メッシュデータを用いた大阪府の人口の社会増減に関連する因子の抽出
高相関データに対する機械学習を用いたボトムアップな特徴量統合手法の提案

瀧澤重志*

Extraction of Factors Related to Social Population Growth and Decline in Osaka Prefecture Using Mesh Data
Proposal of a Bottom-up Feature Integration Method Using Machine Learning for Highly Correlated Data

Atsushi Takizawa*

Abstract: In the context of Japan's population decline, this study investigates social demographic change - specifically migration-related population shifts - using Osaka Prefecture as a case study. Utilizing detailed census-derived data, we apply ensemble tree-based machine learning models, which are robust to multicollinearity, to analyze a large number of highly correlated variables. SHAP values are used to assess variable importance, followed by non-linear dimensionality reduction on the SHAP correlation matrix to extract latent factors. This approach enables a data-driven identification of underlying drivers of social mobility, beyond conventional life event-based explanations.

Keywords: 国勢調査（census），人口の社会増減（Social population growth and decline），機械学習（machine learning），SHAP（SHAP），UMAP（UMAP）

１．はじめに

日本の人口減少に関連して，例えば社人研（2024）は2050年までの予測人口を公表しており，各地の将来像の策定のための基礎資料として活用されている．人口変化は出生・死亡に起因する自然増減と，人の移動による社会増減に分けられる．前者はコホート増減法で精度よく推計できることが知られている．一方社会移動について，年代別のライフイベントの変化に伴う移動は知られているが（高取, 2018），それ以外の要因についてはあまり調べられていない．

本研究は，この社会増減に関して，人口の流入出が活発な大阪府を対象として，詳細な空間解像度で関連要因を抽出することを試みる．そのために，主に国勢調査の派生データを用いて多数の説明変数に投入する．国勢調査のデータは人口数がベースとなっているため相関が高いものが多い．重回帰系のモデルは，相関が高い変数を同時に投入すると多重共線性の問題を生じ，求解や解釈が困難なモデルとなる．そのため，事前に相関の高い変数を削除したり，PCAなどで変数を次元削減するといった前処理が行われる．しかし，多様な情報を捨象しすぎたり，解釈が困難となる危険がある．そこで本研究では，相関が高い変数を投入しても，多重共線性の影響をあまり受けない，機械学習の決定木のアンサンブル型モデルを使用する．変数の寄与度の説明に広く使われるSHAP（Lundberg and Lee, 2017）を用いて，変数選択を行ったのち，残った変数のSHAPの相関行列から非線形次元圧縮を行い，変数をボトムアップに統合する手法を提案する．このような方法で，多数の相関が高い変数から，社会移動に関して，これまでの前提に依らない因子を抽出することを試みる．

２．方法

現在公開されている日本の国勢調査の最新版は2020年のため，一つ前の2015年の国勢調査や関連するデータを多数用意して説明変数として，2015年から2020年にかけて，人口の社会増減を推計する機械学習モデルを作成する．

２.１　データの整備

空間単位は4次メッシュ（500m）のメッシュデータを用いる．まず，2期の国勢調査の男女別年齢階級別人口のデータから，自然増減と社会増減に人口変化数を分解する．この方法は小池（2010）に準ずるが，一部修正を行っている．図1に推計した大阪府の社会増減分布を示す．

表1に使用した説明変数のデータの概要を示す．まず国勢調査データから導かれる人口と関連する属性として，17のジャンル（年齢，世帯，家族類型，住宅，職業，居

*　正会員　　大阪公立大学生活科学研究科（Osaka Metropolitan University）
〒558-8585　大阪府大阪市住吉区杉本3-3-138　E-mail：takizawa@omu.ac.jp

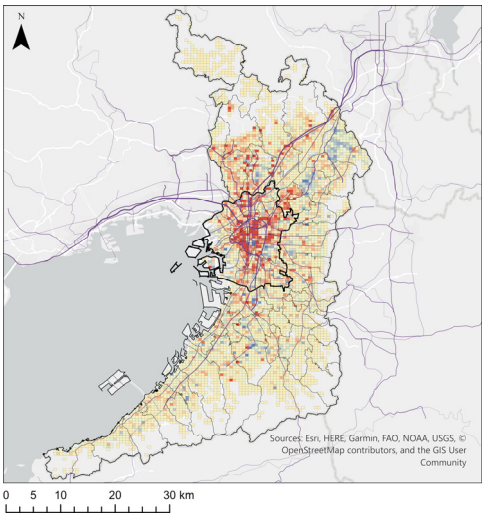


図 1　大阪府の 2015 から 2020 年の社会増減

表 1　使用した説明変数の一覧

タイプ	出典
人口と関連属性 (89)	平成 27 年国勢調査に関する地域メッシュ統計その 1，その 2（2015）
経済活動 (15)	経済センサス 4 次メッシュ（2016）
土地利用 (17)	土地利用詳細メッシュ，都市地域土地利用細分メッシュ（2016）
地価 (1)	地価公示データ，都道府県地価調査データ（2015）
交通 (2)	駅別乗降客数データ，バス停留所データ（2022）
災害 (7)	津波浸水想定データ（2016）
位置 (2)	標準地域 4 次メッシュの中心座標

住歴等）の変数を合計 89 個用意した．都市活動は，規模感を具体的に表すものとして従業員数を用い，居住型の福祉施設の場合は当該メッシュのみの総数を，それ以外の通所型の施設，生活関連施設，教育，医療系施設などは徒歩圏を考慮し当該メッシュと近傍の 8 つのメッシュの合計値を用いた．土地利用は，大阪府内の市街化区域では解像度の高い 50m メッシュを，北部地域は 100mのメッシュを用いて，4 次メッシュ単位で土地利用 17 種類の面積を集計した．地価は大阪府とそれに接する 4 府県の地価公示と都道府県地価調査のポイントデータをマージし，経験ベイズクリギング法により 100m メッシュの解像度でマップを作成し，各メッシュで平均した．交通は各メッシュから 800m の範囲にある駅の 1 日当たり乗降客数の最大値を求めた．さらに，各メッシュに含まれるバス停留所数を合計した．災害は南海トラフに起因する津波の予想浸水深さをマッピングし，7 段階の浸水の階級値の面積をメッシュ単位で集計した．地形は斜面の勾配が歩きやすさなどに影響すると考え，メッシュ内の地形の平均傾斜角度をデータとした．

以上，合計で 133 個の説明変数を用意する．

2.3　分析方法

機械学習モデルとして，一般的に使われる LightGBM（LG）と Random Forrest（RF）に加えて，RF のランダム性を高めた Extra Tree Regressor（ET）の 3 種類を用いて精度を比較する．Python で Scikit Learn の実装を使用した．学習・テスト用のデータをランダムに 8:2 に分け，学習用のデータで 10 回交差検証により R2 を最大化するようパラメータチューニングを行う．

モデル選択は 2 段階で行う．まず，用意した 133 個の説明変数すべてを使い，各モデルでパラメータチューニングを行い，交差検証の精度を求める．加えて全データで学習を行い，各変数の SHAP を求める．SHAP の平均絶対値が小さい 3%の変数を削除したうえで，上記の手順を繰り返し，交差検証の R2 が最大になったモデルを選択する．ここまでが Step 1 である．

SHAP は目的変数に対する加法性があるため，相関が高い変数間に SHAP が分配される傾向があり，変数の寄与度を過小評価する危険がある．そこで Step 2 で，選択されたモデルの相関が高い変数を統合する．まず，選択された変数の集合をAとし，変数$a, b \in A$の SHAP 値の積率相関を$r_{a,b}$とし，変数a, bの非類似度を，相関距離

$$d_{ab} = \sqrt{2(1 - r_{ab})}, \quad \forall a, b, \in A$$

で表す．この相関距離を使って，各変数を UMAP（McInnes et al., 2020）で平面に埋め込み，埋め込み座標を DBSCAN でクラスタリングする．そして，各クラスタの変数を，以下の平均絶対 SHAP の重み付き和として統合する．ここで，A_cはクラスタcに属する説明変数の集合，$shap_a$，x_a，$\mathrm{std}_a()$はそれぞれ，変数$a \in A_c$の SHAP 値，データの値，標準化関数を示す．

$$x_c = \sum_{a \in A_c} \frac{\overline{|shap_a|}}{\sum_{b \in A_c} \overline{|shap_b|}} \cdot \mathrm{std}_a(x_a)$$

最後に，統合した変数を用いて，各機械学習モデルにより，再度パラメータチューニングと精度評価を行う．

3．結果

Step 1 による各モデルの精度を表 2 に示す．各機械学習モデルでのオリジナルモデルの精度評価結果を示す．手順で変数を削除していき CV の R2 が最大となったときのモデルの結果を示している．CV やテストデータの R2 は 0.3 前後で，予測には厳しい．一方，学習データの精度は 0.7～0.8 で，説明のモデルとしては妥当である．CV では LG，テストでは RF の精度が高いが，大きな差はない一方，全データでは ET の精度が高い．

表 2　Step1 での各モデルの精度

Model	$\|A\|$	CV (mean ± std)	Test	Full
LG	68	**0.302 ± 0.044**	0.279	0.712
RF	51	0.285 ± 0.042	**0.284**	0.692
ET	53	0.290 ± 0.048	0.269	**0.822**

次に Step 2 を適用する．例として ET モデルで選択された変数の UMAP による低次元化の結果を図 2 に示す．各属性が平面上で比較的明瞭に分離されており，距離が近い変数同士で，13 個のクラスタを得た．

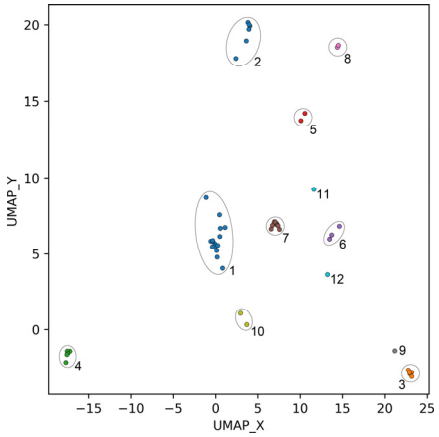


図 2　ET モデルの 53 変数の UMAP と DBSCAN による次元削減

Step2 による精度を表 2 に示す．交差検証の精度は ET，テストは RF，全データの精度は ET が高い．これらのモデルはオリジナルモデルと精度が上昇するか低下が微小であり，変数の統合による情報の欠損の影響は少ないといえる．一方 LB は特にフィッティング精度が大きく下がっており，統合による情報欠損の影響が大きいといえる．以降では，ET による結果が最良とみなし，最終モデルとして，その結果を詳しく分析する．

表 3　Step2 による各モデルの精度

Model	$\|A\|$	CV(mean ± std)	Test	Full
LG	17	0.293±0.042	0.282	0.498
RF	13	0.287±0.068	**0.292**	0.658
ET	12	**0.297±0.035**	0.279	**0.782**

表 4 は最終モデルの変数の要約である．集約された属性は平均絶対 SHAP 値が高い順に掲載している．各変数の summary plot を図 3 に示す．色分けは，赤が SHAP が高いことを青が SHAP が低いことを示している．したがって赤いプロットが 0 より右側に多い属性は，全体として SHAP と人口増減に正の寄与があり，左に多く属性は負の寄与が多いことを示している．さらに上位 3 位までの変数のメッシュ単位での SHAP 空間分布に対して，Getis-Ord Gi*統計量を用いた Hotspot 分析を行った（図 4）．距離バンドは隣接エッジのみで信頼度 90%以上のホットスポットとコールドスポットを抽出した．

表 4　変数統合後のクラスタ名

#	クラスタ名
1	高所得ホワイトカラー層，高流動地域
2	若年世帯の地域内移動者
3	大学生
4	若年単身者向け賃貸住宅
5	中層の公営賃貸住宅
6	超高齢化地域
7	パートタイム，ブルーカラー層
8	50 代から 60 代前半
9	経度
10	津波浸水区域（0.3～2m）
11	空地
12	有料老人ホーム

以上の図表からクラスタの特徴を分析する．最も寄与度が高い#1 は，大阪市内中心から主要鉄道路線沿線の地価が高く，管理職や専門職などのホワイトカラーが多く住んでいる．また，県外からの移住者が多く流動性が高いといった特徴がある．#2 は若年世帯の県内移動者が多いクラスタである．低層もしくは高層の共同住宅の居住者が多かったり，幼稚園従業者が近隣地域に多く，人口流入傾向が高い．立地は大阪市内の中心部に集中している．#3 は大学生のような若者が流出していることを示している．空間分布は有名大規模大学の周囲に多く，卒業とともに引っ越していることが推測される．#4 は若年層・単身者向け賃貸住宅クラスタであり，高密度な低層住宅や，高層マンションへの居住者が多いエリアで，流入の傾向がある．#5 は，公営の賃貸住宅で 5 階建て程度の中層集合住宅が目立つエリアである．これは日本において古くから建設された公営住宅であり老朽化した建物が多く，地域から流出する傾向が高い．#6 は，80 代以上の高齢者が多いエリアである．流入する傾向があり，府内で分散しており，老人ホームへの転居を示唆している．#7 は，パートやブルーカラー，失業者の多さを示すクラスタで，人口流出が目立つ．#8 は中高年のクラスタで，人口流出傾向がある．#9 は経度のクラスタであり，西側ほど人口流入することを示している．#10 は，津波の予測浸水深さが 0.3-2m となるエリアを示している．人工流入の傾向がありリスクを高めている．#11 は空地のクラスタである．大阪府南北の郊外に分布しており，空地だったところが宅地化され人口が流入したことが推測される．#12 は，有料老人ホームの近隣を含む従業者数である．

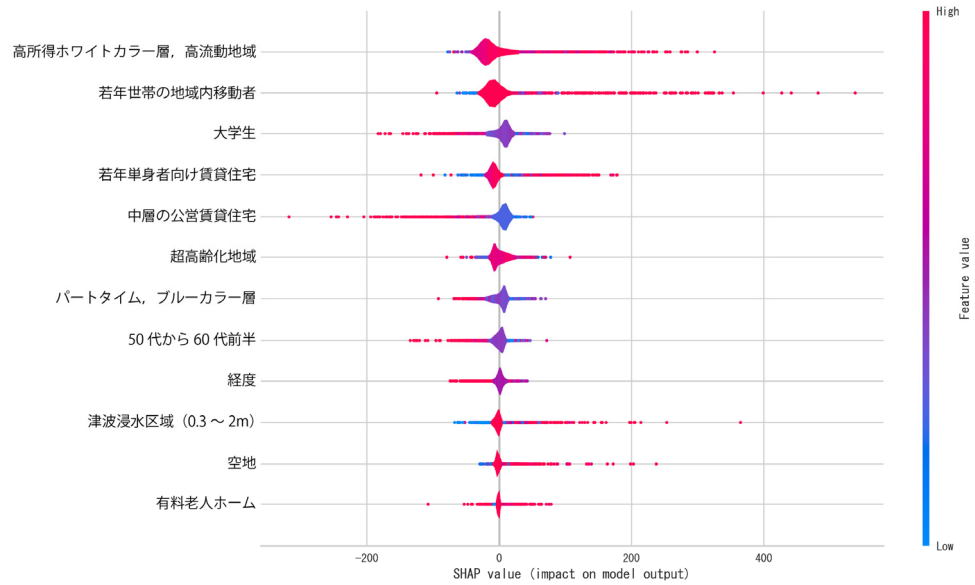


図 3　最終モデル（ET）の SHAP の summary plot

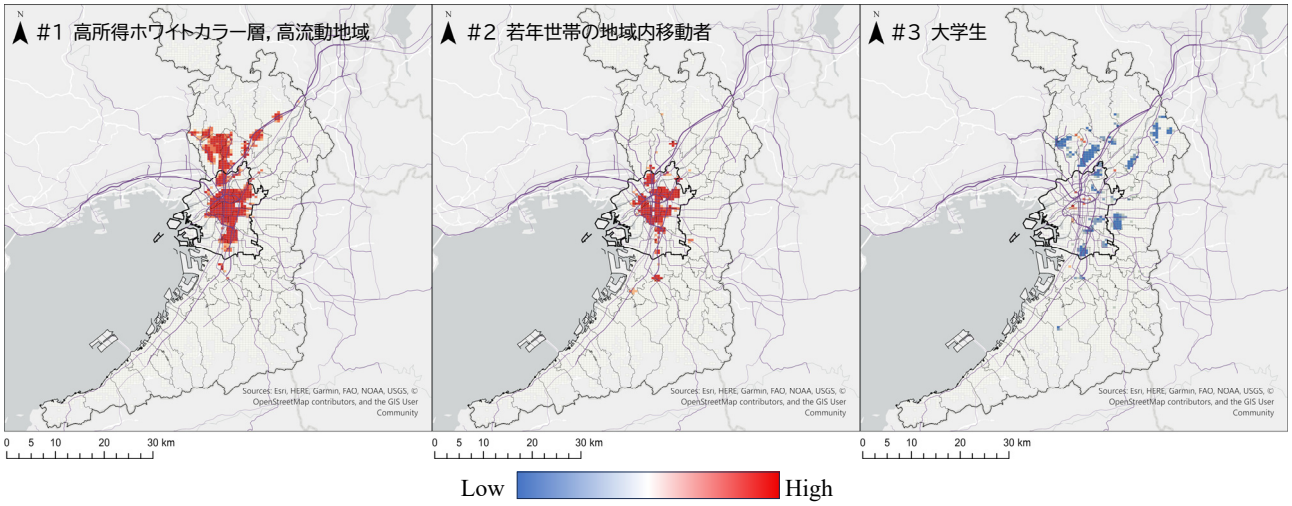


図 4　上位 3 位までの統合変数の Hot/Cold スポットの分布

4．おわりに

本研究は大阪府の直近の国勢調査データからメッシュ単位で社会増減を推計し，多数の相関が高い説明変数から，社会増減と関連する因子を抽出する手法を開発して，分析を行った．これまでは，若者，子育て，老人のライフイベントと関連付けられて理解されることが多かった社会増減に対して，ホワイトカラーの転勤族，老朽化した公共住宅，パート・ブルーカラー，津波浸水地域，空地といった，新規性と解釈性の高い因子が抽出された．今後の課題として，詳細な分析，他地域での検証，人口予測モデルへの結果のフィードバックなどが考えられる．

謝辞

本研究は学術知共創プログラム（#24016566）の補助を受けています．国勢調査と経済センサスデータは東京大学 CSIS の共同研究制度（#1370）で提供いただきました．

参考文献

国立社会保障・人口問題研究所（2024） 日本の地域別将来推計人口-令和 2(2020)～32(2050)年- 令和 5 年推計．人口問題研究資料，349.

高取千佳（2018）人口社会増減と空間指標の関連分析-愛知県を対象として-．日本都市計画学会都市計画論文集，53 (3), 392-399.

Lundberg S. M. ・ Lee, S.-I. (2017) A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS 2017)

小池司朗（2010）首都圏における時空間的人口変化-地域メッシュ統計を活用した人口動態分析-．人口問題研究，66 (2), 26-47.

McInnes L.・Healy J.・Saul N.・Großberger L. (2020) UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 5 (56), 2008.