D-005

分布間距離に基づく類似分布構造を有する地域の抽出
Extraction of Regions with Similar Distribution Structure Based on Distribution Distance

吉田 純 † 伏見 卓恭 †
Jun Yoshida Takayasu Fushimi

1. はじめに

観光は近年著しく成長を遂げており，特に訪日外国人観光客の数は年々増加傾向にある．これらの観光客を都心部だけでなく地方へ誘致するために，各地域ならではの特徴を明らかにし，地域ブランドを創出することは重要な課題である．しかし，行政による地域ブランディングは行政区画を単位としたものが多く，地域をまたがった施策を実施している地域は多くない．行政区分をまたぎ連続性を持った地域において行政区分に固執することは，商業集積機能の分断により効果的な運営を阻害するなどの問題がある．たとえば複数の市にまたがっている湘南エリアでは，海産物が豊富に取れることにより，海産を売りにした飲食店が多く分布している．しかし，複数の関連自治体が協力しあっているという事実はない．逆に，横浜市のように中華街エリアや新横浜エリアなどの複数のブランディングする小地域を抱える市では，市全体ではなく細かいエリアごとでの政策が必要である．まちづくりや地域活性化の計画策定にあたり，行政間協議会の発足などにより一体的な中心市街地活性化を行えるような法的整備も必要である．

そこで本研究では，行政区分にとらわれず，類似する特徴を有する近隣地域を意味する等質地域を抽出する手法を提案する．小さな地域においては，そこでの分布数は全体的に少ない傾向にあるため，大きな地域における分布数との格差により，特性が埋もれてしまう場合がある．そこで，絶対数の規模の格差に隠れてしまう各地域の特性を顕在化させために，飲食店が地域とカテゴリに独立して分布している場合の期待値と実際の分布数である実測値の差を用いて Z スコア計算し，出店数の少ない地域やカテゴリであっても，期待値に比べて偏って多くなっていれば，出店数の格差に隠れた特性を強調する手法を提案する．提案手法では，全体の分布と比較してその地域に統計的有意に多く存在する特徴量により各等質地域をラベリングする（1 参照）．ぐるなび API を用いて取得した飲食店データを用いた評価実験により，行政区分と比較して Z スコアの平均値が高くなるような地域分類ができていること，絶対数では抽出できなかった各地域特有のカテゴリを抽出できていることを確認する．

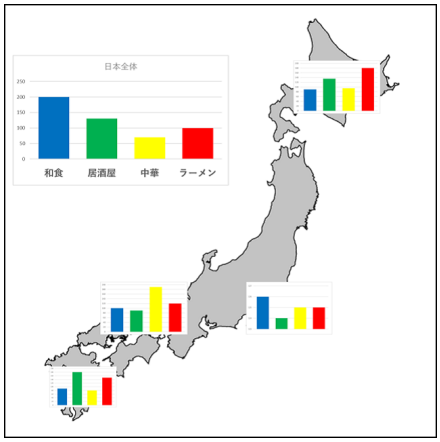


図 1: 全体の分布と地域ごとの分布

2. 関連研究

地域特性を抽出する研究として，SNS 投稿写真の画像特性を抽出し，地域の特徴記述を行い地域間の類似度を求める研究がある [1]．各地域の特性を写真の特徴量を要素としたベクトルを用いて表現している．提案手法では各地域の特性を出店している飲食店のカテゴリ分布から算出した Z スコアで表現している点で異なる．また，位置情報付きコンテンツから地域限定語句を抽出する研究がある [2]．TFIDF をベースとした指標を用いて，特定の地域にのみ出現する単語を抽出している．

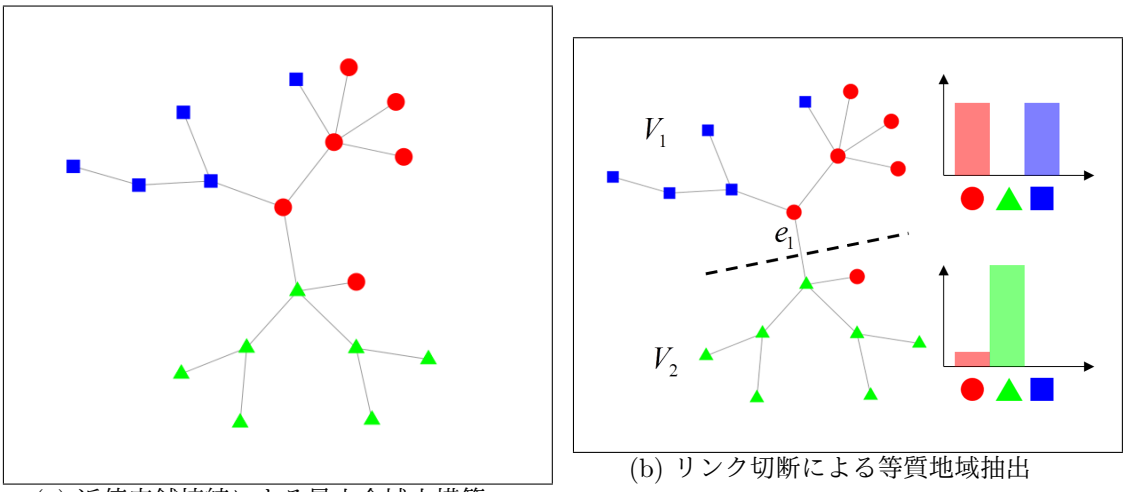
3. 提案手法

提案手法では，共通の特徴を有する隣接エリアを 1 つの地域とした等質地域を抽出し，各地域において特徴的な属性を抽出することで地域をラベリングする．提案手法の手順は以下の通りである：

1. 位置情報から最小全域木を構築；
2. 最小全域木のリンクを切断し，等質地域を抽出；
3. 各地域内における属性分布の偏り度を算出;

本稿では，飲食店の緯度経度を位置情報，飲食店が属するカテゴリを属性として説明する．

† 東京工科大学コンピュータサイエンス学部



(a) 近傍店舗接続による最小全域木構築　(b) リンク切断による等質地域抽出

図 2: 提案手法の処理手順

3.1. 最小全域木の構築

まず N 個の飲食店の集合 $\mathcal{V} = \{v_1, \ldots, v_N\}$ および，これらの位置情報の座標ベクトル群 $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$，$\mathbf{x}_n \in \mathbb{R}^2$ から，図 2(a) のように，近傍に位置する飲食店をつなげることで最小全域木を構築する．最小全域木とは，与えられたオブジェクト集合の中で距離が最小のオブジェクト同士をリンクし，閉路を持つことなく全てのオブジェクト集合をリンクさせたグラフ（木構造）のことを指す．本稿では，飲食店の距離 $d(v_n, v_m)$ は，マンハッタン距離を用いる．すなわち，飲食店 v_n の座標ベクトル $\mathbf{x}_n$ の第 h 次元の値を $x_{n,h}$ とすると，飲食店 v_n と v_m 間の距離を

$$d(v_n, v_m) = \sum_{h=1}^{2} |x_{n,h} - x_{m,h}| \qquad (1)$$

で定義する．式 1 を用いて飲食店間の距離を計算し，最も近傍に位置する飲食店間にリンクを設定することで最小全域木 $G = (\mathcal{V}, \mathcal{E})$ を構築する．

3.2. 等質地域の抽出

次に，上記より得られた最小全域木 $G = (\mathcal{V}, \mathcal{E})$ をある共通特徴をもった K 個の部分集合（等質地域）に分割する．ここで，飲食店集合を等質地域に分割するための切断リンク集合を $\mathcal{E}_{K-1} \subset \mathcal{E}$ とする．k 番目の等質地域を $\mathcal{V}_k$ とすると，$\mathcal{E}_{K-1} = \{e_1, \ldots, e_{K-1}\}$ の要素 e_k は，地域 $\mathcal{V}_{k-1}$ と $\mathcal{V}_k$ を結ぶリンクを示す．よって切断リンク集合 $\mathcal{E}_{K-1}$ から等質地域群 $\{\mathcal{V}_1, \ldots, \mathcal{V}_K\}$ が決定される．飲食店 v_n は一つの地域にのみ属し，複数の地域に属することはないとする：

$$\mathcal{V} = \bigcup_{k=1}^{K} \mathcal{V}_k, \mathcal{V}_k \cap \mathcal{V}_j = \emptyset, \quad \text{for} \quad k \neq j.$$

提案手法では，尤度関数を用いてデータ集合の特徴や関係性に基づいて，等質地域に分割する．いま，飲食店 v のカテゴリを $v.cate$ とし，等質地域 V_a に属する飲食店のうち，カテゴリ c に該当する飲食店数を数え上げ

$$r_{a,c} = |\{v \in \mathcal{V}_a; v.cate = c\}|,$$

とし，地域 V_a 内の全飲食店の該当カテゴリ数の和を

$$R_a = \sum_{c \in \mathcal{C}} r_{a,c}.$$

のように求め，以下の尤度関数 $J(\mathcal{E}_K)$ を定義する：

$$J(\mathcal{E}_K) = \sum_{k=1}^{K} \sum_{c \in \mathcal{C}} r_{a,c} \log \frac{r_{a,c}}{R_a} \qquad (2)$$

尤度関数 2 の最大化により，飲食店集合をカテゴリ分布が大きく変化する箇所で分割でき，特徴的なカテゴリを有する地域の抽出が期待できる．

図 2 において，ノードのマーカー（●，▲，■）が該当する飲食店のカテゴリを表す．説明の簡略化のため，ここでは 1 つの飲食店は 1 つのカテゴリにのみ属するとする．図 2(a) の全域木の場合，尤度関数 $J(\mathcal{E}_1)$ はリンク e_1 で最適化され，飲食店集合は図 2(b) のような等質地域 $\mathcal{V}_1, \mathcal{V}_2$ に分割される．リンク e_1 で分割することにより，$\mathcal{V}_1$ のカテゴリ分布は●と■のみとなり，$\mathcal{V}_2$ は▲の割合が顕著に高くなる．

この結果，$\mathcal{V}_1$ は●と■を，$\mathcal{V}_2$ は▲を特徴的なカテゴリとする等質地域として分割されたことがわかる．このようにして，部分集合のカテゴリ分布が大きく変化する，または特徴的な分布となるような切断リンクを尤度関数 2 により求める．尤度関数 2 を最適化する切断リンク集合 $\mathcal{E}_K$ は貪欲法より求められ，以下にツリー分割による等質地域抽出のアルゴリズムを示す．

1. 初期化：$k = 1, \mathcal{E}_{k-1} = \emptyset$ ；
2. 抽出：$\hat{e}_k = \arg\max_{e \in \mathcal{E}} J(\mathcal{E}_{k-1} \cup \{e\})$ ；
3. 更新：$\mathcal{E}_k = \mathcal{E}_{k-1} \cup \{\hat{e}_k\}$ ；
4. 終了判定：
   - $k = K - 1$ なら $\{\mathcal{V}_1, \ldots, V_K\}$ を出力；
   - さもなければ $k = k + 1$ とし step2 へ戻る；

3.3. 特徴量の偏り度算出

次に，地域ごとカテゴリごとの飲食店数を集計することで Mixing Matrix $\mathbf{F} = [f_{a,c}]$ を構築する．ここで，$f_{a,c} = |\{v \in \mathcal{V}; v.area = a \wedge v.cate = c\}|/N$ は，全飲食店数 $N = |\mathcal{V}|$ に対する地域 a に出店するカテゴリ c の飲食店数の割合である．前述したように，地域 a において $f_{a,c}$ の値が大きなカテゴリ c により地域 a をラベリングするのは適切ではない．そこで，飲食店は地域とカテゴリに独立に分布していると仮定した場合の分布数の期待値 $e_{a,c}$ と実際の分布数の差を用いて Z スコア $z_{a,c}$ を計算する．まず，地域の集合を $\mathcal{A}$，カテゴリの集合を $\mathcal{C}$ とし，地域分布 $p_a = \sum_{c \in \mathcal{C}} f_{a,c}/N$ とカテゴリ分布 $q_c = \sum_{a \in \mathcal{A}} f_{a,c}/N$ を計算する．そして，飲食店の出店数は地域とカテゴリに独立であると仮定して地域 a にカテゴリ c の飲食店が存在する確率は $e_{a,c} = p_a \cdot q_c$ で計算できる．したがって，Z スコアは以下のように計算される：

$$z_{a,c} = \frac{N f_{a,c} - N e_{a,c}}{\sqrt{N e_{a,c}(1 - e_{a,c})}}. \qquad (3)$$

式 (3) の分母は標準偏差であり，Z スコアの値は独立性を仮定した際の期待値と比較して，どの程度有意に多くまたは少なく存在するかを表している．Z スコアが正で大きいほど，カテゴリ c の飲食店が地域 a に統計的有意に多く存在するといえる．Z スコアを用いることにより，出店数の少ない地域やカテゴリであっても，独立性を仮定した際に比べて偏って多く出店していれば大きな値となり，規模の格差により隠れてしまう地域・カテゴリ間の関係の強さを強調できる．

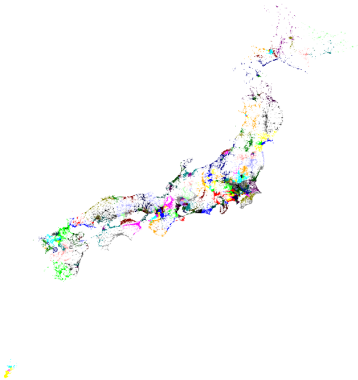


図 3: 提案手法による等質地域抽出結果（$K = 431$）

4. 評価実験

本研究では，ぐるなび API [‡]を用いて取得した全 565,810 件（全 178 カテゴリ）の飲食店データを対象とし，各地域の特性をラベリングする．そして，有名な地域を取り上げ，直感と合致した結果が得られているかを評価する．

表 1 に，関東地方の都県に対して単純な分布数が多い上位 3 つのカテゴリでラベリングした結果を示す．上述したように，日本全国で分布数の多い居酒屋，カフェ，定食がラベルとして用いられる傾向にあり，各地域の特性が現れていない．表 2 は，関東地方の都県に対する Z スコアの上位 3 カテゴリによるラベリング結果である．表 2 より，神奈川県が中華，群馬県がそばでラベリングされており，単純分布数による結果よりは妥当な結果が得られている．他の都県に関しては，必ずしも妥当とは言い難い．これは，地域の分類粒度が大きすぎるためと考えられる．

表 3 は，分類粒度を都道府県より細かくした地域に対するラベリング結果である．ここでは，1 位のカテゴリに対する Z スコアが高い順に表示している．表 3 を見ると，那覇が沖縄料理，京都の祇園が京料理，広島市がお好み焼き，仙台市が牛タンなど，直感に合致した結果が得られているか．さらに，日光が湯葉（単純分布数：16），廿日市市があなご料理（単純分布数：9）のように，絶対数が少なくあまり知られていないが，その地域に偏って分布するカテゴリによりラベリングできている．表 2 では，栃木の Z スコアの 1 位はそばであったため，地域のまとめ方によって Z スコアが上位に来るようなカテゴリが変わることがわかる．

図 3 に，ツリー分割によって抽出した等質地域を示す．行政区分の大分類による 431 地域と比較する．図 3 において，抽出した 431 地域ごとに異なる色で可視化しているが，色が足りないため，異なる地域として抽出されても同じ色

[‡]https://api.gnavi.co.jp/api/

表 1: 分布数によるラベリング（都道府県別）

$Nf_{a,c}$	1 位	2 位	3 位
茨城県	居酒屋:1538	定食:761	そば:649
栃木県	居酒屋:1105	定食:720	そば:704
群馬県	居酒屋:1006	定食:679	そば:620
埼玉県	居酒屋:3908	そば:1236	カフェ:1135
千葉県	居酒屋:3331	カフェ:1096	定食:1025
神奈川	居酒屋:5782	カフェ:2100	中華:1589

表 2: Z スコアによるラベリング（都道府県別）

$z_{a,c}$	1 位	2 位	3 位
茨城県	そば:17.72	ラーメン:11.39	定食:10.07
栃木県	そば:25.47	焼きそば:17.39	ラーメン:13.68
群馬県	そば:22.91	定食:13.24	焼きそば:10.53
埼玉県	そば:19.77	中華:14.43	うどん:12.32
千葉県	ファミレス:12.12	中華:11.26	そば:9.34
神奈川	中華:15.27	イタリアン:9.04	居酒屋:8.99

表 3: Z スコアによるラベリング（行政区分：大分類）

$z_{a,c}$	1 位	2 位	3 位
那覇	沖縄料理:102.16	バー:16.51	しゃぶしゃぶ:7.736
祇園・岡崎・清水寺	京料理:59.44	懐石料理:20.38	割烹:15.18
広島市	お好み焼き:53.22	広島風お好み焼き:41.39	鉄板焼き:17.99
日光・鬼怒川	湯葉料理:53.05	そば:16.67	定食:13.23
廿日市市・大竹市・宮島	あなご料理:44.78	お好み焼き:14.70	牡蠣料理:11.79
仙台市	牛タン:41.20	ショットバー:9.19	牡蠣料理:4.83

表 4: Z スコアによるラベリング（等質地域）

$z_{a,c}$	1 位	2 位	3 位
地域 95（月島近辺）	もんじゃ焼き：127.21	屋形船：15.34	お好み焼き：11.80
地域 429（那覇近辺）	沖縄料理：102.14	メキシコ料理：15.60	アメリカ料理：13.16
地域 428（那覇近辺）	沖縄料理：99.34	バー：19.90	しゃぶしゃぶ：9.03
地域 106（新大久保近辺）	韓国料理：89.44	サムギョプサル：41.68	ネパール料理：18.28
地域 430（那覇近辺）	沖縄料理：72.94	アメリカ料理：10.00	しゃぶしゃぶ：9.00
地域 431（那覇近辺）	沖縄料理：71.23	薬膳料理：9.49	カフェ：9.21

になっている部分もある．行政区分による 431 地域では，Z スコアの平均値が 4.8 であるのに対し，提案手法による等質地域は平均が 6.2 である．同様に，行政区分での各地域の Z スコアの最大値は 102 であるのに対し，提案手法では 127 であることから，提案手法におけるツリー分割では，行政区画にとらわれず共通の特徴を持つ地点を 1 つの地域として抽出できており，かつ，有意に多く存在するカテゴリを抽出できていることがわかる．

表 4 に，提案手法により抽出した等質地域に対するラベリング結果を示す．行政区分では上位に出現しなかった月島近辺が抽出されていること，行政区分でも上位となっていた大久保近辺に対する Z スコアの値が軒並み高いことから，適切な地域分割ができていると言える．一方で，那覇近辺が多く出現している，すなわち，細かく地域分割しすぎていることもわかる．しかし，上位のラベルが同じようなものが多い場合は分類し過ぎであるという知見が適切な分割数を求める手がかりになるとも言える．

5. おわりに

本研究では，行政区分にとらわれない共通の特性を有する等質地域に着目し，各等質地域の地域ブランド創出を目標に，地域に偏在する属性（飲食店のカテゴリなど）を用いてラベリングすることを試みた．提案手法では，飲食店の緯度経度情報から近傍に位置する店舗をつなぐことで最小全域木を構築する．属性の分布が大きく変化するノード間のリンクを切断することで最小全域木を分割し，等質地域を抽出する．そして，各地域に偏在する属性を用いて地域にラベルを付与する．実データを用いた評価実験により，行政区分と比較して，Z スコアの平均値が高くなるような地域分割を実現できることを確認した．さらに，分布数自体は少ないが，偏って分布するカテゴリによるラベリングができることも確認した．今後の課題として，Z スコアを用いて適切な等質地域の数を求めることがあげられる．

謝辞 本研究は，JSPS 科研費 (No.19K20417) の助成を受けたものである．

参考文献

[1] 滝本広樹ほか：SNS 投稿写真の画像内容に基づく地域間の類似度算出に関する検討，信学技報， Vol. 116, No. 73, pp. 83–88 (2016).

[2] 奥 健太ほか：地域限定性スコアに基づく位置情報付きコンテンツからの地域限定語句の抽出，情報処理学会論文誌データベース（TOD），Vol. 5, No. 3, pp. 97–116 (2012).