

マイクロアグリゲーションを用いた匿名化による 確率的潜在空間意味解析

Probabilistic Latent Spatial Semantic Analysis by Anonymization using Microaggregation

井手 絢絵^{*1*2} 川本 達郎^{*2} 山下 和也^{*2} 本村 陽一^{*1*2}

^{*1}東京工業大学 情報理工学院

School of Computing, Tokyo Institute of Technology

^{*2}産業技術総合研究所 人工知能研究センター

Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology

With the era of big data, we have to solve the problem of privacy protection in personal data including a lot of sensitive personal information. Therefore, it is necessary to simultaneously satisfy protecting individual privacy and maintaining the usefulness of data. Various anonymization techniques are used for government statistical micro data, and in recent years, there is research on micro aggregation which is the technique of grouping similar records and replacing attribute values with representative values. In this research, we associated personal data with zip codes and perform micro aggregation with spatial information as an example of micro aggregation using senior citizen questionnaire data. We convert each individual's record into spatial information, and extract potentially meaningful spatial information. I make a proposal this analysis as probabilistic latent spatial semantic analysis by anonymization using microaggregation.

1. はじめに

ビッグデータ時代を迎え、機微な個人情報をもく含むパーソナルデータにおけるプライバシー保護の議論は尽きない。パーソナルデータの提供には特定、連結、属性推定などのリスクが伴う。故に個人に関する情報が含まれている場合はデータの有用性を維持しながら個人のプライバシーを守るという、相反する命題を同時に満たす必要がある。諸外国においては政府統計マイクロデータに対して様々な匿名化技法が用いられており、近年のヨーロッパ諸国では、類似したレコード群をグループ化し属性値を代表値に置き換える操作を行うマイクロアグリゲーションに関する研究が行なわれている。我が国ではマイクロアグリゲーションの実証的な検証として、全国消費実態調査と家計調査を用いて、マイクロアグリゲートされた量的属性への定量的な評価がなされている [伊藤 2014]。

本稿では、高齢者アンケートデータを用いて、マイクロアグリゲーションの一例として空間情報によるマイクロアグリゲーションを行い、パーソナルデータを郵便番号に対応付ける。各個人のレコードを空間情報に変換し確率的潜在意味解析 (pLSA) [Hoffman 1999] を用いて潜在的に意味のある空間情報を抽出し、抽出された地域特性をもとに地理情報に結び付けられた匿名加工情報の分析を行う。

2. 日本老年学的評価研究 (JAGES) プロジェクト

2.1 プロジェクト概要

現在、高齢者ケア政策の基礎となる科学的知見を得る目的で、日本老年学的評価研究 (JAGES) というプロジェクトのもと全国の 100 を超える市町村と共同研究協定を締結し 10 万人以上の高齢者を対象にして大規模なアンケート調査が行なわれている。2017 年 2 月、日本学術会議第 23 期学術の大型研

究計画に関するマスタープランとして本プロジェクトの研究構想「ビッグデータと AI で拓く健康長寿社会」が採択され (臨床医学分野, 計画番号 38), ビッグデータベース (人口・地理情報や健診, 電子カルテ, レセプト, 介護認定・死亡データなど) 構築とデータ解析による保健・医療・介護における需要と供給の予測を目指すと共に, JAGES のデータ解析にあたり産業技術総合研究所人工知能研究センターが参画している。

2.2 データ概要

本分析の対象データは JAGES2010-2013 コホートデータ (2010 横断データに「要介護認定データ」と「介護保険賦課データ」を結合したもの) である。これは 2010 年調査対象の 24 市町の回答者のその後の要介護認定・死亡を約 3 年間追跡している。アンケートの質問群には、コア項目とバージョン項目で構成される。コア項目は全回答者共通の項目であり、コア項目に付随して A-E 版の 5 種類のバージョン項目が均等割付された。アンケート項目の概要を図 1 に示す。

調査項目 2010

コア項目	バージョン項目
身体状況 機能状態	• 罹患 保健行動 • BMI 転倒状況
心理	• うつ 幸福度
社会	• ソーシャル・ネットワーク ソーシャル・サポート
社会経済 的地位	• 年間世帯所得 世帯人数 教育 就職 (最長職) 年金 生活保護
会・グル ープへの 参加	• ボランティア スポーツ 町内会・自治会 趣味 他
地域環境	• (地域に対する) 信頼 互酬性 治安 祭り 近所付き合い
外出	• 外出頻度 交通手段
	A • 家族介護・治療・ 生活様式
	B • 口腔・楽観性尺 度・主観的幸福度
	C • 地域ソーシャル キャピタル・虐待
	D • 主観的生活程度・睡 眠・認知症・過去の様 子・入浴習慣
	E • 運動 (柏市・名古屋 市・神戸市のみ)

図 1: JAGES アンケート項目概要

連絡先: 井手絢絵, 国立研究開発法人産業技術総合研究所 人工知能研究センター, 東京都江東区青海 2-4-7, 03-3599-8914, ide.ayae@aist.go.jp

2.3 先行研究

本センターの井手らは JAGES データから確率的潜在意味構造モデルを構築して高齢者行動の分析を行った [井手 2016]. 確率的潜在意味構造モデルとは pLSA とベイジアンネットワークを組み合わせた確率モデルであり, pLSA により抽出された潜在クラスと説明変数間の関係を非巡回有向グラフで構造化し条件付き確率で定量的に表現したものである. pLSA で高齢者とアンケート項目をクラスターリングし, 潜在クラスを抽出した結果, AIC 基準によりクラス数 7 を採用した. 表 1 が pLSA で得られた結果である. クラス名は抽出された質問項目群より共通する傾向を把握しラベリングを行っており, 所属確率 0.8 以上の項目は項目数に記してあるが, その中で一部抜粋し所属する代表的質問項目として記した.

表 1: 各質問項目が所属するクラス

クラス	各クラスに所属する代表的質問項目	項目数
Z001	友人や近隣とのサポート受領・提供あり	7
Z002	サポート受領・提供無し, 土地勘無し	23
Z003	飲酒頻度, 趣味(釣り), 性別(男性)	7
Z004	同居人(子供・孫), 食事相手(子供・孫)	12
Z005	各交流イベントへの参加頻度, 趣味お稽古	18
Z006	気持ちの混乱や否定的な感情を持つことが無い	5
Z007	介助の有無, 治療中の病名, 否定的な感情	27

上記の pLSA によって得られた潜在クラスと, 質問項目内の習慣状況や社会環境, その他の説明変数間の関係をベイジアンネットワークにより明らかにした. モデルの構築は, ベイジアンネットワーク構築ソフトウェア Bayonet [本村 2003] である. ベイジアンネットワーク上では, 確率推論を実行し, 事後確率の変動から行動傾向を把握することが可能である. 例示した図 2 の確率推論では, Z006 (表 1 参照) のクラスに所属後の, クラスに結びついた質問項目の事後確率の推移を確認した. Z006 に所属する場合, $alcl3.10 = 1$ (飲酒をする) が 0.3462 から 0.1919 に下降し, $ml2sp10 = 1$ (食事相手が配偶者) が 0.6402 から 0.7428 まで上昇し, Z006 に所属するポジティブな高齢者のライフスタイルとして配偶者と食事をし, 飲酒をしないことが推定できる. このように, 説明変数間の関係や行動傾向が定量的に把握され, 高齢者行動モデルを構築した研究である.

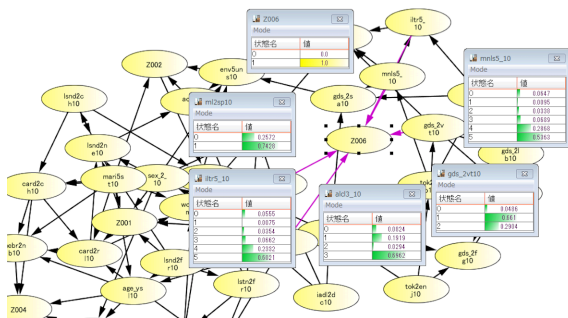


図 2: 構築したベイジアンネットワーク上での確率推論実行

3. 確率的潜在空間意味解析

3.1 確率的潜在意味解析 (pLSA)

確率的潜在意味解析 (pLSA: probabilistic Latent Semantic Analysis) は文書分類の手法として提唱され, 1 つの文書が複数のトピックを持つと仮定するトピックモデルを最尤推定する手法である. 文書 d 中の単語 w が潜在変数 $z \in Z = \{z_1, \dots, z_k\}$ を介して生成されると仮定し, EM アルゴリズムによる尤度最大化で対数尤度の下限を最大化するパラメータを求め, 共起するデータに潜在変数を付随させる. アンケートデータに適用する場合では, アンケート回答者 w が質問項目 d に潜在変数 z を介して回答すると仮定する. 同時確率は以下の式で表される.

$$P(w, d) = \sum_{k=1}^K P(d|z_k)P(z_k|w) \quad (1)$$

その後, 以下の対数尤度関数を EM アルゴリズムによって最大化することで, 潜在クラスを抽出する.

$$L = \sum_{i=1}^N \sum_{j=1}^M n(w_i, d_j) \log P(w_i, d_j) \quad (2)$$

アンケートデータにおける回答の似通っている回答者群と質問項目群を抽出することが可能であると同時に, 高齢者と質問項目を分析可能な数のクラスに圧縮することで, 情報量を維持したまま特性の抽出が可能である.

3.2 空間情報によるマイクロアグリゲーション

マイクロアグリゲーション [Defays, D., and M. N. Anwar 1998] とは, ある数値属性に着目してレコードをグループ化し, その属性値を代表値に置き換える匿名化技法である. 本研究では, 近傍に居住する人々は回答傾向が類似するという仮定の下に, 各郵便番号のレコードをグループ化し高齢者 ID を郵便番号に置き換えた.

3.3 確率的潜在空間意味解析の提案

空間情報によりマイクロアグリゲートされたデータから pLSA を用いて, 個人ではなく空間についての潜在クラスを抽出する方法, 確率的潜在空間意味解析を提案する. 郵便番号に対応する地区 w が質問項目 d に潜在変数 z を介した回答傾向があると仮定し, 潜在的に意味のある空間構造を抽出した.

4. JAGES データにおける潜在空間意味構造の抽出

4.1 確率的潜在空間意味解析結果

前述したように, 本研究では JAGES2010-2013 コホートデータ (対象市町村は愛知県常滑市, 名古屋市, 知多北部, 阿久比町, 武豊町, 美浜町, 南知多町, 碧南市, 西尾市, 一色町, 旧吉良町, 旧幡豆町, 長崎県松浦市, 山梨県中央市, 三重県大会町, 千葉県柏市, 北海道大雪, 青森県十和田市, 宮城県岩沼市) を用いた. 本データセットに郵便番号は本来付随していないが, 本研究の意義に賛同して頂き, 郵便番号が結合されたデータセットを提供して頂いた. JAGES2010-2013 コホートデータ総人数は 74264 人, うち郵便番号が結合されたのは 53801 人 (結合率 72.4%) であった. この郵便番号結合データに pLSA を用いてクラス抽出をした. クラス数は AIC 基準で決定した. pLSA の尤度計算に EM 法を用いるため, 初期値依存性がある. そのため, 初期値を 5 回変更して潜在クラスを 4 から 45 まで増加させ各クラス数での最小値を比較し

た。その結果、 $K = 25$ の時に AIC が最小値をとるので、クラス数 25 を採用した。抽出された 25 クラスのうちの一部を例に挙げる。抜粋した 4 つのクラスターの所属確率上位 10 位質問項目の表 2 に列挙した。Area02 では主に家から 1 キロ以内の地域環境、Area06 では一人暮らしの高齢者に関する項目、Area17 では地域で感じた変化と地域交流の頻度、Area24 では子どもや孫と同居が想定される項目により地域が特徴付けられている。

表 2: 各質問項目が所属するクラス

Area02		Area06	
$P(d/Area)$	質問項目	$P(d/Area)$	質問項目
0.764	あなたの住む地域で感じた変化 (転入者の増加)	0.523	住宅の種類 (社宅・寮)
0.396	ボランティアグループへの参加頻度 (ほぼ毎日)	0.520	婚姻状態 (未婚)
0.346	あなたの住む地域で感じた変化 (地域経済の活性化)	0.424	手段的サポート受領 (友人)
0.293	家から 1 キロ以内に交通事故の危険が多い道路や交差点がたくさんある	0.403	世帯人数 (一人)
0.277	業界団体・同業者団体への参加頻度 (ほぼ毎日)	0.403	あなたの同居人 (一人暮らし)
0.271	家から 1 キロ以内に落書きやゴミの放置が目立つところがたくさんある	0.371	この 1 年の出来事 (一人暮らしを始めた)
0.264	家から 1 キロ以内に夜の一人歩きが危ない場所がたくさんある	0.357	住宅の建て方 (長屋建)
0.221	地域の治安について不安を強く感じている	0.348	手段的サポート提供 (友人)
0.217	あなたの住む地域で感じた変化 (祭りの復活)	0.320	あなたの住む地域で感じた変化 (地域経済の沈滞)
0.176	家から 1 キロ以内に新鮮な野菜や果物が手に入る商店・施設がたくさんある	0.285	野菜や果物を食べる頻度 (週 1 回)

Area17		Area24	
$P(d/Area)$	質問項目	$P(d/Area)$	質問項目
0.445	あなたの住む地域で感じた変化 (地域住民の活動や交流の衰退)	0.673	世帯人数 (6 人)
0.434	あなたの住む地域で感じた変化 (祭りの衰退)	0.642	あなたの同居人 (孫・ひ孫)
0.273	あなたの住む地域で感じた変化 (行政サービスの低下)	0.624	世帯人数 (5 人)
0.271	スポーツ関係のグループやクラブ (年に数回)	0.594	あなたの同居人 (子どもの配偶者)
0.256	あなたの住む地域で感じた変化 (失業者の増加)	0.561	世帯人数 (7 人)
0.242	あなたの住む地域で感じた変化 (所得の格差の拡大)	0.552	食事相手 (孫)
0.235	あなたの住む地域で感じた変化 (貧困者の増加)	0.444	情緒的サポート受領 (同居の子ども)
0.220	この 1 年の出来事 (経済的な困難が増した)	0.442	情緒的サポート提供 (同居の子ども)
0.207	町内会・自治会への参加頻度 (年に数回)	0.433	世帯全体の合計年間収入金額 (1000-1200 万円未満)
0.191	業界団体・同業者団体への参加頻度 (月に 1-2 回)	0.432	手段的サポート受領 (同居の子ども)

対象市町村のうち最も回収数の多い名古屋市について考察する。名古屋市を構成する各 16 区において、各 25 クラスに所属する郵便番号の数を集計し、郵便番号の分布が最上位のクラスをその行政区を代表するクラスとし、コロプレス図を作成し図 3 に示した。各クラスに所属する質問項目は以下の通りである (抜粋)。

Area03{ 趣味関係のグループへの参加頻度 (週 1 回程度、月 1-2 回)、この 1 年の出来事 (新しい友人ができた)、よく会う友人や知人との関係 (趣味や関心が同じ友人)、食事相手 (友人)、趣味お稽古 (絵画、グランドゴルフ)}

Area09{ 家から 1 キロ以内に魅力的な景色や建物がある、家から 1 キロ以内に運動や散歩に適した公園や歩道がある、家から 1 キロ以内に坂や段差など歩行が大変なところがある、趣味お稽古 (パソコン、囲碁、写真撮影)}

Area25{ 趣味関係のグループへの参加 (参加していない)、よく会う友人や知人との関係 (いない)、この 1 か月何人の友人・知人と会ったか (1-2 人)}

図 3 に示すように、近傍の区域は代表するクラスが等しく、居住区域が近接していると回答傾向が類似的であることが確認された。すなわち、JAGES 調査票から把握可能な高齢者

の心理・社会的状況や会・グループへの参加等が居住区域と相関があると推定される。

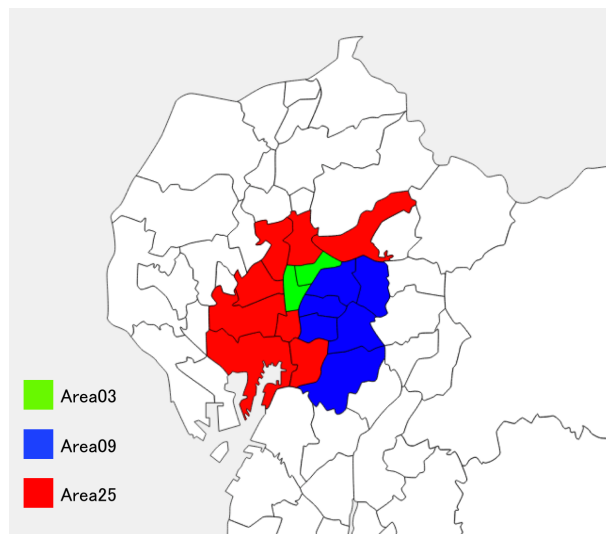


図 3: 名古屋市区別クラス分布

5. まとめ

郵便番号でマイクロアグリゲーションした匿名加工情報から、地域特性の潜在クラスを抽出した。公衆衛生のコンテキストにおいて地理情報は非常に重要であり、本研究の分析結果においても高齢者の生活行動と居住地域の関連が確認され、公衆衛生における地理情報利活用の意義を裏付ける形となった。今後の課題としては、(1) 郵便番号界地図を用いた詳細分析並びに、(2) 郵便番号から把握可能な空間内の施設の把握、(3) JAGES の縦断データを分析することによる潜在的な時空間の意味解析を試み、地域の経年変化がもたらす高齢者の生活行動の変化を明らかにしたいと考えている。また、JAGES の研究成果を地域での介入に反映するため、GIS(地理情報システム) 活用で地域に最適化された介入を可能とする社会実装を含めた研究を展開する予定である。

6. 謝辞

本研究行うにあたり、日本老年学的評価研究 (the Japan Gerontological Evaluation Study, JAGES) プロジェクトのデータを使用しました。JAGES 内における研究会でのご指導や人工知能ワーキンググループの立ち上げに協力してくださいました代表の近藤克則先生、郵便番号結合データセットを提供して下さいました辻大士先生をはじめとした研究者、事務局の方々など、JAGES プロジェクトに関わる多くの方々にご助力いただきましたことを、心より感謝申し上げます。

また、本研究は NEDO 委託事業「人間と相互理解できる次世代人工知能技術の研究開発」の支援を受けたものです。

参考文献

[Defays, D., and M. N. Anwar 1998] Defays, D., and M. N. Anwar. "Masking microdata using micro-aggregation." Journal of Official Statistics 14.4 (1998): 449.

[Hoffman 1999] Hofmann, Thomas. "Probabilistic latent semantic analysis." Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1999.

[井手 2016] 井手絢絵, 本村陽一:高齢者の生活行動推論のための大規模アンケートデータを用いた確率的潜在意味構造モデル構築, 人工知能学会「社会における AI」研究会第 27 回研究会 (2016).

[伊藤 2014] ITO, Shinsuke, Mariko MURATA, and Masahiro TAKANO. "マイクロデータにおける匿名化技法の適用可能性の検証—全国消費実態調査と家計調査を用いて—伊藤伸介, 村田磨理子, 高野正博." 統計研究彙報 (2014): 83.

[本村 2003] 本村陽一. "ベイジアンネットソフトウェア BayoNet." 計測と制御 42.8 (2003): 693-694.