トピックモデルによる土地利用分類

広島大学　学生会員　○塚野裕太
広島大学　正会員　塚井誠人

1. はじめに

日本では，地理情報データの整備が進んでおり，土地利用や人口動態などに関して，多種類の情報が詳細な地点別に入手できるようになった．しかし，従来の分析手法である主成分分析や因子分析などには，データが疎構造化（「0」観測値が極めて多い）する細かな地点単位の分析が難しいという問題点がある．本研究ではトピックモデルを用いて，土地利用特性の抽出を検討する．トピックモデルは，テキストデータへの適用が多くみられるものの，地理情報データへの適用例については筆者の知る限り存在しない．本研究では，トピックモデルの適用可能性を検証するため，主成分分析との性能比較を行う．

2. Latent Dirichlet Allocation(LDA)

確率的トピック生成モデル(LDA)の概略を示す．属性別の階級別属性wに対応する潜在トピックを表す潜在変数zを導入する．モデルの確率構造は，以下のように仮定する．潜在トピックkごとの属性vの分布をφ，メッシュdごとのトピックkの分布をθとする．分布φからは，トピックごとに属性の出現確率がわかり，トピックの特徴を推定できる．分布θからは，メッシュごとにトピックとの関連の強さがわかる．LDA の全確率を式(1)に，グラフィカルモデルを図 1 に示す．図 1 より，LDA は，属性とトピックの出現確率がそれぞれ多項分布に従い，それらの多項分布のパラメータがそれぞれ Dirichlet 分布に従う構造をとる．多項分布と Dirichlet 分布の式を式(2)，(3)にそれぞれ示す．

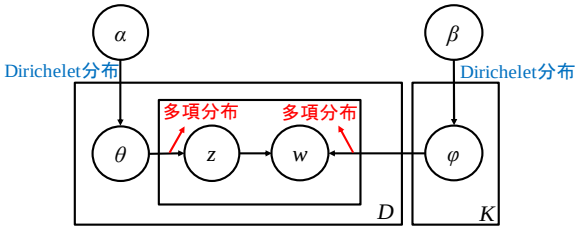

図 1 LDA のグラフィカルモデル

$$p(w,z,\theta,\varphi|\alpha,\beta)=p(w|z,\varphi)p(z|\theta)p(\theta|\alpha)p(\varphi|\beta) \quad (1)$$

$$p(\{n_k\}_{k=1}^{K}|\pi_k)=\frac{n_k!}{\prod_{k=1}^{K}n_k!}\prod_{k=1}^{K}\pi_k^{n_k} \quad (2)$$

$$Dir(p|\boldsymbol{\alpha})=\frac{\Gamma(\sum_{k}^{k}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}p^{\alpha_k-1},\quad k=1,\ldots,K \quad (3)$$

3. 地理情報データの加工

本研究では，国土数値情報ダウンロードサービスなどから，3 次メッシュレベルで 23 種類の属性データを収集した．対象地域は，広島市，山口市，松江市，鳥取市，岡山市とした．収集した属性データの一覧を**表 1** に示す．

地理情報データをトピックモデルによって解析するために，データを Bag of words 形式とする．ここで，Bag of words は，文書を行，語彙を列にとり，文書単位で語彙の出現頻度をカウントした行列である．地理情報データでは文書を地点，語彙をデータの種類とすると，語彙が少なく情報量が不足する．そこで，属性データ別に階級を設定して，その該当/非該当をダミー変数で与えた．具体的には属性ごとに自然分類によって 8 段階に階級分けを行い，分類には自然分類を用いた．ただし， 8 階級に分けられないデータは，階級数を減らした．以上の方法によって，23 属性から，181 階級別属性を得た．

4. トピックの抽出

トピックモデルでは，最適なトピック数 K を提示する必要がある．所与の K の下で算出したトピックモデルの適合度は，尤度比とトピック類似度によって判定する．先述したように，上位 30 位までのトピックごとの階級別属性の分布から，トピックにタイ

キーワード　地理情報データ　LDA　離散化　トピックの空間分布
連絡先　〒739-8527　東広島市鏡山 1-4-1 広島大学大学院工学研究科　社会基盤環境工学専攻　事務室
TEL：082-424-781

表 1 属性データ一覧

大分類	属性データ
土地	田
	他農用地
	建物用地
	標高差
世帯	核家族世帯数
	核家族以外世帯数
	6 歳未満世帯員のいる世帯数
	一戸建世帯数
	共同住宅世帯数
就業者	第 1 次産業就業者総数
	第 2 次産業就業者総数
	卸売・小売・金融・保険・不動産・物品賃貸・生活関連・娯楽業就業者総数
	医療・福祉就業者総数
	公務就業者総数
事業所	全産業_事務所数
	全産業_従業者数
	小売業事務所数
人口	小学校・中学校在学者総数
	高校在学者総数
	居住期間出生時から総数
	居住期間5年未満総数
	65歳以上人口
	昼夜間人口比率

表 2 トピックごとの階級別属性の分布

トピック	階級別属性
都心	建物用地8, 卸売・小売~就業者総数8, 第2次産業就業者総数8, 高校在学者総数8
近郊	建物用地8, 小売業事務所数3, 卸売・小売~就業者総数5, 核家族世帯数5, 医療・福祉就業者総数5
郊外	卸売・小売~就業者総数4, 核家族世帯数4, 第2次産業就業者総数4, 一戸建世帯数5,
低密度居住地+商業地	卸売・小売~就業者総数3, 一戸建世帯数4, 全産業_事務所数2, 第2次産業就業者総数3
農地+低密度居住地	田8, 建物用地3, 一戸建世帯数3, 核家族世帯数2
急傾斜地域+低密度居住地	標高差6, 一戸建世帯数2, 建物用地2, 核家族世帯数2, 核家族以外世帯数2
急傾斜地域+農地	標高差8, 田3, 他農用地3, 第1次産業就業者総数2
可住不適地	標高差8, 階級1の属性多数

トルを付けたところ，「都心」，「近郊」，「郊外」，「低密度居住地+商業地」，「農地+低密度居住地」，「急傾斜地域+低密度居住地」，「急傾斜地域+農地」，「可住不適地」の 8 トピックを得た．タイトルの命名に参考にした階級別属性を表 2 に示す．

5. トピックの空間分布

抽出された 8 トピックにそれぞれ固有の色を割り当てた．さらに，メッシュごとに帰属確率が最も高いトピックで塗り分けた．広島市のトピック分布を図 2 に示す．同図より，広島市の中心部から，都心→近郊→郊外というトピックの分布が見られる．さらに，都

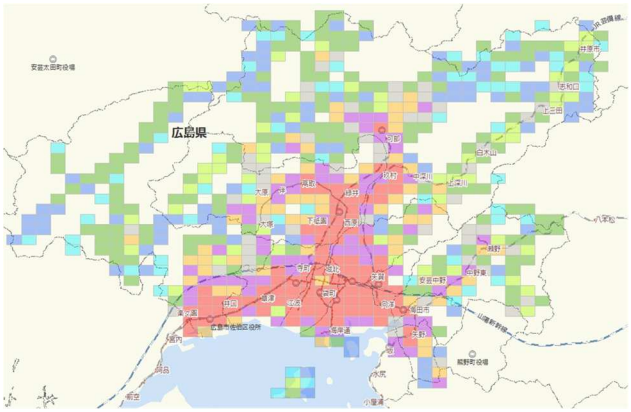

図 2 トピックの空間分布

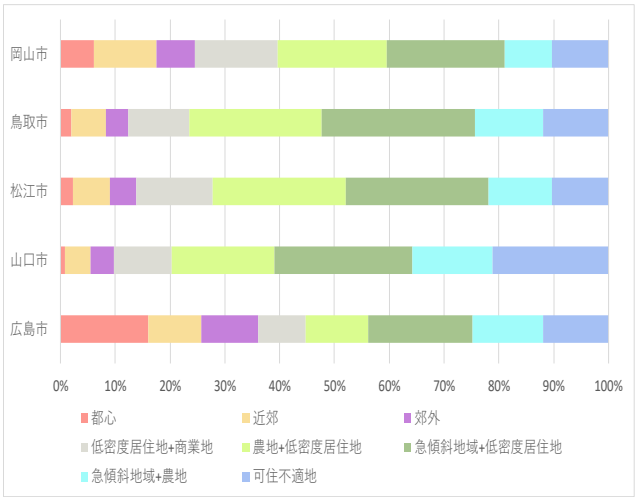

図 3 都市別のトピック構成比

市別のトピック構成比を図 3 に示す．鳥取市と松江市のトピック構成比は類似している．

6. 結論

トピックモデルを適用したところ，中国地方五県の県庁所在地を対象に，8 トピックが抽出された．さらに，紙数の都合で詳細を示すことができなかったが，従来の分析手法として主成分分析との比較を行い，トピックモデルの有用性を示した．

今後は属性データの拡張や，データの追加収集を行って，2 時点間のトピックの推移を考察する．

参考文献

(1) 塚井誠人，椎野創介：討議録に対するトピックモデルの適用，土木計画学研究・講演集(CD-ROM)，Vol.52，(2015)

(2) 岩田具治：トピックモデル，講談社，(機械学習プロフェッショナルシリーズ)，(2015)