# 位置情報付きソーシャルメディアデータの意味・空間結合クラスタリング手法の提案

名倉　幸大†　　村山　太一†

† 横浜国立大学 大学院環境情報学府〒 240–0067 神奈川県横浜市保土ケ谷区常盤台７９－１
E-mail: †{nagura-kodai-cz,murayama-taichi-bs}@ynu.jp

**あらまし**　都市計画やマーケティングにおいて，人々の実際の活動に基づく「都市機能領域」の推定に対する需要が高まっている．しかし，従来手法では，都市構造を考慮せずに一律なグリッド等で空間を分割することが多く，実際の活動の連続性が分断されるという課題がある．一方，位置情報付きソーシャルメディアデータは，活動の場所とその意味的文脈の双方を含むため，都市空間を多層的に捉える可能性を持つが，これらを統合して自律的に領域を推定する確立された方法論は存在しない．そこで本研究では，投稿から得られる空間情報と意味情報を確率的に統合し，柔軟な領域形状をデータ駆動的に推定するモデルを提案する．提案手法は EM アルゴリズムに基づいており，データの分布形状に合わせて意味的な一貫性と空間的な広がりの双方を同時に満たす機能領域を抽出する．東京都渋谷区の X（旧 Twitter）データを用いた実証実験では，単純な特徴量結合による K-means 法との比較を行った．その結果，提案手法は，定量的には空間的な凝集性と意味的な一貫性の均衡を保ちつつ，定性的には地理的な包含関係にある広域エリアと局所的な商業施設の分離や，ライブハウス密集地における文脈に基づく微細な領域分割が可能であることを示した．

**キーワード**　地理情報，ソーシャルメディア，計算機社会科学

## 1　は じ め に

近年，都市計画やマーケティングの分野において，人々の実際の活動に基づいて都市の機能領域を把握する需要が高まっている．都市機能領域の抽出には，従来の都市計画や行政施策において土地利用や土地被覆分類といった枠組みが用いられてきたが，これらは主に行政区画や用途地域などの固定的な境界に基づいている．しかし，現代都市における人々の生活圏や活動範囲は流動的かつ複合的であり，これらは必ずしも固定的な境界線とは一致しないため，既存の固定的な区分だけでは，実際の都市利用の実態を十分に捉えきれないという課題がある．そこで近年，トップダウンな区画ではなく，人々の実際の活動データに基づいた動的な都市機能領域を把握する手法が求められる．

本稿では，都市機能領域の抽出を，人々の移動履歴や滞在パターンといった活動データに基づき，類似した機能や役割を持つ空間を同定するプロセスと定義する．このプロセスにより，固定的な境界に縛られない実質的な活動エリアを可視化することが可能となる．こうした解析には，GPS による人流データや Point of Interest（POI）データが広く活用されているが，これらは人々の移動量や物理的な施設分布を把握する上で強力である反面，その場所が実際にどのような文脈で利用されているかという意味的な側面を捉えることには限界がある．この点において，位置情報付きソーシャルメディアデータは，物理的な投稿位置と活動内容や印象を反映するテキストが紐づいているため，その場所が持つ意味的な文脈も捉えることが可能となる．

しかし，このような都市機能領域の抽出においては，データの種類に関わらず，計算の効率性や扱いやすさから，対象地域を地域メッシュ（グリッド）などの固定的な単位で分割し，その単位ごとに集計を行う手法が一般的である．このグリッドベースの手法には，都市機能の抽出において二つの本質的な課題がある．第一に，人為的な境界設定により，実際の活動領域が分断される点である．グリッドは，道路網や地形などの都市構造を無視して一律に空間を分割するため，実際の人の動きや物理的な区切りとは必ずしも一致しない．その結果，商店街や公園といった，本来連続しているはずの活動領域がグリッド境界によって分断され，一つのまとまりある機能として抽出できず，実態との乖離が生じる．第二に，グリッドの粒度が固定されることにより，異質な機能が混在してその領域の特性が平滑化される点である．都市機能は，局所的なスポットから広域的なエリアまで多層的なスケールで存在するが，固定化されたグリッドはこれら固有の空間的な広がりに適応して境界を変化させることができない．その結果，単一のグリッド内に複数の異質な機能が混在した場合，その地域の特性が平均化され，局所的な特徴が埋没してしまう．

そこで本研究では，事前定義された境界やグリッドに依存しない，データ駆動的な都市機能領域の抽出を行う手法を提案する．具体的には，位置情報付きソーシャルメディアデータに含まれる位置情報（空間特徴）とテキスト（意味特徴）の双方を確率変数として扱い，両者の同時確率を最大化するように領域推定する確率モデルを構築する．これにより，空間的に近接し，かつ活動内容が類似した領域を都市機能領域として柔軟に推定することが可能となる．本手法は，データの分布形状そのもの

から境界を決定するため，従来のグリッド手法では捉えきれなかった複雑な形状の領域や，地理的に重複・近接していても意味的な文脈が異なる多層的な機能領域を，人為的なバイアスなく分離・抽出することが可能である．

本稿では，東京都渋谷区のソーシャルメディアデータ（X, 旧 Twitter）を用いた実証実験を行い，単純な特徴量結合による K-means 法との比較を通じて，提案手法の有効性を検証する．実験の結果，定量的には，単純な結合手法において著しく低い値となっていた地理空間上の凝集度を大幅に改善し，空間的なまとまりと意味的な一貫性の双方の指標において均衡の取れた結果を示した．また，目視で観察したところ，地理的に包含関係にある広域的な飲食エリアと局所的な商業施設を分離して抽出できたほか，ライブハウス密集地域において文化的な文脈に基づき領域を細分化できるなど，従来の距離ベースの手法では捉えきれない都市の多層的な機能構造を明らかにできることが示された．

## 2 関 連 研 究

### 2.1 空間単位の選択における従来手法の課題

都市機能領域や人流パターンの分析において，空間単位の選択は結果の妥当性を左右する．従来の研究では，行政区画やグリッドといった固定的な単位が広く採用されてきた．例えば，Rowe ら [1] は，パンデミック前後の人口移動を分析する際，人口密度に基づき分類された広域的なメッシュ単位で集計を行い，大都市圏から地方への人口流出傾向をマクロな視点から論じている．また，Kissler ら [2] は，ニューヨーク市内の行政区レベルで移動データを統合し，通勤行動の減少と，SARS-CoV-2 有病率の間に強い負の相関が見られることを明らかにした．さらに，Liu ら [3] は都市機能領域の推定において，道路網で区切られた空間単位を定義し，これらに POI や人流データを統合することで各空間単位内の機能識別を行った．これらはマクロな傾向把握には有効であるが，Dark and Bram [4] や Akbari ら [5] が指摘するように，これらの人為的な境界設定は可変単位地区問題（Modifiable Areal Unit Problem; MAUP）を引き起こす．具体的には，本来一体であるべき機能領域が分断されてしまうゾーニング効果や，集計単位が大きすぎるために地域内の異質な特性が平均化されてしまうスケール効果が生じる．実際の都市活動は道路網や地形に沿って連続的かつ不均一に広がっているため，これらの固定的な境界設定は，都市の微細な活動文脈を捉え損ねるという課題を抱えている．

### 2.2 位置情報と意味情報の統合アプローチ

グリッドの制約を回避し，かつ場所の意味を考慮するために，様々なデータソースを用いた新たな手法が提案されている．テキスト以外のデータソースを用いた研究として，Che ら [6] はストリートビュー画像や衛星画像を用いた景観分類を行い，Liu ら [3] は POI・人流データを統合した都市機能の推定手法を提案している．これらの手法は，都市の物理的な構造や，施設分布に基づく機能区分を把握する上で非常に有効である．しかし，その場所で人々がどのような文脈で活動しているかという，動的な意味情報を直接的に捉えることには限界がある．

一方で，人々の活動内容や文脈を捉えるために，ソーシャルメディアのテキスト情報を用いた研究も行われている．Tantoush ら [7] は，Twitter データの位置情報，投稿テキスト，投稿日時を用いて ST-DBSCAN（時空間密度クラスタリング）を適用し，グリッドレスに活動領域を抽出するフレームワークを提案した．これは，位置と時間に基づいてクラスタを形成し，その後に LDA トピックモデルを用いて意味を付与する段階的なアプローチである．しかし，このような段階的アプローチでは，テキストの意味情報を考慮せずに空間的なクラスタリングを行った後に意味を付与するため，最終的な抽出精度が初期のクラスタリングの境界設定に強く依存してしまうという問題が残る．

これに対し本研究は，位置情報とテキスト情報を同時確率として単一のモデルで扱う点に独自性がある．空間的な凝集性と意味的な一貫性を同時に最適化することで，都市機能の文脈に則した領域分割を実現する．

## 3 提 案 手 法

本章では，位置情報とテキスト情報を確率的に統合し，都市機能領域を抽出するための確率モデルについて述べる．具体的には，各投稿が潜在的な機能領域から生成されると仮定した混合モデルを構築し，EM アルゴリズムを用いてモデルパラメータを推定する．これにより，人為的な境界設定に依存することなく，データの分布特性に基づいて，空間的な広がりと意味的な特徴の双方を反映した領域抽出を実現する．提案手法の全体図を図 1 に示す．

### 3.1 データ前処理と特徴量抽出

ソーシャルメディアデータには，広告やボット，あるいは極端に短い投稿などのノイズが多く含まれる．都市機能の推定において，ユーザーの実際の体験や文脈を反映していない投稿は分析の妨げとなるため，以下の基準に基づき，フィルタリングを行う．

1. bot の除去：自動生成される定型文を含む投稿の削除
2. HTML リンクの削除
3. 他ユーザーへのメンションの削除
4. ハッシュタグの削除
5. 投稿末尾に付加された位置情報の削除（例：[投稿本文] @東京都渋谷区）
6. 空白文字の削除
7. 日本語以外の投稿の削除
8. 短すぎる投稿の削除
9. 重複投稿の削除

前処理後のテキストデータに対し，Sentence-BERT を適用し，投稿単位での埋め込みベクトルを取得する．本研究では，日本語に特化した事前学習済みモデル [1]を採用し，768 次元の

1：sonoisa/sentence-bert-base-ja-mean-tokens-v2

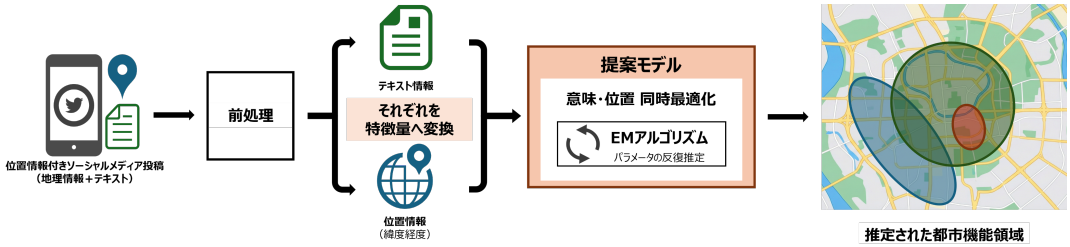


図 1　提案手法の全体処理フロー．位置情報とテキスト情報をそれぞれ独立した特徴量としてベクトル化し，同時確率モデル（EM アルゴリズム）を用いて統合することで，空間的な凝集性と意味的な一貫性の双方を考慮した機能領域を抽出する．

テキスト特徴量を得る．しかし，768 次元のテキスト特徴量は 2 次元の位置特徴量と比較して極めて高次元であり，そのまま確率モデルに組み込むと，次元数の不均衡により，位置情報の影響が過小評価される恐れがある．そこで本研究では，テキスト特徴量を主成分分析（PCA）を用いて次元圧縮を行い，10 次元のベクトルへと変換した．さらに，これを各座標軸に対して平均 0，分散 1 となるように標準化を行ったものをテキスト特徴量 $t_n$ とする．

位置情報は緯度経度（単位：度）で与えられているが，本研究では，位置情報の近接性をユークリッド距離に基づいて評価するため，対象地域に適した平面直角座標系（JGD2011 / Plane Rectangular Coordinate System IX, EPSG:6677）への投影変換を行う．これにより，位置情報をメートル単位の 2 次元ユークリッド座標に変換する．さらに，テキスト特徴量とのスケールの均衡を保つため，各座標軸に対して平均 0，分散 1 となるように標準化を行ったものを位置特徴量 $l_n$ とする．

### 3.2　モデルの定式化

本節では，ソーシャルメディア投稿に含まれた位置特徴量 $l_n \in \mathbb{R}^2$ とテキスト特徴量 $t_n \in \mathbb{R}^{10}$ を統合してクラスタリングを行うための確率モデルを提案する．各投稿は $x_n = (l_n, t_n)$ と表し，$n = 1, \ldots, N$ とする．

#### 3.2.1　潜在変数の導入

各投稿 $x_n$ がいずれのクラスタに属するかを表す潜在変数 $\mathbf{z}_n$ を導入する．ここで，$\mathbf{z}_n$ は $K$ 次元の One-hot ベクトル $\mathbf{z}_n = (z_{n1}, \ldots, z_{nK})^T$ であり，以下の制約を満たす．

$$\sum_{k=1}^{K} z_{nk} = 1 \tag{1}$$

$\mathbf{z}_n$ は混合係数 $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ に従うカテゴリ分布によって生成される．

また，クラスタが与えられた条件下での位置情報とテキスト情報は条件付き独立であると仮定する．

これによって，各投稿の生成確率が，意味空間と位置空間の両面から独立にモデル化できる．その条件付き確率は、以下のような積の形で記述できる。

$$p(l_n, t_n | z_{nk} = 1) = \mathcal{N}(l_n | \boldsymbol{\mu}_k^{(l)}, \boldsymbol{\Sigma}_k^{(l)}) \mathcal{N}(t_n | \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}) \tag{2}$$

ここで，$\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ は多変量ガウス分布を表す．$\boldsymbol{\mu}_k^{(l)}, \boldsymbol{\Sigma}_k^{(l)}$ および $\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}$ は，それぞれクラスタ $k$ における位置とテキストの平均ベクトルおよび共分散行列である．

#### 3.2.2　周辺尤度

潜在変数 $\mathbf{z}_n$ について周辺化することで，観測データ $x_n$ の周辺分布は混合ガウス分布として表される．

$$p(l_n, t_n) = \sum_{\mathbf{z}_n} p(\mathbf{z}_n) p(l_n, t_n | \mathbf{z}_n) \tag{3}$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(l_n | \boldsymbol{\mu}_k^{(l)}, \boldsymbol{\Sigma}_k^{(l)}) \mathcal{N}(t_n | \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}) \tag{4}$$

投稿群 $X = \{x_1, \ldots, x_N\}$ 全体に対する対数尤度は，以下の式で与えられる．

$$\ln p(X | \boldsymbol{\theta}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(l_n | \boldsymbol{\mu}_k^{(l)}, \boldsymbol{\Sigma}_k^{(l)}) \mathcal{N}(t_n | \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}) \right\} \tag{5}$$

ここで，$\boldsymbol{\theta}$ はモデルの全パラメータ集合を表す．本手法では，この対数尤度を最大化するパラメータ $\boldsymbol{\theta}$ を推定することを目的とする．

### 3.3　パラメータ推定 (EM アルゴリズム)

対数尤度 $\ln p(X | \boldsymbol{\theta})$ の最大化は解析的に困難であるため，EM アルゴリズム（Expectation-Maximization Algorithm）を用いる．

#### 3.3.1　E ステップ（負担率の計算）

現在のパラメータ値を用いて，各投稿 $x_n$ がクラスタ $k$ に属する負担率 $\gamma(z_{nk})$ を計算する．

$$\gamma(z_{nk}) \equiv p(z_{nk} = 1 | x_n) \tag{6}$$

$$= \frac{\pi_k \mathcal{N}(l_n | \boldsymbol{\mu}_k^{(l)}, \boldsymbol{\Sigma}_k^{(l)}) \mathcal{N}(t_n | \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(l_n | \boldsymbol{\mu}_j^{(l)}, \boldsymbol{\Sigma}_j^{(l)}) \mathcal{N}(t_n | \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})} \tag{7}$$

式（ 7）において，負担率は位置およびテキストそれぞれの確率密度の積として算出される．実装上は，数値計算の安定性を確保するため，対数領域での計算を行う．

#### 3.3.2　M ステップ（パラメータ更新）

次に，E ステップで得られた負担率 $\gamma(z_{nk})$ を用いて，対数尤度を最大化するようにパラメータを更新する．条件付き独立性の仮定により，位置とテキストのパラメータはそれぞれ独立

に更新される．

まず，各クラスタの実効的なデータ数 $N_k$ を以下のように定義する．

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{8}$$

これを用いて，各パラメータの更新式は以下のように与えられる．

混合係数の更新：

$$\pi_k^{new} = \frac{N_k}{N} \tag{9}$$

位置パラメータの更新：

$$\boldsymbol{\mu}_k^{(l),new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) l_n \tag{10}$$

$$\boldsymbol{\Sigma}_k^{(l),new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(l_n - \boldsymbol{\mu}_k^{(l),new})(l_n - \boldsymbol{\mu}_k^{(l),new})^T + \epsilon I \tag{11}$$

テキストパラメータの更新：

$$\boldsymbol{\mu}_k^{(t),new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) t_n \tag{12}$$

$$\boldsymbol{\Sigma}_k^{(t),new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(t_n - \boldsymbol{\mu}_k^{(t),new})(t_n - \boldsymbol{\mu}_k^{(t),new})^T + \epsilon I \tag{13}$$

ここで，$\epsilon I$（本研究では $\epsilon = 10^{-6}$）は共分散行列の正定値性を保証するための正則化項である．

#### 3.3.3 アルゴリズムの全体像

本アルゴリズムは，初期化を行った後，対数尤度が収束するまで E ステップと M ステップを交互に反復する．具体的な処理の流れは以下の通りである．

1. **初期化 (Line 1)**:
   混合係数 $\boldsymbol{\pi}$，平均ベクトル $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{\Sigma}$ をランダムに初期化する．
2. **E ステップ (Lines 4–8)**:
   現在のパラメータを用いて，全データ点に対して各クラスタへの負担率 $\gamma(z_{nk})$ を計算する（式 (7)）．
3. **M ステップ (Lines 9–14)**:
   算出された負担率を用いて，各クラスタの実効データ数 $N_k$ を求めた後，パラメータ $\boldsymbol{\theta}$ を更新する（式 (9)–(13)）．
4. **収束判定 (Lines 15–19)**:
   更新後のパラメータを用いて対数尤度を計算し（式 (5)），前回との差分が閾値 $\delta$ を下回った場合に収束とみなしてループを終了する．

**Algorithm 1** Proposed EM Algorithm

**Input:** Observed data $X = \{x_1, \ldots, x_N\}$, Number of clusters $K$
**Output:** Estimated parameters $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}^{(l)}, \boldsymbol{\Sigma}^{(l)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}\}$

▷ *Initialize parameters*

1: Initialize $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ randomly
2: $\mathcal{L}_{old} \leftarrow -\infty$

▷ *Iterate until convergence*

3: **while** not converged **do**
   /* **E-Step**: Calculate Responsibilities */
4:  **for** $n = 1$ to $N$ **do**
5:  **for** $k = 1$ to $K$ **do**
6:  Calculate $\gamma(z_{nk})$ using Eq. (7)
7:  **end for**
8:  **end for**
   /* **M-Step**: Update Parameters */
9:  **for** $k = 1$ to $K$ **do**
10:  Calculate $N_k \leftarrow \sum_{n=1}^{N} \gamma(z_{nk})$
11:  Update $\pi_k$ using Eq. (9)
12:  Update $\boldsymbol{\mu}_k^{(l)}, \boldsymbol{\Sigma}_k^{(l)}$ using Eq. (10), (11)
13:  Update $\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}$ using Eq. (12), (13)
14:  **end for**
   /* Check Convergence */
15:  Calculate log-likelihood $\mathcal{L}_{new}$ using Eq. (5)
16:  **if** $|\mathcal{L}_{new} - \mathcal{L}_{old}| < \delta$ **then**
17:  **break**
18:  **end if**
19:  $\mathcal{L}_{old} \leftarrow \mathcal{L}_{new}$
20: **end while**
21: **return** $\boldsymbol{\theta}$

## 4 実験設定

本章では，X（旧 Twitter）の投稿データを用いた東京都渋谷区における都市機能領域の抽出実験における実験設定について述べる．

### 4.1 実験データ

実験には，X（旧 Twitter）から収集した位置情報付き投稿データを用いた．対象地域は東京都渋谷区全域とし，対象期間は 2019 年 1 月 1 日から 2020 年 12 月 31 日までの 2 年間である．収集されたデータは 657,498 件であったが，第 3 章で述べた前処理プロセスに基づきデータのフィルタリングを実施したところ，最終的に実験に用いるデータ件数は 160,037 件となった．

### 4.2 パラメータ設定

提案モデルの学習における主要なパラメータ設定を以下に示す．抽出された機能領域の数（クラスタ数）$K$ は渋谷区内の多様な活動を十分に分離できる粒度として，予備的な検討に基づいて $K = 100$ に設定した．モデルの EM アルゴリズムの収束判定閾値は $1.0 \times 10^{-4}$ とし，最大反復回数は 100 回とした．

### 4.3 比較手法

提案手法の有効性を検証するため，ベースライン手法として以下の K-means 法を設定する．なお，比較手法におけるクラスタ数は提案手法と同様に $K = 100$ とした．

本手法は，提案手法と同様の前処理を経た位置特徴量（2 次元）とテキスト特徴量（10 次元）を単純に結合し，12 次元のベクトルとしたものを入力として K-means 法によるクラスタリングを行うものである．本比較を通して，単純な距離ベースの結合に対する，提案手法の特徴の違いを検証する．

### 4.4 評価指標

抽出された機能領域の妥当性を評価するために，本研究では定量的，定性的の両面から評価を行う．

#### 4.4.1 定量的評価指標

本実験では，各手法によって得られたクラスタリング結果に対し，地理空間および意味空間のそれぞれにおいて，クラスタの凝集度と分離度を評価する．これにより，手法がどちらの情報の空間構造をより強く反映しているか，あるいは双方のバランスを保てているかを検証する．本タスクには正解ラベルが存在しないため，以下の 3 つの内部指標を用いて，地理空間および意味空間のそれぞれにおける評価値を算出する．

1. **シルエット係数（Silhouette Coefficient）**
   クラスタ内の凝集度と他クラスタとの分離度を示す指標．-1 から 1 の値をとり，値が大きいほど適切にクラスタが分離されていることを示す．
2. **Calinski-Harabasz Index（CHI）**
   クラスタ間分散とクラスタ内分散の比率に基づく指標．値が大きいほど，クラスタが密で，かつ互いに離れていることを示す．
3. **Davies-Bouldin Index（DBI）**
   各クラスタの散らばり具合と，クラスタ間の距離の比率に基づく指標．クラスタ内の類似性が高く，クラスタ間の異質性が高いほど小さな値となる．

## 5 結　　果

### 5.1 定量評価

本研究では，都市機能領域の妥当性を検証するため，地理空間における空間的凝集性と，テキスト空間における意味的一貫性の 2 つの観点から評価を行った．各手法におけるクラスタリング指標の比較結果を表 1 に示す．

全体的な傾向として，比較手法はテキスト空間における指標で相対的に良好な値を示す一方，地理空間における指標では著しく低い結果となった．対照的に，提案手法はテキスト空間の指標では比較手法に劣るものの，地理空間の指標において良好な値を示しており，空間と意味の双方において均衡のとれた結果となった．

まず，テキスト空間における評価では，比較手法が提案手法よりも良好な値を示している．これは，入力ベクトルの次元構成に起因すると考えられる．比較手法では，2 次元の位置情報と 10 次元のテキスト特徴量を等価に結合して 12 次元ベクトルとして扱っているため，クラスタリングの基準となる距離計算において，次元数の多いテキスト情報の影響が支配的となる．その結果，比較手法は意味的な類似性を優先したクラスタリングを行ったことで意味空間での凝集度が高くなったと考えられる．しかし，これは地理的なまとまりの欠如を意味しており，抽出された機能領域が空間的に連続せず，全体に散在していることを示唆している．

一方，提案手法は，テキスト空間の指標では K-means 法に及ばないものの，地理空間の指標においては良好な値を示していることが確認できる．提案手法は，位置とテキストを独立した確率変数として扱い，それぞれの分布の同時確率を最適化しているため，次元数の不均衡に左右されることなく，地理的な凝集性と意味的な一貫性の双方をバランスよく満たす領域分割を実現できている．以上の結果から，提案手法は特定の指標を過度に優先することなく，本研究の設計趣旨である「空間的なまとまりと意味的な文脈の双方を考慮した領域分割」を適切に反映した挙動を示したと評価できる．

### 5.2 定性評価

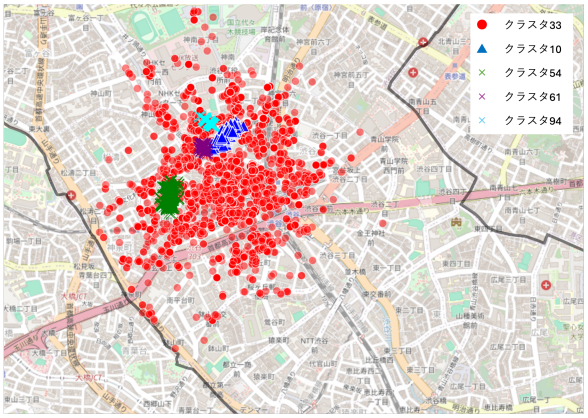


図 2　抽出された主要な機能領域の空間分布．(赤: クラスタ 33, 青: クラスタ 10, その他: クラスタ 54, 61, 94)

本実験により抽出された機能領域の中から，提案手法の特徴である，意味的文脈に基づく空間分離が顕著に表れている事例として，商業施設と広域エリアの包含関係の分離，および，類似機能の文脈による細分化，の 2 点について考察する．各クラスタの空間分布を図 2 に，代表的なトピックを表 2 に示す．

#### 5.2.1 地理的包含関係の機能の分離

クラスタ 33（赤色）は，道玄坂エリアを中心に広域に分布しており，頻出単語を確認すると，飲食全般に関連する汎用的な話題で構成されている．一方，クラスタ 10（青色）は，クラスタ 33 の分布領域の内部（渋谷 PARCO 周辺）に局所的に存在している．頻出単語を確認すると，特定の商業施設やイベントに関連する語彙が支配的であり，クラスタ 33 とは明確に異なる文脈を持っている．地理的には，クラスタ 10（特定の商業施設）はクラスタ 33（広域の飲食エリア）に包含される位置関係にある．地理情報のみに依存する従来手法では，このように

表 1　提案手法と比較手法の定量評価比較

| | 地理空間における評価（Geo） | | | テキスト空間における評価（Text） | | |
|---|---|---|---|---|---|---|
| 手法 | Silhouette | CHI | DBI* | Silhouette | CHI | DBI* |
| 結合 K-means | -0.310 | 1427.046 | 37.421 | **0.033** | **1014.349** | **3.291** |
| **提案手法** | **-0.132** | **4702.468** | **5.100** | -0.153 | 225.677 | 7.385 |

* DBI のみ値が小さいほど良好な結果であることを示す.

表 2　定性評価に用いた各クラスタの意味的特徴と代表的な投稿

| クラスタ ID | 推定される機能領域 | 頻出単語 | 代表的な投稿例 |
|---|---|---|---|
| 33 | 道玄坂・飲食エリア | ラーメン，焼肉，カレー | 渋谷で焼肉ランチ。これだけ食べれてお値段 1,600 円ほど。 しかも、牛タンが柔らかくて美味い! |
| 10 | 渋谷 PARCO | パルコ，任天堂，グランドオープン | 新しくなった PARCO 劇場で初観劇。 |
| 54, 61, 94 | ライブハウス群 | ワンマン，ツアー,（具体的なライブハウス名） | （各クラスタで異なる規模感やジャンルの公演への言及） |

空間的に重なり合う都市機能領域は同一クラスタに埋没しやすい．しかし提案手法は，クラスタ 10 が持つ特異的な意味特徴と，クラスタ 33 がもつ汎用的な意味特徴の差異を捉えることで，地理的に包含関係にあっても，異なる機能領域として明確に分離・抽出することに成功している．

**5. 2. 2**　文脈による同種機能の細分化

クラスタ 54, 61, 94 は，いずれも「ワンマン」「ツアー」，具体的なライブハウス名といった単語が含まれており，広義には「ライブハウス・イベント」機能を持つ領域である．これらのクラスタには，イベントへの参加報告やアーティストに対する熱量の高い感情表現など，類似した文脈が共通して見られる．しかし，空間的な分布とテキストの詳細を確認すると，これらは単一の領域としては抽出されず，物理的な施設集積ごとに明確に分離されている．具体的には，クラスタ 54 は円山町・道玄坂エリアの大規模なライブハウスやクラブ，クラスタ 61, 94 は宇田川町エリアの中規模施設や特定の拠点といったように，所在地，施設の規模感，および開催されるイベント内容の差異に基づいて，それぞれ独立したクラスタを形成している．これは，提案手法が単にライブという大まかなトピックだけで分類を行っているのではなく，微細な地理的差異とそれに基づく施設の属性を捉えていることを示している．意味的に類似した活動であっても，物理的な場所や施設の性質が異なることを識別し，高解像度に分離・抽出することに成功している．

## 6　おわりに

本研究では，ソーシャルメディア投稿に含まれる位置情報とテキスト情報を確率的に統合し，都市空間内の潜在的な機能領域を抽出する手法を提案した．東京都渋谷区を対象とした実証実験の結果，提案手法はグリッド等の人為的な境界設定に依存することなく，人々の活動実態に基づいた柔軟な領域分割を実現した．

本研究の主な貢献は以下の 2 点である．第一に，地理的な近接性と意味的な文脈の双方を考慮することで，都市の多層的な構造を明らかにした点である．定性評価において，広域的な飲食エリアに包含される特定の商業施設の分離や，地理的に密集するライブハウス群を施設の規模や文化的な文脈に基づいて微細に分離することに成功した．これにより，従来の単純な位置クラスタリングでは埋没していた，物理的配置と意味的文脈が複雑に絡み合った都市機能の構造を可視化することが可能となった．第二に，位置情報とテキスト情報の同時確率を最適化するアプローチの有効性を示した点である．定量評価の結果，単純なベクトル結合を用いた K-means 法がテキスト情報の次元数に引きずられ地理的な凝集性を損なう傾向にあったのに対し，提案手法は地理的なまとまりと意味的な一貫性のバランスを保った領域分割を実現した．これは，都市機能領域の抽出において，空間的な連続性と意味的な類似性の双方を同時に満たす確率モデルの構築が不可欠であることを示唆している．

今後の課題として以下の点が挙げられる．第一に，データの代表性とバイアスに関する限界が挙げられる．ソーシャルメディアの利用者は年齢層や属性に偏りがあり，必ずしも実際の都市人口の分布を正確に反映しているとは限らない．特に，若年層や特定の趣味嗜好を持つ層の行動が過大評価されている可能性があるため，異なる属性特性を持つデータと統合したマルチモーダルな解析や，統計的なバイアス補正手法の導入が必要である．第二に，抽出精度の向上とノイズへの対応である．実験結果において，特定の店舗やボットに由来する一点集中のクラスタや，解釈が困難なクラスタが一部確認された．これらは，位置情報やユーザーの重複，日常会話などの汎用的なテキストが原因で形成されたと考えられる．これらのノイズを適切に処理し，より純度の高い機能領域を抽出するアルゴリズムの改良が求められる．

## 謝　辞

本研究は JSPS 科研費 JP23K16889 の助成を受けたものです.

## 文　献

[1] Calafiore A. Arribas‐Bel D. Samardzhiev K. Fleischmann M.Rowe. Urban exodus? understanding human mobility in britain during the covid‐19 pandemic using meta - facebook data. *Population, Space and Place*, 2022.

[2] Kishore N. Prabhu M. et al. Kissler, S.M. Reductions in commuting mobility correlate with geographic differences in sars-cov-2 prevalence in new york city. *Nature Communication*, 2020.

[3] Xingyu Liu, Yehua Sheng, and Lei Yu. A data-synthesis-driven approach to recognize urban functional zones by integrating dynamic semantic features. *Land*, Vol. 14, No. 3, p. 489, 2025.

[4] Shawna J Dark and Danielle Bram. The modifiable areal unit problem (maup) in physical geography. *Progress in physical geography*, Vol. 31, No. 5, pp. 471–479, 2007.

[5] Kamal Akbari, Stephan Winter, and Martin Tomko. Spatial causality: A systematic review on spatial causal inference. *Geographical Analysis*, Vol. 55, No. 1, pp. 56–89, 2023.

[6] Lin Che, Yizi Chen, Tanhua Jin, Martin Raubal, Konrad Schindler, and Peter Kiefer. Unsupervised urban land use mapping with street view contrastive clustering and a geographical prior. In *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems*, pp. 28–38, 2025.

[7] Mahmud Tantoush, Ulysses Sengupta, and Liangxiu Han. Exploring city dynamics through tweets: A framework for capturing urban activities as complex spatiotemporal patterns. *Cities*, Vol. 162, p. 105894, 2025.