人流データのクラスタリング手法に対する一考察

藤原直哉

On a Clustering Method of Human Mobility Data

Naoya FUJIWARA

Abstract: 時間および空間分解能が高い地理空間データを用いて，恣意性の少ない手法で圏域を抽出することは重要な課題である．先行研究により，人流データを用いて複雑ネットワークのクラスタリング手法によって自動的に階層的圏域構造の抽出が可能であることが知られている．本研究では，集計に用いるメッシュサイズに対するクラスタリング結果を簡単なモデルと人の流れデータを用いて検討し，手法の発展に関する基本的知見を得る．

Keywords: 人流データ（human mobility data），ネットワーク（network），クラスタリング (clustering)

1. はじめに

近年，スマートフォンの普及などによって，大規模かつ長期間にわたる詳細な人の行動データが取得可能になりつつある．それらのデータを解析するための情報学的手法，および解析結果によって明らかになる知見は地理情報科学において非常に重要である．とりわけ，都市圏，通勤圏などの圏域の設定は古くから研究が行われている重要な課題であるが（金本・徳岡，2002，山田・徳岡，1983，森川，1990，Kawashima et al.，1993，駒木，2004），これらの詳細な人流データを有効活用できれば，これまでよりも詳細な解析が可能であると期待される．

本研究は，先行研究によって優秀な性能を示したネットワーククラスタリングの手法について詳しく解析を行い，手法の持つ特徴，問題点，そして将来の改善点を明らかにすることを目的とする．

藤原直哉　〒277-8565　千葉県柏市柏の葉 5-1-5
東京大学空間情報科学研究センター
Phone：04-7136-4297
E-mail：fujiwara@csis.u-tokyo.ac.jp

2. 手法の概略

2.1　複雑ネットワークのクラスタリング手法

圏域設定の問題は，数理的には，集計単位をクラスタリングする問題であると考えられる．特に，集計単位をノードとし，トリップの存在するノード間をリンクでつないだネットワークであると考えることが可能である．それゆえ，圏域設定の問題は人の流れによるネットワークのクラスタリングの問題であると解釈できる．ネットワーククラスタリングは，コミュニティ分割とも呼ばれており（Fortunato，2010），複雑ネットワーク科学の発展に伴い，近年盛んに研究されている．

2.2　Infomap

ネットワーククラスタリングとして，様々な手法が提案されており，問題の性質に応じて適切な手法を適用する必要がある．都市圏における人流は動的な性質なので，ネットワーク上での流れを用いたクラスタリング手法が適切であると考え

られる．

Rosvall & Bergstrom (2008)によって提案されたInfomap は，ネットワーク上のランダムウォークに基づいたクラスタリング手法である．近年，ベンチマークテストにおいてもよい成績をあげており，本研究ではこの手法を用いる．

Infomap では，ネットワーク上のランダムウォーカーの軌跡を符号で記述する．この記述長を最小化するようにクラスタ構造を決めることがInfomap の基本的な考え方である．最も重要なアイディアは，クラスタ構造を考え，ランダムウォーカーの位置を，所属するクラスタの符号およびクラスタ内の符号で表現することで，記述長を短縮することができるという点にある．具体的には，1 ステップあたりの記述長（Map equation）

$$L(M) = -q\,H(Q) - \sum_{\{i=1\}}^{m} p_i H(P_i)$$

を最小化するクラスタ構造を求めることになる．ここで，定常状態におけるランダムウォーカーのノードαへの滞在確率をp_α，ランダムウォーカーがクラスタiへ流入するイベントの発生確率をq_i，$q = \sum_{i=1}^{m} q_i$とし，

$$H(Q) = \sum_{i=1}^{m} \frac{q_i}{q} \log\left(\frac{q_i}{q}\right)$$

$$H(P^i) = \frac{q_i}{q_i + \sum_{\beta\in i} p_\beta} \log\left(\frac{q_i}{q_i + \sum_{\beta\in i} p_\beta}\right) + \sum_{\alpha\in i} \frac{p_\alpha}{q_i + \sum_{\beta\in i} p_\beta} \log\left(\frac{p_\alpha}{q_i + \sum_{\beta\in i} p_\beta}\right)$$

である．この手法の拡張として階層的なクラスタリングも提案されている（Rosvall & Bergstrom, 2011）．クラスタ数m及び階層数は記述長を最小化するものとして自動的に決定される点も，この手法の利点である．

2.3　人流ネットワークへの適用時の問題点

本研究では，本来は連続的な人流データを集計して離散的なネットワークを構成した．そのため，クラスタリング結果が集計単位の大きさに依存する可能性がある．このような，離散的なネットワークにおいては通常考慮されない，離散化に伴う問題が地理的な問題には常に存在する．逆に，集計単位に極力依存しない情報量規準を提案することができれば，ネットワーククラスタリングの適用可能な対象が広がる可能性があり，さらなる研究が必要である．本研究では，集計単位の大きさを変化させた時の記述長の変化を考え，その足がかりとする．

3. ネットワーク分割に対する記述長の変化

3.1　簡単な例

前節で定義した Infomap の記述長の，集計メッシュサイズに対する依存性について議論する．本節では簡単な例として，m個のクラスタからなるネットワークを考える．図-1 では$m = 2$の場合を示す．本稿では，各クラスタがそれぞれ内部構造を有すると仮定し，それぞれのノードを分割することを考える．ここでは，分割によって Infomap の記述長がどのように変化するかを考察する．なお，ネットワークがさらに小規模な内部構造を入れ子的に持っているとすると，ネットワークはフラクタル性を持つ．

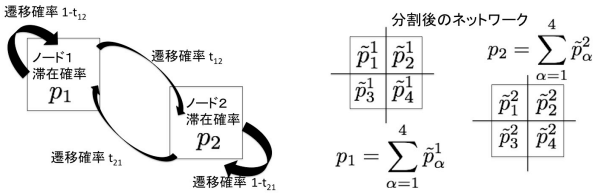

図- 1　ネットワーク分割の簡単なモデル．分割前（左）のノードが分割後（右）のクラスタとなり，各クラスタが４分割される．

分割していないネットワークの，定常状態でのノードiにおけるランダムウォーカーの滞在確率をp_i，ノードiからjへの遷移確率をt_{ij}と書く．分割していないネットワークの記述長は，

$q_i = \sum_{j \neq i} t_{ji} p_j = (1 - t_{ii}) p_i$に注意すると，それぞれ $H(Q) = \sum_{i=1}^{m} \frac{(1-t_{ii})p_i}{\sum_{j=1}^{m}(1-t_{jj})p_j} \log\left(\frac{(1-t_{ii})p_i}{\sum_{j=1}^{m}(1-t_{jj})p_j}\right)$，

$H(P^i) = \frac{1-t_{ii}}{2-t_{ii}} \log\left(\frac{1-t_{ii}}{2-t_{ii}}\right) + \frac{1}{2-t_{ii}} \log\left(\frac{1}{2-t_{ii}}\right)$となる．

次に，分割したネットワークを考える．クラスタiに属するノードαにおける定常状態での滞在確率を$\tilde{p}_\alpha^i$とおき，このノードからクラスタjに属するノードβへの遷移確率を$\tilde{t}_{\alpha\beta}^{ij}$と書く．ここで，簡単のため，クラスタ間の遷移確率は一様$\tilde{t}_{\alpha\beta}^{ij} = \tilde{t}_{\gamma\beta}^{ij} = t_{ij}\tilde{p}_\beta^j / p_j$とし，クラスタ内の遷移確率は発ノード$\alpha$と$\beta$が同一クラスタに属する場合，$\tilde{t}_{\alpha\beta}^{ii} = t_{ii}\,\tilde{p}_\beta^i / p_i$とする．すると，$t_{ii} = \sum_{\beta \in i} \tilde{t}_{\alpha\beta}^{ii}$が成り立つ．遷移確率をこのようにとることで，分割した場合の滞在確率と元のネットワークの滞在確率の間に，$p_i = \sum_{\alpha \in i} \tilde{p}_\alpha^i$なる関係が成り立つ．トリップ数で遷移確率を決定する場合，このような配分は自然である．

分割ネットワークの記述長は，$\tilde{H}(Q) = H(Q)$および$\tilde{H}(P^i) = H(P^i) + \frac{1}{2-t_{ii}}[\sum_\alpha \frac{\tilde{p}_\alpha^i}{p_i} \log(\tilde{p}_\alpha^i) - \log p_i]$である．すなわち，クラスタ構造が変化しない限り，ネットワーク分割によって$H(Q)$は変化しない．クラスタ内の移動を記述する記述長のうち，クラスタを退出する情報についても，クラスタ構造が変化しなければ変化しない．一方，クラスタ内の移動に関する記述長は増加している．1つのメッシュをn個のメッシュに分割する際，記述長が最も増加するのは，n個のメッシュに等確率でランダムウォーカーが存在する場合であり，この時の記述長の増加量は$\frac{1}{(2-t_{ii})n} \log n$である．

以上の考察により，集計単位を変更することによって，Map Equationの各項の変化のしかたが異なることにより，クラスタリング結果が変わる可能性が存在することがわかった．

3.2　人の流れデータを用いた解析

さらに，人の流れデータにInfomapの手法を用いてクラスタリングを行った．筆者らは，全国規模でGPSデータを用いた場合のクラスタリングをすでに報告しているが（桜町ほか，2015），GPSデータでは，利用者にバイアスがある，個人の属性や移動目的などが不明である，などの問題点がある．よって，パーソントリップ調査に基づいた人流データを用いクラスタリングを行い，GPSデータでの結果と比較することは有意義である．

京阪神の人の流れデータを用い，3次メッシュを集計単位として関西圏のクラスタリングを行った結果（図-2）は桜町ほか（2015）で報告されたクラスタリング結果と大きく異ならないものであり，データセットの差異によってクラスタリング結果は大きく変わらないことがわかった．

一方，集計単位を2次メッシュとした場合（図-3）には最上位階層（図-3左）のクラスタリング結果が3次メッシュを集計単位とした場合（図-2左）と異なる箇所が存在する．

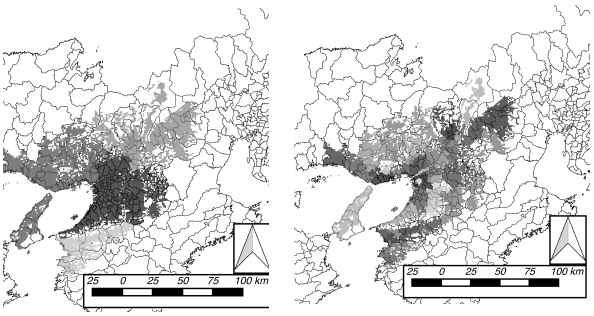

図-2　関西の3次メッシュで集計した場合のクラスタリング結果．左が最上位階層の4つのクラスタ，右が第2階層のクラスタ．

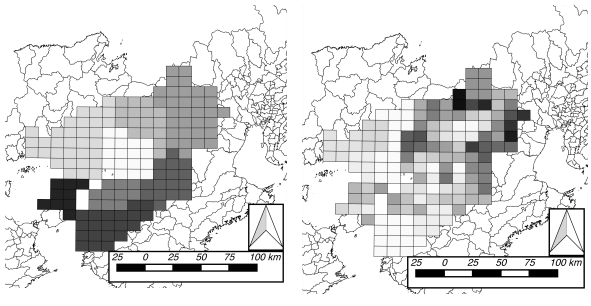

図-3　2次メッシュで集計した場合のクラスタリング結果．最上位階層が7つのクラスタに分割される．

4. まとめと考察

本研究では，人流データを用いた地域のクラスタリングにおいて，空間的に連続な構造を離散化することによって発生する問題について考察を行った．特に， Infomap の記述長を表す式の各項ごとに集計単位によって振る舞いが異なる具体例を構成し，人の流れデータのクラスタリングにおいて，結果がメッシュサイズに依存する例を示した．その一方で，Infomap はネットワーククラスタリング手法としては優秀であることが知られており，事実，人の流れデータに適用した際にも感染症拡大パターンを近似できるなど，様々な分野への応用可能性がある．本研究で明らかになった問題点を解決し，メッシュサイズ依存性が少ない情報量規準を提案するなど，クラスタリング手法を改善することは今後の重要な課題である．

謝辞

本研究は JSPS 科研費 15K16061，および公益財団法人二十一世紀文化学術財団の学術奨励金の助成を受けた．また，本研究は，東京大学 CSIS 共同研究（No. 315）による成果である．

参考文献

金本良嗣・徳岡一幸（2002）：日本の都市圏設定基準，応用地域学研究，7，1-15.

山田浩之・徳岡一幸（1983）：わが国における標準大都市雇用圏：定義と適用－戦後の日本における大都市圏の分析（2）－，経済論叢，132，3・4，145-173.

森川洋（1990）：広域市町村圏と地域的都市システムの関係，地理学評論，63，A-6，356-377.

Kawashima, T., Hiraoka, N., Okabe, A. and Ohtera, N., 1993. *Metropolitan analysis: Boundary and future population changes of functional urban regions. Gakushuin Economic Papers*, **29**, 3-4, 205-248.

駒木伸比古（2004）：通勤・消費行動からみた東京大都市圏の空間構造，日本地理教育學會會誌，52，1，1-15.

Fortunato, S., 2010. *Community detection in graphs. Physics Reports,* **486**, 75-174.

Rosvall, M. and Bergstrom, C. T., 2008. *Maps of information flow reveal community structure in complex networks. Proceedings of the National Academy of Science*, **105**, 1118-1123.

Rosvall, M. and Bergstrom, C. T., 2011. *Multilevel Compression of Random Walks on Networks Reveals Hierarchical Organization in Large Integrated Systems. PLoS ONE,* **6**, e18209.

桜町律・藤原直哉・藤嶋翔太・秋山祐樹・柴崎亮介（2015）：人の流れを考慮した都市圏の定義手法に関する研究，第 24 回地理情報システム学会講演論文集.