移動系列データのベクトル化に基づく地域分類

伊藤 翔*　井上 亮**

Regional classification based on vectorizing human flow series data

Kakeru Ito, Ryo Inoue

Most studies on regional classification based on people flow data have analyzed OD data and have not considered the diversity of travel patterns, such as sequence of visits and purpose of stay. In this study, we proposed a new classification method to extract the characteristics of regions from people's travel history data by applying a natural language processing algorithm called Word2Vec, which obtains a vector representation of words from data of sentences. Through the application of Word2Vec to pseudo human flow data, the analysis utilizing Word2Vec proved to be effective in classifying regions. It was also confirmed that a more detailed understanding of regional characteristics can be obtained by analyzing travel history data that describes the purpose of travel.

Keywords: 地域分類（regional classification），移動系列データ（flow series data），Word2Vec

１．はじめに

現代の都市圏には，オフィス街，住宅街，歓楽街など，異なる性格を持つ地域が点在し，道路や鉄道などの張り巡らされた交通網がその間の移動を可能にしている．ある地域への流出入量やその時間帯毎の分布，目的の割合などの移動パターンの特徴は地域毎に多様で，それらは社会経済活動の特徴を表している．また，地域間の流動量の大小は，それらの関連性の強弱を表している．そこで，地域間流動データに注目して地域の拠点性やつながりが強い地域間を分析し，地域の社会経済活動の特徴を把握する研究が古くから盛んに行われてきた．例えば，都市圏や商圏などの「機能地域」抽出に関しては，因子分析法（例えば Goddard, 1970）や，モジュラリティ最大化法（例えば Farmer & Fotheringham, 2011; 福本・岡本, 2012），スペクトラルクラスタリング（例えば Noulas et al., 2011），グラフエンベッディング（例えば Hajiseyedjavadi et al., 2018）など多くの分析手法が応用され, 地域分類結果の有効性が議論されてきた．

既往分析の大半は，国勢調査やパーソントリップ調査などで観測された，発地・着地間の流動を表す OD データに基づく．しかし，個人の一日の移動には，通勤に伴う自宅と勤務地との往復から，買い物や娯楽に伴う回遊行動まで，様々な種類があるが，OD データに基づき移動の目的を踏まえて地域分類を行う分析には限界がある. 地域間の移動に加えて，その目的や，一日の移動の中で複数の地域を訪問する場合にその順序やタイミングも考慮した分析ができれば，地域の特徴をより詳細に把握し，分類に活用することが可能だと考える．

移動系列データに基づく分析としては，系列マイニングを応用した Versichele (2014) や，ハフマン符号化を応用した Inoue & Tsukahara (2016) などがあるが，既往分析の多くは観光など非日常的な移動系列を扱っているほか，対象領域を明確に分割できる解釈が容易な結果が得られる研究は少ない.

本研究では，自然言語処理における単語のベクトル化手法である Word2Vec に注目し，個人の移動系列データを活用した地域分類手法を提案する．Word2Vec は, 単語の系列である文章データをもとに, 前後の単語や文脈が類似している単語間を表すベクトルが類似の値を持つように学習する手法である．個人の一日の移動系列データに応用すると，前後の滞在地域に加え，移動系列の中で訪問目的や滞在時間帯が類似した地域を把握できると期待される．

本研究では, Word2Vec の移動系列データ分析への

*　学生会員　東北大学　大学院情報科学研究科（Tohoku University）
〒980-8579　宮城県仙台市青葉区荒巻字青葉 6-6-06　E-mail: kakeru.ito.s1@dc.tohoku.ac.jp
**　正会員　東北大学　大学院情報科学研究科（Tohoku University）

応用可能性を検証するため，疑似人流データを用いた実験を行う．既存の OD データ分析手法との比較を通して, Word2Vec による地域分類の特徴を把握し, Word2Vec の分析結果を活用した新たな地域分析の可能性について議論する．

２．提案手法

2.1.で，自然言語処理に活用される Word2Vec のアルゴリズムを説明し，得られる単語ベクトル表現の特徴を解説する．次に，2.2.で，移動系列データへの適用で得られると予想される，地域のベクトル表現の特徴を説明した後，それを活用した地域分類手法の可能性について述べる．

2.1. Word2Vec（skip-gram モデル)

Word2Vec は，Mikolov et al. (2013) が提案した自然言語処理手法で，単語をベクトルで表現する手法の 1 つである．本手法は，単語の意味は周囲の単語によって形成されるという分布仮説に基づいている．各単語のベクトル表現を第一層の重み層に記録し，文章中でその単語の前後に現れる単語を推測するタスクを行うニューラルネットワークを学習すると，単語の意味の近さが，単語に対応するベクトル表現の近さとして表される重み層が得られる．

(1) skip-gram モデル

skip-gram は Word2Vec の代表的な二層ニューラルネットワークモデルで，一単語を入力として与え，文章の中でその前後に表れる単語の確率を出力する．Skip-gram の重み層の学習の目的は,「$w_1, w_2, \dots, w_T$」という意味を持った単語列（文章）に対して，文章中の一単語w_t $(t = 1,2, \dots, T)$をニューラルネットワークに入力した際に，その周辺単語が出力される確率を最大化することである．

$$maximize \quad p(w_{t-C}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+C} | w_t) \qquad (1)$$

なお, Cはウィンドウサイズで, 予測する前後単語の範囲を決定するハイパーパラメータである．

skip-gram の入力層は$1 \times W$ (語彙数) の行列で，入力する単語に対応する成分が 1，それ以外は 0 を要素とする one-hot ベクトルが与えられる．したがって，入力層と中間層との間の重み層の各行ベクトルは単語ごとに固有のものとなり，中間層には入力した単語固有のベクトル値が出力される．これが入力単語のベクトル表現であり，重み層の各行に，対応する単語のベクトル表現が格納される．中間層から出力層にかけては，予測する前後の単語の位置に$2C$個の重み層があり，入力単語の周辺に表れる単語の確率を，ソフトマックス関数を用いて出力する．

$$p(w_O | w_I) = \frac{\exp\left({v'}_{w_O}{}^{\top} v_{w_I}\right)}{\sum_{w=1}^{W} \exp\left({v'}_{w}{}^{\top} v_{w_I}\right)} \qquad (2)$$

ここで，w_Iが入力語，w_Oが出力語，v_{w_I}が入力単語w_Iのベクトル表現，${v'}_w$が中間層から出力層への重み層の，単語wに対応するベクトルを表す．

出力層へ出力したのち，入力した文章である正解データをもとに重み層を更新する．skip-gram にはウィンドウサイズに応じた複数の出力層が存在するため，各出力の損失の合計を 1 つの文章データにおける損失とする（式(3)）．

$$-\frac{1}{T}\sum_{t=1}^{T}\sum_{-c \le j \le c, j \ne 0} \log p\left(w_{t+j} | w_t\right) \qquad (3)$$

この損失の最小化を目的関数とし，ニューラルネットワークの学習過程の代表的手法である確率的勾配降下法を用いて最適化を行う．

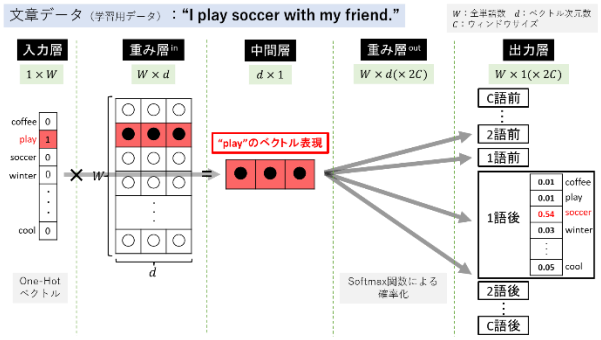


図-1　skip-gram のニューラルネットワーク

(2) ベクトル表現

skip-gram が入力語の前後単語を推測するニューラルネットワークであることから，類似した前後単語を持つ単語間では近いベクトル表現を学習する．したがって，単語の意味は周囲の単語によって形成されるという分布仮説に基づけば, Word2Vec のアルゴリズムにより，意味が類似する単語間で類似した

ベクトル表現を得ることができる．

単語間のベクトル表現の類似度は，コサイン類似度（cos 類似度）で評価できる．表-1(a)に，学習済みのベクトル表現を利用して，類似度が高い単語の一例を示す．“January”とコサイン類似度が高い単語として，同じく冬期の月を表す“February”や“December”などが挙げられる．文章中の前後に出現する単語が似ていることが評価され，意味が類似する単語と近いベクトル値を持つ結果が得られる．

また，Word2Vec で得られるベクトル表現の大きな特徴は，ベクトル表現の加減算によって単語間の意味関係を理解できることである．表-1(b)は演算の一例を示す．日本の首都 “Tokyo”から“Japan”を引き“France”を加えて得られるベクトル値と最も cos 類似度が高い単語は，フランスの首都である“Paris”となる．またその cos 類似度は，“Tokyo”と“Paris”の cos 類似度よりも大きく，ベクトル表現の演算により“Tokyo”という単語が持つ国の概念を変化させられることを示している．

表-1 Word2Vec によるベクトル表現の単語間類似性

(a)“January”

順位	単語	cos類似度
1	February	0.986
2	December	0.975
3	October	0.971
4	November	0.970
5	March	0.970

(b)“Tokyo”-“Japan”+“France”

順位	単語	cos類似度
1	Paris	0.835(0.547)
2	Toulouse	0.754(0.327)
3	Marseille	0.738(0.379)
4	Rennes	0.699(0.333)
5	Marseilles	0.696(0.348)

カッコ内は"Tokyo"とのcos類似度

※fastText (https://fasttext.cc/docs/en/crawl-vectors.html) で公開されている学習済みベクトル表現を使用．

2.2. 移動系列データへの応用

本研究では，Word2Vec を移動系列データに適用し，移動実態から明らかになる地域の特徴をベクトルで表現し，その結果に基づいて地域を分類することを提案する．具体的には，ある個人の「【地域 A】→【地域 B】→【地域 C】→【地域 A】」という移動系列があるとき，これを 4 単語からなる文章とみなし，skip-gram による地域のベクトル表現の学習データとして用いる．1 日の滞在地域をカバーできる大きさにウィンドウサイズを設定して学習したベクトル表現は，一日の移動系列の中で前後の滞在地域が類似する地域間で近い値をとる．本研究では，地域のベクトル表現間のユークリッド距離に基づいて，階層クラスタリング手法であるウォード法によって地域を分類する．

また，「【自宅 A】→【通勤 B】→【買い物 C】→【自宅 A】」のように，各地域の滞在目的を表すラベルを移動系列に付与して分析すると，滞在目的を考慮して地域の特徴を把握できると期待される．また，ラベルごとの地域のベクトル表現が，ラベル間の関係性についても学習できるなら，ベクトル表現間の加減算ができ，また，ラベル間の関係性の差異を活用した地域分類が可能であると考える．例えば，「【買い物 A】－【通勤 A】＋【通勤 B】」という演算を行うと，「【買い物 B】」に近いベクトル表現が得られると期待される．さらに，各地域の【買い物】ベクトルから【自宅】ベクトルを引いたベクトル値は，各地域の「商業地」としての性質と，「住宅地」としての性質の関係性を表すと考えられ，それを活用した地域分類は，住宅地や商業集積地などの地域の特徴を抽出できる可能性がある．

これら Word2Vec の移動系列データ分析への応用可能性について，次章で検証する．

3．疑似人流データを用いた地域分類

本章では，提案手法によって得られる地域分類結果の特徴について，疑似的な移動系列データを用いた分析により考察する．3.1.では実験に使用する「疑似人流データ」の概要や，本章共通の分析設定について述べる．次に 3.2.では，提案手法が既往手法とは異なる分類結果を抽出できることを確認するために，ラベルを与えない移動系列データを作成し，分類結果について考察する．最後に，3.3.ではトリップに「通勤」，「買い物」といった目的を付与した系列データを作成し，目的ごとにベクトル表現の学習，地域分類を行う．この節では(1)で目的ごとに異なる分類結果を出力することを確認するとともに，(2)にて目的間，地域間のベクトル演算を検証し，差分ベクトルを活用した集積地抽出を試みる．

3.1. 使用データ

東京大学空間情報科学研究センター「人の流れプロジェクト」よりご提供いただいた，「疑似人流・トリップデータ　東京都データセット」を用いて地域分類を実験する．本データは東京都居住者の一日のトリップを疑似的に生成したものであり，トリップ前後の緯度経度座標に加え，時間，交通手段，目的，個人属性が付与されている．そのうち本研究では，外出していないデータや東京都外へのトリップを含むデータを除いた，8,681,031 人分のデータを使用した．また，東京都市圏パーソントリップ調査の 416 小ゾーンを分析単位の地域として設定した．

3.2. 提案手法の適用

本節では，提案手法によって得られる地域分類結果について，既往研究で多く用いられるモジュラリティ最大化法との比較を交えながら考察する．

提案手法の適用では，疑似人流データから，移動目的などのラベルを付与しない移動系列を作成し，地域ごとに計 416 個のベクトル表現を学習した．このとき，ベクトル表現の次元数は 100，ウィンドウサイズは全データ中，最大の系列長である 20 とした．図-2 は得られたベクトル表現の特徴の一例として，丸の内地域と他地域の間の cos 類似度の分布を表したものである．丸の内地域に対して，八重洲地域や大手町地域が高い類似を示している．この結果は，丸の内地域の滞在前後に訪れる地域と，八重洲・大手町地域の滞在前後に訪れる地域が，類似していることを示す．

次に，学習したベクトル表現に基づく地域分類結果を図-3(a)に示す．また比較のため，モジュラリティ最大化法（Fortunato, 2010）の結果を図-3(b)に示す．

モジュラリティ最大化法は，ネットワークのコミュニティ分割の質を表す値であるモジュラリティが，最大となる分割をする手法で，機能地域抽出をはじめとする地域分類問題に多く応用されている．ここでは疑似人流データから地域間の無向 OD 行列を作成し分析した．なお，分類数は，モジュラリティ最大化法において最良の 6 を両手法で用いた．

分析の結果，提案手法は飛び地の無い解釈しやすい地域群を抽出した．ただし，モジュラリティ最大化法の結果とは一部に違いが見られる．モジュラリティ最大化法では，杉並区の小ゾーンと武蔵野市の小ゾーンを異なる地域群に分類しているが，本手法では同一の地域群に分類している．モジュラリティ最大化法では，杉並区が新宿区や代々木区など都心部との間で流動が多いこと，多摩東部が三鷹市や西東京市など多摩東部の他地域との間で流動が多いことにより，それぞれのクラスタ内でのつながりの強さが評価されているものと考えられる．一方，提案手法では，系列全体を比較した際に，杉並区と武蔵野市はいずれも互いの市区や，新宿区，三鷹市など

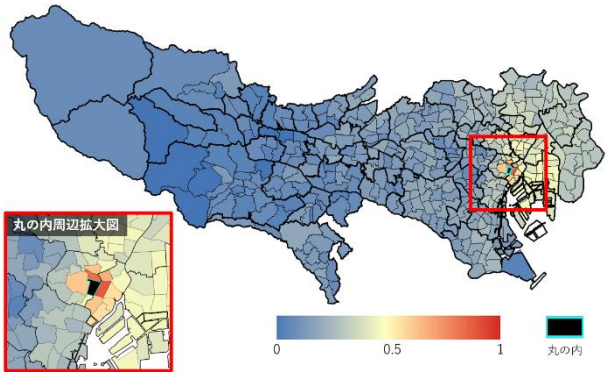
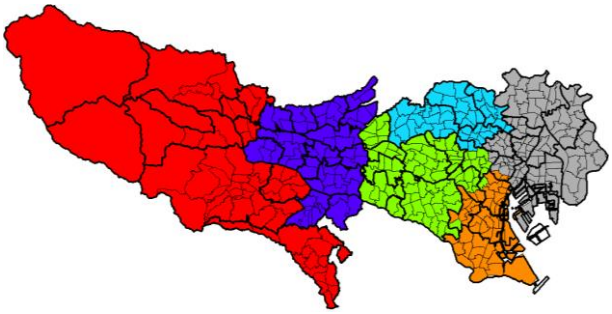
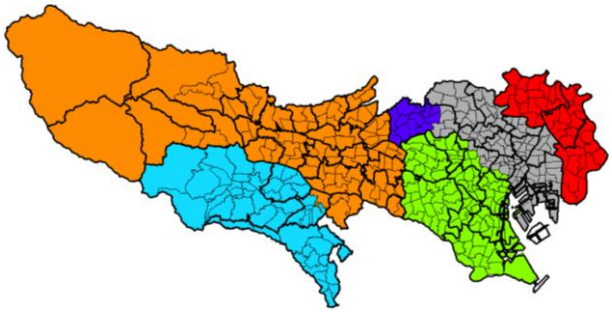


図-2 千代田区丸の内のベクトル表現に対する各地域の cos 類似度



(a) 提案手法（Word2Vec）



(b) モジュラリティ最大化法

図-3 各手法における分類結果（6 分割）

の地域への滞在を系列内に含んでいることが評価され，杉並区と武蔵野市が同一の地域群として抽出されたものと考えられる．このように系列全体を分析対象としている提案手法では，既往研究で用いられている手法とは異なる地域群を抽出した．これは，モジュラリティ最大化法など既往研究のほとんどが地域間の流動の多さに注目して地域を分類するのに対し，提案手法による地域分類では，前後の滞在地域が類似している地域を類似していると評価し，同一地域群として抽出するためと考察される．

3.3. 目的を区別した系列データによる分析

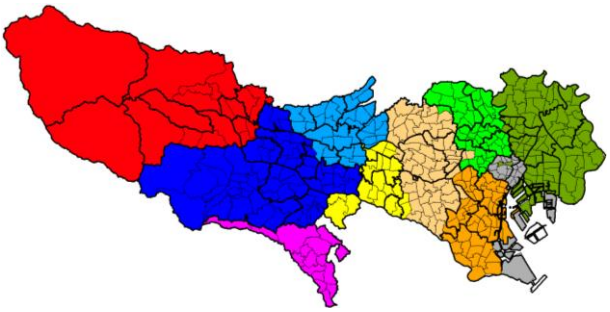
疑似人流・トリップデータに付与されている 7 種類の目的「自宅」「通勤」「通学」「買い物」「食事」「通院」「業務」「フリー」を滞在地域に付与して移動系列を作成し，地域ごとに目的別 7 種類，計 2,912 のベクトル表現を学習した．また，分析条件については 3.2.と同様，ベクトル表現の次元数を 100，ウィンドウサイズを 20 とした．

本節ではまず，ベクトル表現に付与するラベルによって，類似する地域が異なることを地域分類を通して示す．次に，これらのベクトル表現間で，演算が可能であることを示す．

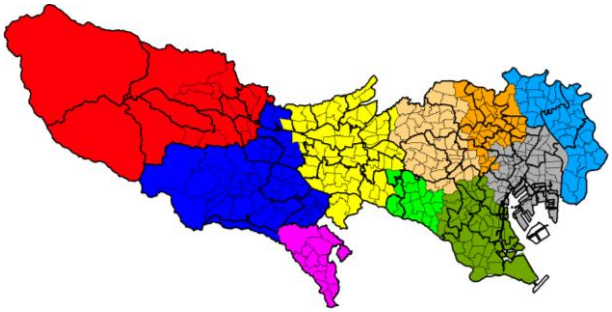
(1) 目的ごとの地域分類

図-4 は各目的のベクトル表現で東京都を 10 地域に分類した結果を示す．いずれも飛び地がほとんどない，解釈しやすい分類結果を得ることができた．私事目的である食事と買い物では分類結果の全体的な傾向が似ているが，それ以外の目的では，分類結果が異なることが観察される．

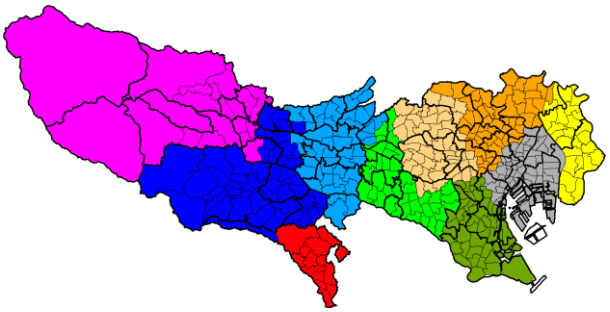
例えば自宅ベクトルによる分類結果は，町田市を 1 つの地域群として抽出したのに対し，他の目的による分類結果や目的を区別しない分類結果（図-3(a)）では，町田市西部を八王子市と同一の地域群に分類した．これは，町田市西部に業務や買い物などで訪れる系列の中には，八王子市への滞在もあるのに対し，町田市西部を自宅，つまり出発地とする系列には，八王子市への滞在が現れることが少ないことが顕れたと考察される．



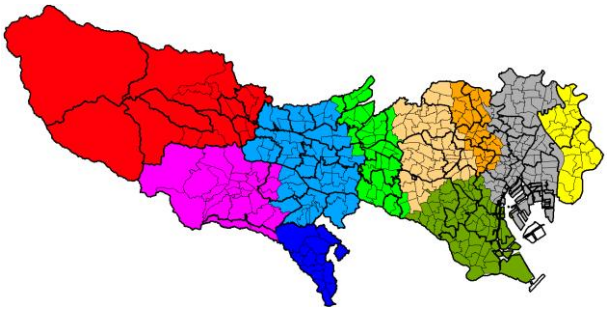
(a) 自宅ベクトル



(b) 食事ベクトル



(c) 買い物ベクトル



(d) 業務ベクトル

図-4 目的別地域分類

(2) 目的別地域ベクトルによる演算

得られたベクトル表現を用いた演算を行った結果，試行したほとんどすべての演算において，解釈のしやすい結果が得られた．表-2 にその一例を示す．表-2(a) の計算では，計算式より「新宿駅東口【買い物】」が出力されることが期待されるが，本手法ではこれ

を最も類似したベクトルとして出力できた．表-2(b)でも同様に，板橋と中野という異なる地域間，【自宅】と【通勤】という異なる目的間で，解釈のしやすい演算ができている．このことから，ラベル別のベクトル表現を用いた演算が可能であるとともに，【買い物】【自宅】といったラベル間のベクトル値の差分が，各地域である程度類似していると考えられる．

本研究ではこの差分ベクトルを地域分類に活用する分析として，以下の二つを試行した．

- 買い物ベクトルと自宅ベクトルの差分ベクトルを用いた地域分類による，商業集積地の抽出（図-5(a)）．
- 通勤ベクトルと自宅ベクトルの差分ベクトルを用いた地域分類による，ビジネス集積地の抽出（図-5(b)）．

分析の結果，両方の分析とも赤色で示されている地域群について，山手線内側の地域や羽田空港周辺といった，商業，ビジネスが集積している地域を抽出することに成功した．しかし，それ以外の色の地域群については近隣地域同士で地域群が形成されており，地域群内部でどういった特徴が類似しているのかは考察が難しいほか，赤色で抽出された地域は，住宅としての利用がほとんどないエリアであるため，こういった地域では，自宅ベクトルが住宅街としての地域の性質を学習できていない可能性がある．以上のことから，ベクトルの差分を利用した地域分類は集積地検出のような発展的な分析へ応用できることが期待されるが，その結果の意味解釈や正当性についてはさらなる手法改良や結果に対する考察の余地が残るといえる．

表-2　目的別ベクトル表現の演算例

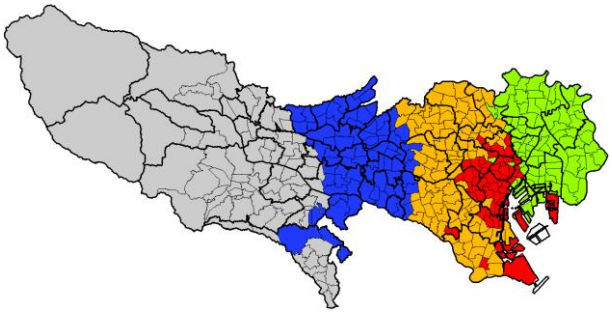
(a)「丸の内【買い物】」
－「丸の内【業務】」
＋「新宿駅東口【業務】」

順位	単語	cos類似度
1	新宿駅東口【買い物】	0.845
2	新宿駅東口【食事】	0.814
3	新宿駅東口【フリー】	0.792
4	千駄ヶ谷【買い物】	0.774
5	新宿駅東口【病院】	0.771

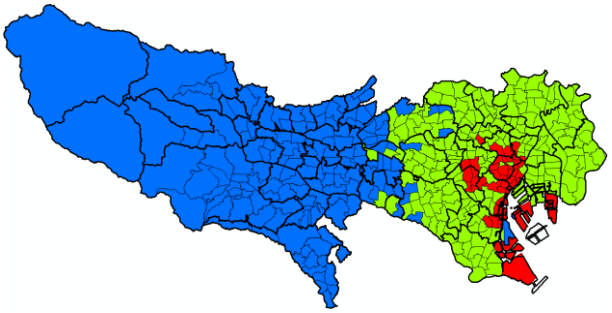
	新宿駅東口【通勤】	0.656

(b)「板橋【自宅】」
－「板橋【通勤】」
＋「中野【通勤】」

順位	単語	cos類似度
1	中野【自宅】	0.647
2	沼袋【自宅】	0.548
3	中野区中央【自宅】	0.542
4	中野【食事】	0.540
5	高円寺【自宅】	0.538



(a)　「買い物ベクトル」―「自宅ベクトル」



(b)　「通勤ベクトル」―「自宅ベクトル」

図-5 ベクトル演算を活用した地域分類

４．まとめと今後の課題

本研究では，自然言語処理における単語ベクトルの学習手法である Word2Vec を応用し，移動系列データをもとに地域分類を行う手法，およびラベル別地域ベクトルの演算を活用した分析手法を提案した．疑似人流データにおける実験の結果，提案手法は移動系列に基づく地域間の類似を評価し，既存手法とは異なる地域分類結果を出力した．また，ベクトル演算やそれを活用した地域分類は，検討の余地があるものの，解釈のしやすい結果を出力することができた．

今後の研究課題としては，遠方からの移動をより評価できるように手法を改良することが挙げられる．本手法により得られた分類結果はいずれも近隣同士で同じ地域群に分類される傾向が強かった．これはトリップ数が距離に比例して減少することによるものであると考察され，例えば買い物のための移動は，商業集積地への距離の長い移動よりも，近所の商店への短い移動の数が多いため，近隣の地域同士で買い物ベクトルは類似が評価されやすい．そのため集積地検出のような地域分類を行うためには手法の工夫が必要である．具体的には，長距離トリップを含

む系列データの，繰り返し学習の回数を増やすことや，分析対象領域の拡大により，集積地への長距離移動を含む系列数を増やすことを検討したい.

謝辞

本研究は，東京大学 CSIS 共同研究（No. 1184）による成果である（利用データ: 擬似人流・トリップデータ 東京都データセット（CSIS 人の流れプロジェクト事務局提供)）.

参考文献

Farmer, C. J. Q. and Fotheringham, A. S. (2011) Network-based functional regions, *Environment and Planning A*, Vol.43, pp. 2723–2741.

Goddard, J. B. (1970) Functional regions within city centres: A study by factor analysis of taxi flows in central London. Transact, *Inst. Brit. Geogr*, Vol.49, pp.161–181.

Hajiseyedjavadi, S., Lin, Y. R., and Pelechrinis, K. (2018) Discovering functionality of urban regions by learning low-dimensional representations of a spatial multiplex network, *In Proceedings of the Third Mining Urban Data Workshop*, MUD 2018.

Inoue, R. and Tsukahara, M. (2016) Travel pattern analysis from trajectories based on hierarchical classification of stays, In: *Ninth International Conference on GIScience Short Paper Proceedings*, pp.150–154.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013) Distributed representations of words and phrases and their compositionality, *Advances in neural information processing systems,* 26.

Noulas, A., Scellato, S., Mascolo, C., Pontil, M. (2011) Exploiting semantic annotations for clustering geographic areas and users in location-based social networks, In: *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5, No. 3, pp. 32–35.

Versichele, K., Groote, L., Bouuaert, M., Neutens, T., Moerman, I., and Weghe N (2014) Pattern mining in tourist attraction visits through association rule learning on Bluetooth tracking data: A case study of Ghent, Bergium, *Tourism management,* Vol. 44, pp. 67–81.

福本潤也，岡本住洋 (2012) コミュニティ抽出法と空間相互作用モデルを組み合わせた機能地域区分手法の提案，土木学会論文集 D3, Vol.68, pp.427-436.