

明示的な形状制約を必要としない空間的クラスタリング手法の開発

東北大学 学生会員 ○氏家 晃仁
東北大学 正会員 福本 潤也

1. はじめに

近年、町丁目単位や小メッシュ単位で集計された詳細な空間データの利用が容易になっている。本研究ではこうした詳細な空間データから集積領域を検出する問題に着目する。

既存の集積検出手法では、集積領域について円形や凸性等の形状制約を置いたり、集積領域を構成する地理的単位間の連結性を仮定している。集積領域の連結性の仮定が妥当ではない詳細な空間データに既存手法を適用すると、集積でない地理的単位が集積領域に含まれたり、本来1つの集積領域になるべき地理的単位が複数の集積領域に別れる恐れがある。しかし、幾何形状に関する仮定を置かなければ、空間的にまとまった集積を検出できない。

本研究では空間データの生成過程を確率モデルで表し、モデルベースクラスタリングの枠組みを通じて集積領域を検出する手法を開発する。空間データの生成過程を表す確率モデルの中で集積領域の幾何形状への仮定を表現し、厳格な形状制約や連結性を仮定しない集積検出手法を開発する。

2. 既存研究

代表的な集積検出手法として、Kulldorff and Nagarwalla¹⁾が提案した空間スキャン統計がある。Kulldorff and Nagarwalla の手法は集積領域の幾何形状を円形と仮定し、統計的に厳密に集積を1つ検出する。この手法には、1)円形の集積しか検出できない、2)複数の集積の検出に向かない、という課題がある。これに対し、Duczmal and Assunção²⁾等によって形状制約のない空間スキャン統計が提案されている。さらに、Mori and Smith³⁾や井上・糟谷⁴⁾は複数の集積を検出でき、集積領域の幾何形状への制約を緩めた手法を提案している。しかし、いずれの手法も集積領域を構成する地理的単位間の厳密な連結性を仮定している。

3. 提案手法

画像処理分野でモデルベースクラスタリングによって領域分割を行う研究が進められている (e. g. Cucala and Marin⁵⁾)。本研究ではモデルベースクラスタリングの枠組みに基づく集積検出手法を開発する。

N 個の地理的単位から成る地理空間を考える。地理的単位 i の観測値を f_i で表す。観測データ全体を $f = (f_1, \dots, f_N)^T$ で表す。分析範囲に集積領域以外の領域（以下、非集積領域）1つと K 個の集積領域が存在すると仮定する。地理的単位は $(K+1)$ 個の領域の1つに所属する。地理的単位 i が所属する領域を a_i で表し、 $a = (a_1, \dots, a_N)^T$ と定義する。 a をクラスタ分割と呼ぶ。データ f は次の過程を経て生成されると仮定する。

- 1) あるクラスタ分割 a が事前分布 $\Pr(a|\alpha)$ に従って実現する。（ α はパラメータ）
- 2) 実現したクラスタ分割 a の下で、尤度 $\Pr(f|a, \theta)$ に従ってデータ f が生成される。（ θ はパラメータ）

本研究では、モデルベースクラスタリングにおけるモデル選択基準として Biernacki⁶⁾が提案した ICL を最大にするクラスタ分割を求めることで集積領域の検出を行う。本研究の集積検出問題は式(1)で定式化される。

$$(a^*, \alpha^*, \theta^*) = \arg \max_{a, \alpha, \theta} ICL(a, \alpha, \theta) = \ln p(f|a, \theta) \Pr(a|\alpha) - v/2 \ln N \quad (1)$$

v はモデルに含まれるパラメータの総数である。式(1)は地理的単位の組合せ最適化問題であり、大域的最適解を求めることは困難である。本研究では遺伝的アルゴリズムを用いて近似的に式(1)の問題を解く。

事前分布は Q 状態ポッツモデルを用いて式(2)で定義する。

キーワード 空間集積検出, 形状制約, モデルベースクラスタリング

連絡先 〒980-8579 仙台市青葉区荒巻字青葉 6-6-06 人間・環境系教育研究棟 406 号室 TEL 022-795-7498

$$\Pr(a|\alpha) = \frac{1}{Z(\alpha)} \exp \left(\sum_{k \in \{0,1,\dots,K\}} \sum_{(i,j) \in L_{C_k}} \frac{1}{2} \alpha w_{ij} - \text{PenaltyA} - \text{PenaltyB} \right) \quad (2)$$

$k (\in \{0,1,\dots,K\})$ は領域番号, L_{C_k} は k 番の領域に所属する地理的単位をノード, 地理的単位間の距離をリンクの重みとするネットワークの最小全域木 T_k を成すリンクの集合, α はパラメータ, w_{ij} は地理的単位 (i, j) 間の空間重み係数である. PenaltyA は距離を隔てた地理的単位が同じ集積領域に所属するクラスタ分割の事前確率を小さくするペナルティである. 各集積領域の最小全域木 T_k に含まれるリンクの長さが一定値を超えるとペナルティを課す. PenaltyB は多数の集積領域が入り乱れたクラスタ分割の事前確率を小さくするペナルティである. 集積領域毎に時間距離ネットワーク上の凸包を求め, 凸包の重複部分にペナルティを課す. 指数の第一項は近接した地理的単位が同じ領域に所属しやすいことを表す. 近接した地理的単位が同じ領域に所属し, ペナルティが 0 のクラスタ分割ほど事前確率が大きくなる.

尤度は式(3)の観測値間の独立性を仮定した多変量正規分布で定義する.

$$p(f|a, \mu_k, \sigma) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(f_i - \mu_{q_i})^2}{2\sigma^2} \right\} \quad (3)$$

μ_{q_i} と σ はパラメータである. 式(3)は観測値が類似する地理的単位が同じ領域に所属するほど大きな値を取る.

4. ケーススタディ

平成 21 年経済センサスの 2 分の 1 地域メッシュの全産業事業所密度を観測データとした. 地理的単位間の距離はメッシュの重心間の直線距離とする. 重心間の距離が d_{ij} のメッシュのペアに対して空間重み係数 w_{ij} を $w_{ij} = 1/d_{ij}^2$ と設定する. また, 今回は事前分布のパラメータ α を $\alpha = 2$ と与えて分析した.

図-1 に東京都都心部の 483 メッシュの事業所密度を示す. 黒の濃いメッシュほど事業所密度が高い. 鉄道路線網を併せて示した. 図-2 は図-1 に示した分析範囲から集積領域を検出した結果である. 同じ模様のメッシュが同じ領域に所属する. 無地のメッシュは非集積領域に所属する. 図-2 より, 次の 2 点に分かる. 第一に, 連結していない集積を検出できる. 東京駅周辺の集積領域が事業所密度の低いメッシュを含まず, 連結していない. 第二に, 複数の集積を検出できる. 東京駅周辺と新宿駅周辺の 2 つの集積領域が検出されている.

5. おわりに

本研究では厳格な形状制約や連結性を仮定しない集積検出手法を開発した. ケーススタディを通じて, 本手法により地理的単位が連結していない集積を検出できることを確認した. 今後, より広範囲のデータを用いたケーススタディを蓄積していく予定である.

参考文献

- 1) Kulldorff, M. and Nagarwalla, N. : Spatial disease clusters: Detection and inference, *Statistics in Medicine*, Vol. 14, pp. 799-810, 1995.
- 2) Duczmal, L. and Assunção, R. : A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters, *Computational Statistics & Data Analysis*, Vol. 45, Issue 2, pp. 269-286, 2004.
- 3) Mori, T. and Smith, T. E. : A probabilistic modeling approach to the detection of industrial agglomerations, *KIER Discussion Paper*, 2010.
- 4) 井上亮, 糟谷志帆: 複数の柔軟な形状を持つ点事象集積地域の検出法の開発, 土木計画学研究・概要集, Vol. 49, CD-ROM, 2014.
- 5) Cucala, L. and Marin J.-M. : Bayesian inference on a mixture model with spatial dependence, *Journal of Computational and Graphical Statistics*, Vol. 22, Issue 3, pp. 584-597, 2013.
- 6) Biernacki, C., Celeux, G. and Govaert, G. : Assessing a mixture model for clustering with the integrated completed likelihood, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 22, Issue. 7, pp. 719-725, 2000.

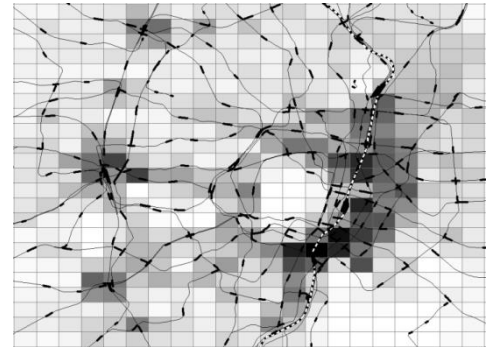


図-1 事業所密度

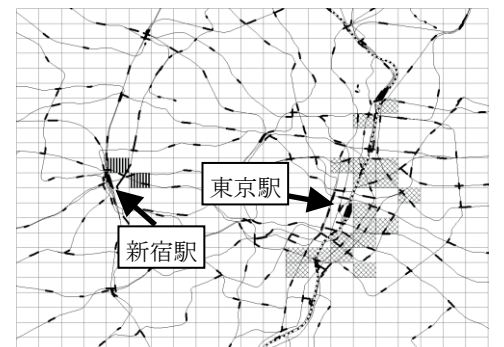


図-2 集積検出結果