1/d_{ij}^{\alpha} \qquad (3)$$

$$W_{ij}=\exp(-\alpha d_{ij}^2) \qquad (4)$$

(4) 帰無ネットワークの定義

帰無ネットワークの定義にあたり，Mori and Smith と同じく，地理的単位 i にある事業所が立地する確率が，地理的単位 i の経済面積に比例する多項分布に従うという帰無仮説を採用する．この時，帰無ネットワークにおけるリンクの重みの期待値が

$$P_{ij}=\frac{e_ie_j}{E^2}X\left(X-1\right) \qquad (5)$$

に比例する．ただし，e_i は地理的単位 i の経済面積，E は分析対象エリア全体の経済面積，X は分析対象地域内に立地する事業所数の総数である．

(5) 産業クラスタ検出問題

(3)で定義した A_{ij} と(4)で定義した P_{ij} を用いると，産業クラスタ検出に用いるモジュラリティを

$$Q=\sum_{i,j}\left(A_{ij}-\gamma P_{ij}\right)\delta\left(C_i,C_j\right) \qquad (6)$$

と定義できる．ただし，解像度パラメータ γ は，次の条件を満たすものとする．

$$\gamma\geq\frac{\sum_{ij}A_{ij}}{\sum_{ij}P_{ij}} \qquad (7)$$

(6) アルゴリズム

モジュラリティを最大化するアルゴリズムとして，以下では，簡便な Greedy Algorithm を改良した手法を用いる．アルゴリズムの手順は次の通りである．

i) 全てのノードを独立した仮クラスタとして定義する．
ii) 任意の仮クラスタペアについて，結合した場合のモジュラリティの増加量を計算する．
iii) 最大のモジュラリティ増加量が正の場合には当該仮クラスタペアを結合し，1つの新しい仮クラスタを生成する．その後，ii)に戻る．最大のモジュラリティ増加量が正の場合には，iv)に進む．
iv) 全ての仮クラスタについて，部分モジュラリティ Q_c を計算し，部分モジュラリティ Q_c が正の

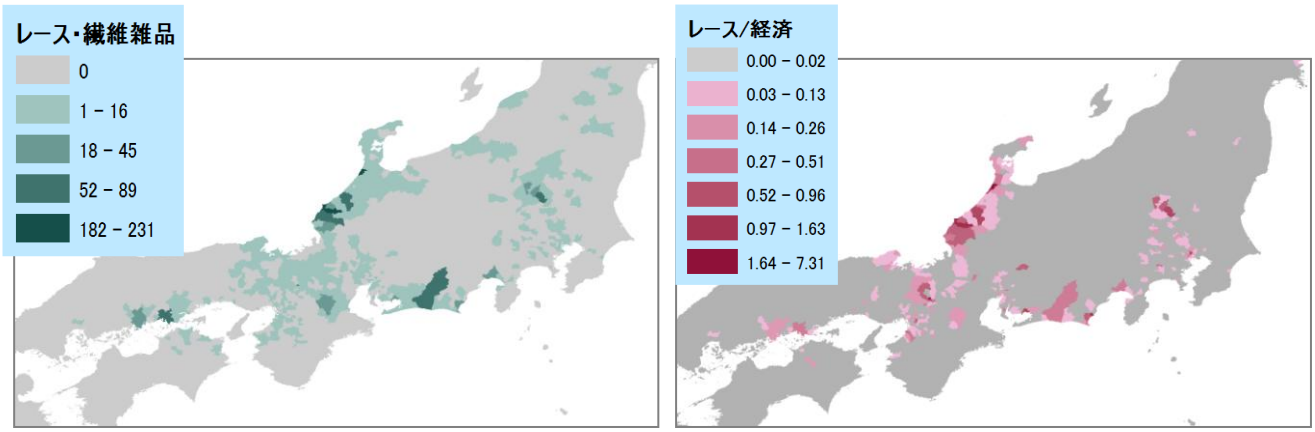

図-1　レース・繊維雑品製造業の事業所数(左)と単位経済面積当たりの事業所数(右)

仮クラスタを検出されるクラスタとする．ただし，$Q_c \equiv \sum_{i,j\in c}(A_{ij}-\gamma P_{ij})$ である．

3.　パラメータ設定法

2. で提案した産業クラスタ検出法を適用する場合，空間重み行列に含まれるパラメータ α とモジュラリティの定義式に含まれる解像度パラメータ γ の値を決める必要がある．本研究では，パラメータを設定するための指標として，Mori and Smith でも使用した BIC(Bayesian Information Criterion)基準に着目する．BIC は以下の式(8)で定義される．

$$BIC(\alpha,\gamma) = L(\alpha,\gamma) - \frac{k_C}{2}\ln(n) \tag{8}$$

第一項は，パラメータ α,γ を用いて得られたクラスタ検出結果の対数尤度である．事業所分布が多項分布に従うと仮定すると，次式で表される．

$$L(\alpha,\gamma) = \sum_{j=0}^{k_C} n_j \ln\left(\frac{n_j}{n}\right) + \sum_{j=0}^{k_C}\sum_{r\in C_j} n_r(\ln a_r - \ln a_{C_j}) \tag{9}$$

ただし，k_C は検出されたクラスタ数，n は分析対象範囲内の事業所総数，n_j はクラスタ C_j 内の事業所数，n_r は地理的単位 r 内の事業所数，a_{C_j} はクラスタ C_j に含まれる地理的単位の経済面積の合計，a_r は地理的単位 r の経済面積である．一方，第二項は，クラスタ数の増加（すなわち，推定パラメータの増加）によるペナルティ項である．

BIC 基準に基づくパラメータ設定の手順は以下の通りである．まず，空間重み行列のパラメータ α と解像度パラメータ γ の複数の組み合わせを格子状に設定し，それぞれを用いて産業クラスタ検出を行う．次に，それぞれの検出結果について BIC を計算する．最後に，最も BIC が大きくなる検出結果を探索する．

4.　ケーススタディ

(1)　ケーススタディの条件

ケーススタディでは，日本全国の1,858市区町村を分析対象範囲とするレース・繊維雑品製造業の産業クラスタ検出問題に提案手法（産業クラスタ検出法とパラメータ設定法）を適用する．同産業を取り上げた理由は，事業所が本州の関東から中国地方に主に分布しており，事業所が比較的集積しているため，産業クラスタが検出しやすいと考えられたからである．事業所数と単位経済面積当たりの事業所数の分布は図-1に示した通りである．下野（足利），北陸（福井・石川），遠江（浜松市），京阪（京都・大阪），備後（倉敷・福山）などに集積がみられる．

仮想ネットワークと帰無ネットワークの定義にあたり，事業所数データは，平成18年の事業所・企業統計調査の調査結果を用いる．地理的単位間の空間距離は，それぞれの市区町村の市区役所・町村役場の直線距離を用いて定義する．市区町村の内々距離は，栗田・越塚の領域間平均距離の近似公式を用いて定義する．経済面積は，統計局・政策統括官・統計研修所の2012年「統計でみる都道府県・市町村」の「可住地面積」データを用いる．

パラメータ設定については，空間重み行列のパラメータ α と解像度パラメータ γ の組み合わせを格子状に設定する代わりに，空間重み行列のパラメータ α と次式で定義されるパラメータ γ' の組み合わせを格子状に設定する．ただし，$\gamma' \geq 1$ である．

$$\gamma = \gamma' \frac{\sum_{ij} A_{ij}}{\sum_{ij} P_{ij}} \tag{10}$$

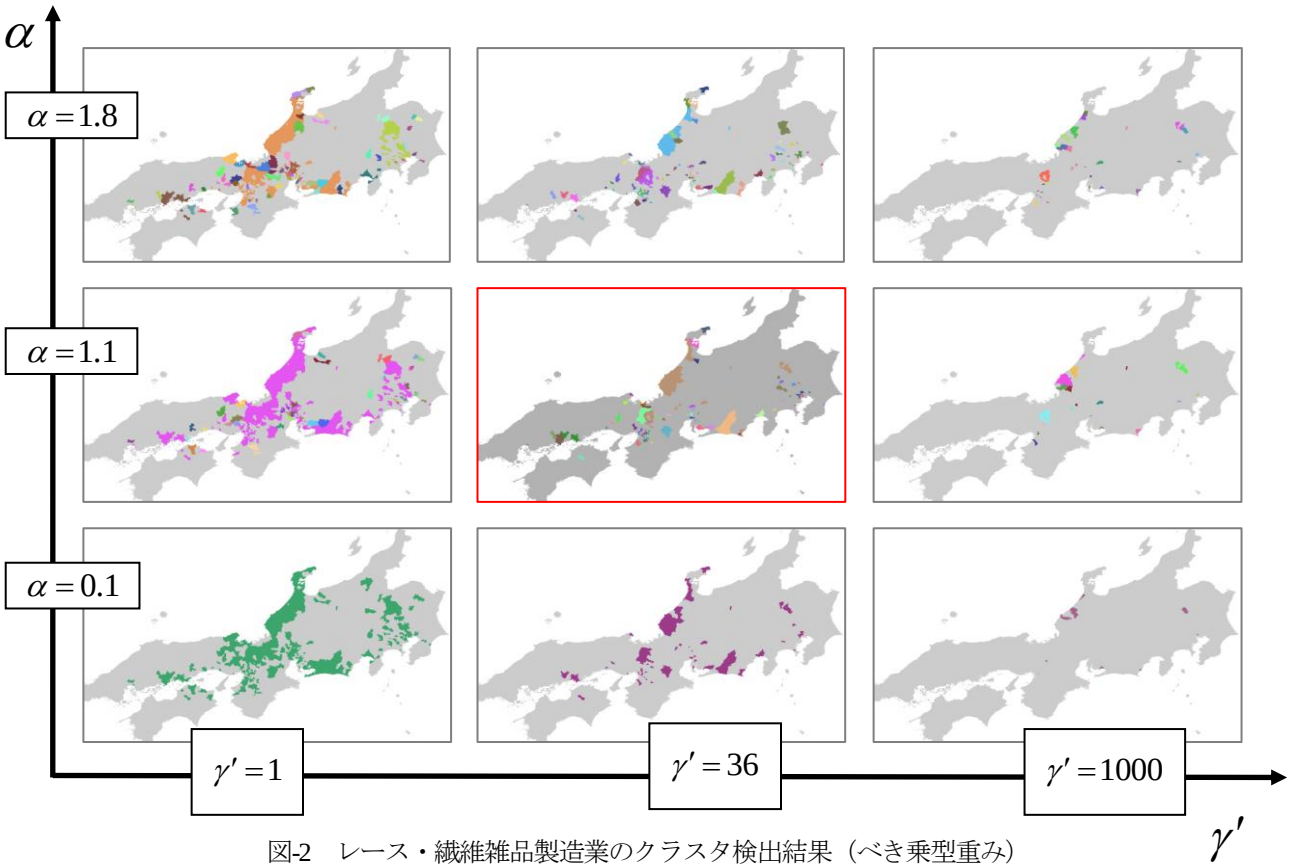

図-2　レース・繊維雑品製造業のクラスタ検出結果（べき乗型重み）

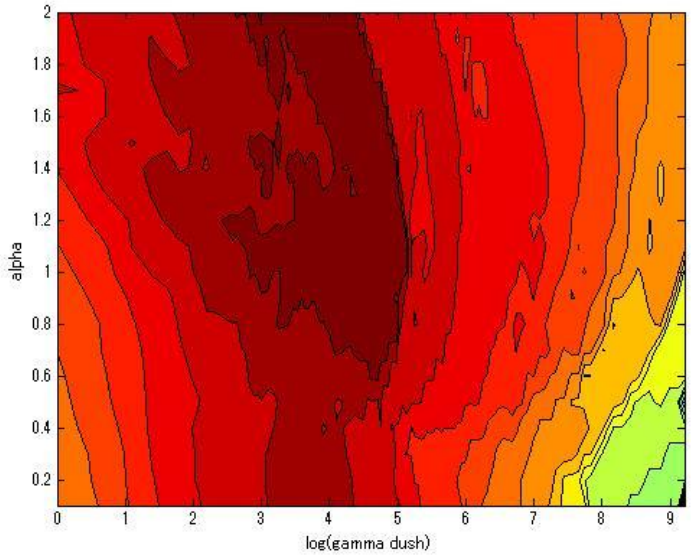

図-3　レース・繊維雑品製造業のクラスタ検出結果に対する BIC（べき乗型重み）

(1)　べき乗型の重みを用いる場合

式(3)で空間重みを定義した場合のクラスタ検出結果を図-2，BIC の計算結果を図-3に示す．図-2より，α が大きくなるほど小規模のクラスタが多数検出されること，γ' が大きいほど検出されるクラスタに含まれる市区町村数が少なくなることが観察できる．その理由は，モジュラリティの定義式を変形することで理解できる．モジュラリティの定義式より，2つの市区町村 i, j がは以下の条件を満たす場合，結合して同一のクラスタを形成しやすい．

$$W_{ij}\frac{x_i}{e_i}\frac{x_j}{e_j} > \gamma\frac{X(X-1)}{E^2} \qquad (11)$$

右辺は市区町村によらない定数である．これより，γ が大きいほど市区町村 i, j がクラスタを形成しにくくな

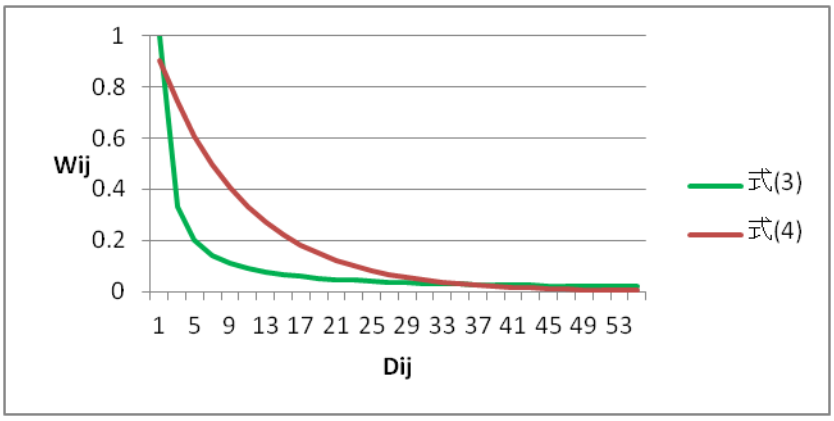

図-4　空間重みの空間距離による低減

り，最終的に検出されるクラスタに含まれる市区町村数が少なくなることがわかる．一方，α が大きいほど W_{ij} が小さくなるため，地理的に離れた市区町村ほど結合しにくくなることもわかる．

図-3からは読み取りにくいが，BIC は α =0.8~1.4，log γ' =3~5 において最も大きい値を示している．これらに対応する産業クラスタの検出結果は，図-2の赤枠で囲んだ結果に対応する．この結果を見ると，事業所の集積が見られる石川県・福井県周辺や滋賀県周辺，静岡県・愛知県周辺にクラスタが検出されており，本手法が産業クラスタ検出手法としておおむね良好であることが確認できる．ただし，細かく見ると，福井県周辺の市区町村と栃木県足利市が同一のクラスタを形成している一方で，足利市周辺の市町村が足利市とは異なるクラスタに所属しているなど，直感的な理解とは異なる検出結果になっていることがわかる．その理由は，分析に用いた空間重みの距離 D_{ij} による減衰率が不十分だったためであると考えられる．べき乗の空間重み行列を用いると，福井県の市区町村と栃木県足利市のように遠方に位置する場合でも W_{ij} が0よりも有意に大きい．その結果，それぞれの市区町村内の事業所数が大きければ式(11)の条件が成り立ち，結合して同一のクラスタを形成してしまう．さらに，福井県の市区町村と足利市が結合してしまい，（足利市ほど事業所数が多くない）足利市周辺の市区町村と福井県の市区町村間では式(11)の条件が成り立たないため，足利市と足利市周辺の市区町村が異なる産業クラスタに所属する結果となる．

(2)　指数型の重みを用いる場合

べき乗型の重みを用いた場合に，地理的に離れた福井県と栃木県足利市の市区町村が同一のクラスタを形成した理由は，空間重みの距離 D_{ij} による減衰率が不十分だったためである．式(4)の指数型の空間重みは，べき乗型の空間重みと比較して，近接する市区町村間の重みが相対的に大きくなり，遠方に位置する市区町村間の重みがゼロに収束しやすい（図-4を参照）．そのため，べき乗型を使用した場合に生じた問題は回避できると考えられる．

式(4)の指数型の空間重みを用いた場合のクラスタ検出結果を図-5，BIC計算結果を図-6に示す．べき乗型の空間重みを用いた場合と同様に，α が大きいほどクラスタ同士が互いに結合せずに細分化されること，γ' が大きいほどクラスタの検出される範囲が小さくなることが観察できる．また，図-6からBICは log α = -2~ -4，log γ' =2~5において大きな値を示していることが分かる．最も大きいBICを示した結果を図-5において赤い枠で囲まれた結果として示している．

BICを最大化する検出結果において，クラスタが細分化された理由は，BICの定義式上，事業所数に差のあるクラスタ同士は結合しない方が，BIC値が大きくなるからである．例えば，福井県から石川県かけての一帯は事業所数が相対的に多く立地しており，産業集積を形成している．ただし，図-1の[事業所数 / 経済面積]のコロプレスマップから分かる通り，市区町村で単位経済面積当たりの事業所数が異なる．その結果，一つのクラスタを形成するより，複数のクラスタに分割した方が，BICが大きくなる．

一方，図-5で α =1/900の場合の検出結果に目を向けると，北陸や京阪，備後，下野などの地域ごとにまとまった産業集積が検出されていることがわかる．これらは，直感的には産業クラスタとして解釈しやすい結果である．パラメータの設定により，本研究で提案した産業クラスタ検出手法でも直感的に解釈しやすい産業クラスタの検出結果が得られることが分かる．パラメータの設定法については，BIC基準が必ずしも有効であるとは限らないため，別の手法の利用可能性を検討していく必要がある．

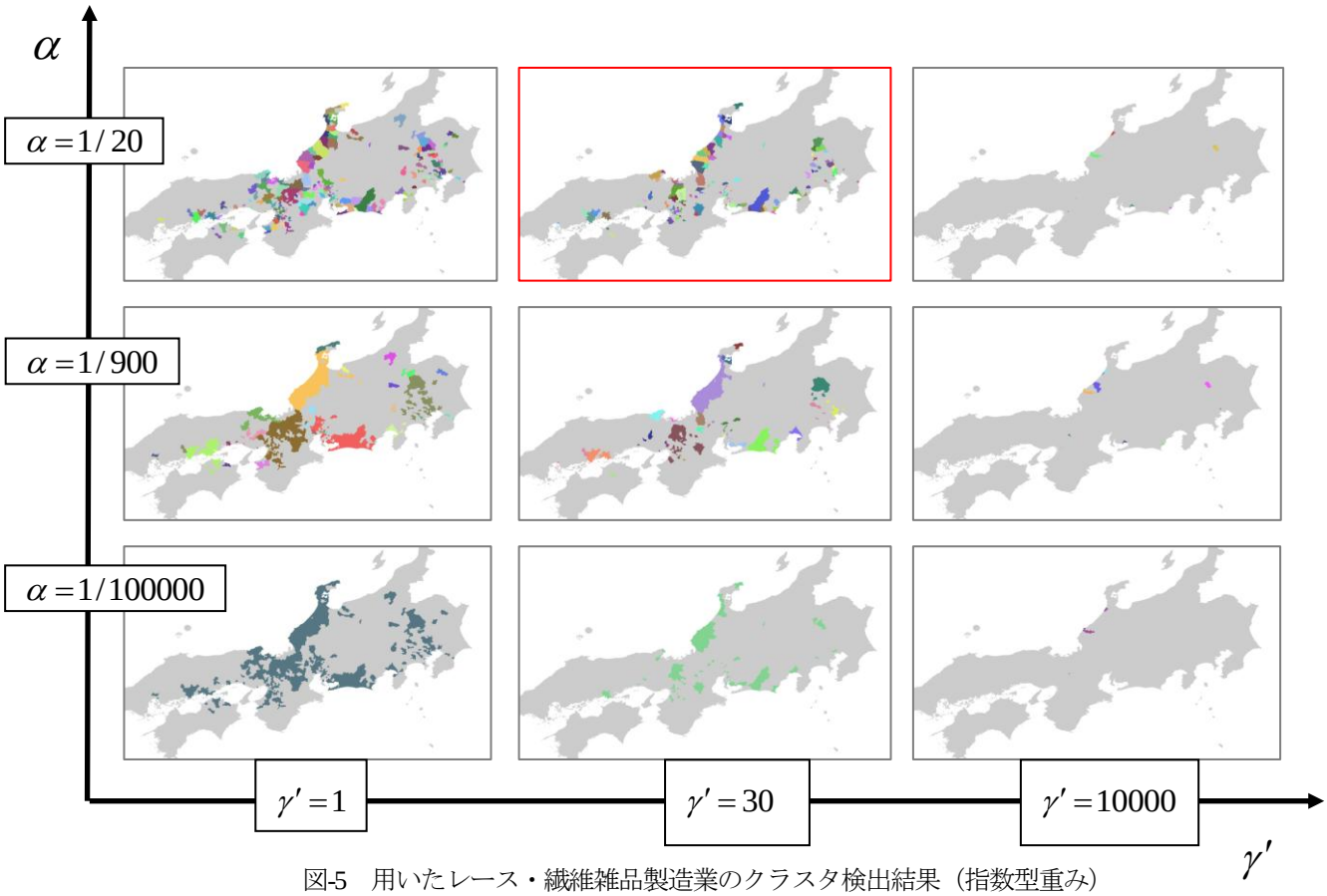

図-5　用いたレース・繊維雑品製造業のクラスタ検出結果（指数型重み）

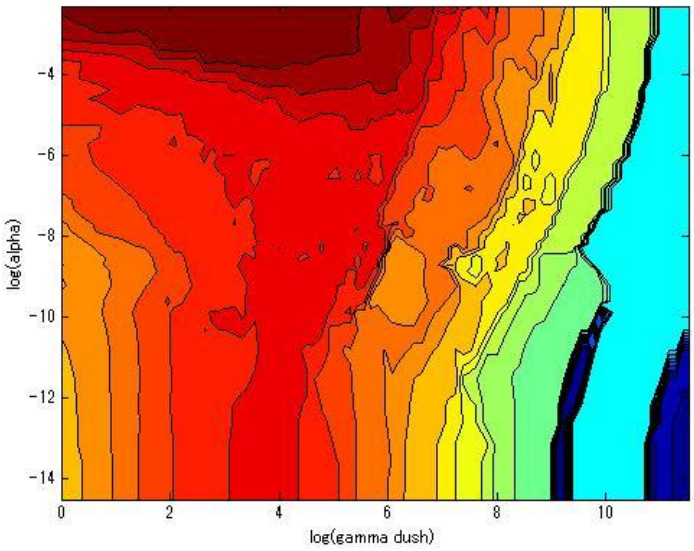

図-6　レース・繊維雑品製造業のクラスタ検出結果に対する BIC（指数型重み）

5.　おわりに

本研究では，地理的単位ごとの事業所数と地理的単位間の空間距離を用いて定義したネットワークに対してコミュニティ抽出法を適用することで産業クラスタを検出する手法を提案した．ケーススタディを通して，コミュニティ抽出法に基づく産業クラスタの検出手法の有効性を確認した．空間重み行列の定式化を工夫することで，直感的にもっともらしい結果を得ることができるという知見も得た．

ただし，パラメータをどのように設定すべきであるかという点が今後の課題として残された．本研究では，BICに基づくパラメータの設定法を検討した．しかし，BICの性質上，事業所が集積しているクラスタ候補地域であっても，地域内で事業所数が異なる場合には，クラスタが細分化されてしまうという結果が得られた．この点については，BICの定式化の変更や，事業所分布の生成プロセスとして用いた確率分布（多項分布モデルの見直し）の見直し，BIC基準とは異なるパラメータの設定方法などを検討していく必要がある．

参考文献

1) Mori, T. and Smith, T. E.: A probabilistic modeling approach to the detection of industrial agglomerations, 2010.
2) 栗田治，越塚武士: 領域間平均距離の近似理論とその応用，都市計画論文集，No.23，pp.43-48, 1988.

（2013. 5.7修正）