【研究ノート】

シビックプライド醸成に繋がる住民価値の掘り起こしと貢献度の検証に関する研究

～ uncovering potential resident value ～

山本裕・高田晃希・黒羽晟

橋本沙也加[*1]・橋本尚子[*1]・岡田ゆかり[*1]

Research on uncovering resident values that lead to the fostering of civic pride and the verification of their contribution

Hiroshi Yamamoto, Koki Takada, Jo Kuroha,
Sayaka Hashimoto[*1], Shoko Hashimoto[*1] and Yukari Okada[*1]

Abstract: In this research, we use the case of a local government to address the issue f fostering civic pride. From the results of resident questionnaires of existing local governments, we will use multivariate analysis and machine learning to uncover potential "new resident value" that will lead to the cultivation of civic pride. We will evaluate the contribution of "new resident value" to fostering civic pride, and derive the current living environment conditions to raise awareness of "new resident value".

Keywords: principal component analysis, analysis of qualification type II, cramer's coefficient of association, machine learning

１．はじめに

近年、日本は少子高齢化に伴って東京一極集中が進み、地方の過疎化・高齢化が進行している。地方公共団体には、こうした地方の衰退を抑制する課題がある。そこで、本課題解決のアプローチとして、住民が一丸となって自治行政を推進することで社会増減などに歯止めをかけ、持続可能なまちづくりを推進するために、住民の地域への参加意識、愛着度などを高める活動がある (シビックプライドの醸成)。住民意識のモニタリング方法として、住民アンケートが一般的であり、多変量解析やテキストマイニングなどによりアンケート回答を分析し、シビックプライドの醸成に繋がる「住民が感じる地域の潜在的価値」を導出・評価する研究が拡がっている。研究の多くは潜在的な住民価値を模索したデータ解析であるが、既存のアンケートから導出した潜在的な住民価値とアンケート項目とを関

*1 株式会社百代

連付けて、自治体施策を模索する研究事例は少ない。

本研究では、ある自治体殿を 1 事例としたシビックプライド醸成という課題に対して、既存の自治体住民アンケート結果から、多変量解析や機械学習を用いてシビックプライドの醸成につながる潜在的な「新住民価値」を掘り起こし、「新住民価値」のシビックプライド醸成への貢献度の評価と、「新住民価値」に関する住民意識を高めるための現状の居住環境条件の導出を行う。本研究は今回の事例以外の多くの自治体での訴求も目的とする。

2．自治体住民の新価値の創出

2.1 本研究で取り組む課題

既存の住民アンケートから、新住民価値を掘り起こし、住民意識を高める居住環境条件を導出するために本研究で取り組む課題は以下である。

（a）住民アンケートから潜在的な「新住民価値」を掘り起こす（潜在的テーマの発掘）
（b）「新住民価値」が住民意識を高める変数になっていることの評価
　（潜在的テーマの評価）
（c）「新住民価値」の変数の住民意識を高めるために寄与する居住環境条件の導出
　（潜在的テーマからの施策導出）

2.2 主な研究アプローチ

本研究の目的は、自治体住民のシビックプライド醸成のために、自治体施策や居住環境が住民の新しい価値となるような潜在的なテーマを発掘し、当該テーマに関して住民意識を高めるような自治体の具体的施策を導出することである。進め方としては、ある自治体殿で運用している住民アンケートから、多変量解析や機械学習を活用して住民意識を向上するために寄与する潜在的な「新住民価値」となるテーマを発掘し、発掘テーマに関する住民意識を向上させる環境条件を発見するプロセスとする。具体的には、ある自治体殿にて 2018 年以降に実施された住民アンケートを対象にして、上記 2.1 に示した課題に対応し、（a）潜在的テーマの発掘、（b）潜在的テーマの評価、（c）潜在的テーマに対して住民意識を高める居住環境条件の明確化に対応した事例研究を進める。本事例研究全体の概念図を図 1 に示す。

今回の研究報告は、上記（a）および（b）の評価準備の範囲とし、本範囲において以下の方針で進める。

（1）解析対象のデータは、自治体が従来から運用中の住民アンケートを適用する。
（2）アンケート項目回答と自由記述回答の両方を総合的に解析する手法として、多変量解析（数量化Ⅱ類、相関分析）、テキストマイニング及び機械学習のモデルを適用する。
（3）新住民価値を導出する方法として主成分分析を活用する。合成変数（主成分）の意味付けを行い寄与度の評価を行う。

アンケートデータから導くシビックプライド醸成のための新住民価値は、シビックプライド醸成に寄与する特徴量全体を要約できるような軸であることが求められるため、主成分分析を用いて新住民価値を導く。主成分分析はある多変量データの情報を最大限

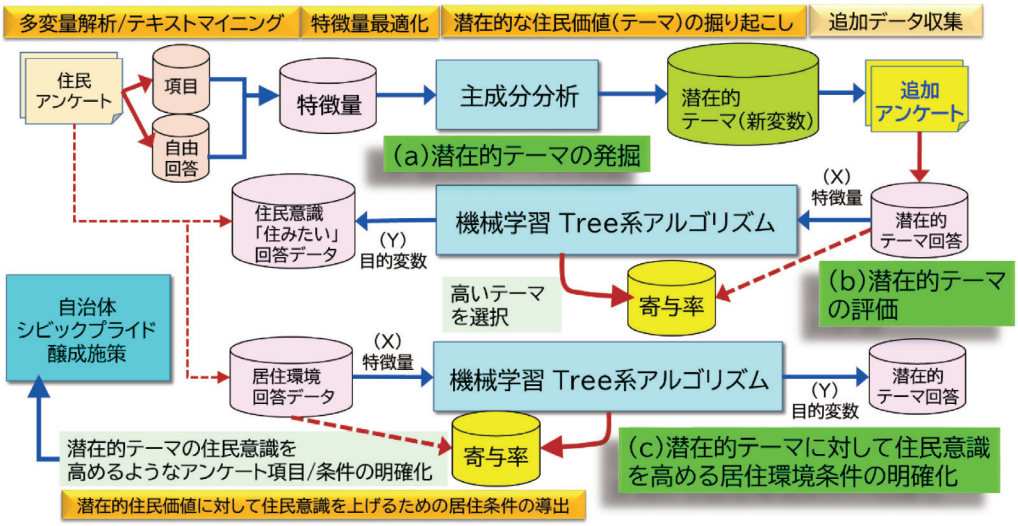


図 1. 全体概念図

保持するように、軸いわゆるベクトルをうまく選ぶ手法である。主成分分析によって導出される軸は主成分といい、主成分は射影されたデータの分散を最大化するように選ぶ。

(4) 新住民価値はシビックプライド醸成に寄与するベクトルであることが望ましい。従って、新住民価値を主成分分析の手法を用いて導出する際、入力する説明変数（特徴量）は、「住み心地」回答に影響度・重要度が高いアンケート項目および自由記述回答を選択する。以下のプロセスにより主成分分析を行うための特徴量を選択する。

(4-1)「住み心地」回答に影響度が高いアンケート項目を選択するために、数量化Ⅱ類を適用しカテゴリウェイトを評価する。また特徴量の相関分析を行い多重共線性を排除する。

(4-2)「住み心地」回答に影響度が高い自由回答を選択するために、テキストマイニングを適用し、TF-IDF 値を指標とする

(4-3)「住み心地」回答を目的変数、(4-1) 及び (4-2) で選択した特徴量をマージしたものを説明変数として、機械学習モデルを適用し説明変数の重要度を評価する。重要度が高い説明変数を選択し、主成分分析に入力する説明変数とする。

上記の方針に従い以下のプロセスで研究及び具体的検証を進める。

2.3 研究プロセスの概要

今回対象とする（a）潜在的テーマの発掘（b）潜在的テーマの評価準備のプロセス概要を示す。

(a) 潜在的テーマの発掘

住民アンケート回答から、居住意識向上に繋がる潜在的な新住民価値を導出する。導出する方法としては、アンケート結果から主成分分析を行い主成分の意味付けを行う。(a-2 新住民価値の概念の掘り起こしと検証)

また、主成分分析における主成分導出の精度を高めるため、事前に「住み心地」に関するアンケート項目回答に重要度が高い特徴量を選択しておく。選択する方法としては、

tree 系の機械学習モデルの説明変数の重要度を測る手法を適用する。つまり、主成分分析を行う前処理として、シビックプライド醸成に重要度が低い特徴量を排除する。手法としては、「住み心地（住み心地良い＝ 1 , 住み心地悪い =0）」を目的変数、「『住み心地』のアンケート項目回答に影響度が大きいアンケート項目及び自由回答内容」を説明変数とした機械学習モデルを適用して、重要度が高い説明変数を選択する。（a-1 新住民価値導出のための特徴量の選択）

機械学習の前処理として、「住み心地」のアンケート項目回答に影響度が高い「アンケート項目及び自由回答内容」を事前に選択しておく。選択手法として、「住み心地」回答に影響度が高いアンケート項目回答に関しては、数量化Ⅱ類を適用し特徴量のカテゴリウェイトのレンジ幅を評価する。また、「住み心地」回答に影響度が高いアンケート自由回答に関しては、テキストマイニングを適用し自由回答の各々に対する特徴語を選択する。選択基準として各回答の TF-IDF 値を指標とする（a-1　新住民価値導出のための特徴量の選択）

（a-1）新住民価値導出のための特徴量の選択

表 1 に「住み心地が良い」「住み続けたい」アンケート項目回答に影響度が高いアンケート項目（特徴量）を選択するプロセスの概要を示す。

表 1.「住みたい」回答に貢献度が高い特徴量を選択するプロセス

目的 / プロセス	手法	手法の概要
アンケート項目回答の解析（特徴量選択）	数量化Ⅱ類	「住み心地」の群を最も分離するアンケート項目（特徴量）の重み付け、特徴量の重み（「住み心地」に対する寄与率）評価、寄与率ランキングでアンケート回答項目を選択（カテゴリウェイト / レンジが 0.15 以上）
（多重共線性排除）	相関分析	数量化Ⅱ類で選択した特徴量に対してクロス集計を行い、変数間のクラメール連関係数を評価
アンケート自由記述回答の解析（特徴語選択）	テキストマイニング	形態素解析 [KH Coder] で、特徴語の TF-IDF を評価（コーパスごとに）、特徴語ごとの TF-IDF 値ランキング 50 を選択。選択した特徴語を数量化（One-hot encoding）
主成分分析のための特徴量選択	機械学習（教師あり分類）	数量化Ⅱ類で選択した特徴量とテキストマイニングで選択した特徴語をマージ。機械学習入力特徴量を生成。 ［1 次選択］ 精度が高いアルゴリズムを評価選択（PyCalet）。 「住み心地」回答変数を目的変数、上記を説明変数とし機械学習モデルで学習 / 予測 / 精度評価。 特徴量の重要度を評価し 50 個の特徴量を選択 ［2 次選択］

本プロセスで特徴量を選択し、次プロセスの主成分分析の説明変数として入力する。

(a-2) 新住民価値の概念の掘り起こしと検証

表 2 に新住民価値の概念（変数）を掘り起こし、検証の準備をするプロセスを示す。

表 2. 新住民価値の概念の掘り起こしのプロセス

目的 / プロセス	手法	手法の概要
新価値概念に相当する主成分の導出と意味付け	主成分分析	「住み心地」アンケート項目回答に対して重要度が高い特徴量から新住民価値の概念に相当する主成分を導出 機械学習の重要度評価で選択した特徴量を標準化。 sklearn の pca.fit_transform() で主成分算出。 主成分の意味付けと累積寄与率を評価。
新住民価値の概念の検証	住民アンケート項目追加	主成分分析で導出した新概念が住民意識向上に繋がるかどうかの検証。導出した主成分の意味付けから、その内容を検証する設問を 2022 年度定例住民アンケートに追加する。

選択した特徴量から主成分を導出して意味づけを行う。意味付けした主成分に関して、住民アンケートに追加設問を行い、アンケート回答から主成分の意味付けの妥当性を検証する。

(a) 潜在的テーマの発掘を行うプロセスの全体像を図 2 に示す。

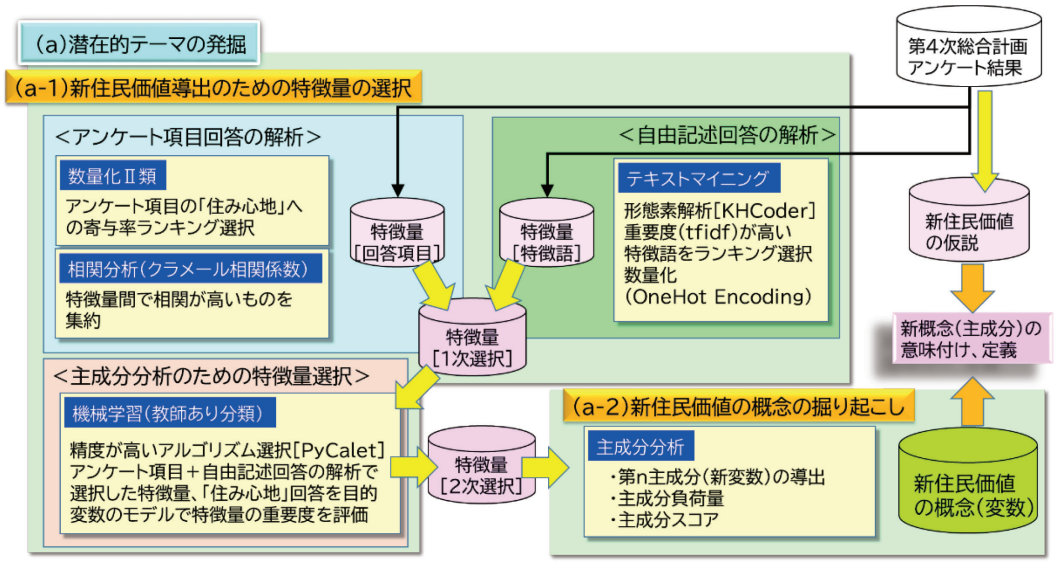


図 2 潜在テーマの発掘を行うプロセスの全体像

(b) 潜在的テーマの評価準備

主成分分析で導出した主成分の意味付けの検証を行う。検証方法は、自治体の 2022 年度の住民アンケートに対して、主成分の意味付けに対応する設問を追加する。本追加アンケート結果を分析して、導出した主成分の意味付けが正しいかどうかを検証する。また、

追加したアンケート項目と「住み心地」との相関を測り（相関分析、機械学習）追加アンケート項目が住民意識の向上に寄与していることを検証する。

３．新住民価値導出のための特徴量の選択

3.1　アンケート項目回答の解析（機械学習モデルの学習に向けた前処理）

新住民価値に相当する主成分を導出する主成分分析を行うために、機械学習のモデルを活用して、「住み心地」のアンケート項目（目的変数）に対して重要度が高い特徴量を選択するが、この機械学習の精度を高めるための前処理として「住み心地」アンケート項目の回答に寄与率が大きい特徴量を選択する。アンケート項目の中でシビックプライドの醸成、町への愛着度を高める上で必要条件となるのが住み心地であると考えた。住み心地が良ければ町への愛着度や住み続けたい意識が向上し、シビックプライドが醸成されることが期待できる。

本プロセスにおける特徴量選択の手法として数量化Ⅱ類および相関分析を適用する。

3.1.1　数量化Ⅱ類での解析

（1）解析プロセスの内容

まず、質的データの判別分析である数量化Ⅱ類を用いて「住み心地が良い」「住み心地が悪い」回答の相関比を最大にするカテゴリウェイトを算出し、「住み心地」回答への寄与率が高い特徴量を選択する。次に、全ての説明変数の組み合わせでクロス集計表を作成し、クラメール連関係数を算出し、クラメール連関係数による相関分析により相関係数が高い特徴量（説明変数）を集約する。

数量化Ⅱ類は、ダミー変数を導入して質的データを数値化することにより、判別分析を行う手法である。判別分析と同様に、説明変数間の関係を加味しながら目的変数を予測するが、予測に対して重要な影響を及ぼす説明変数を明確化する。判別する内容に応じた群データで与えられる目的変数（今回は「住み心地」の良し悪し）と質的データで与えられる説明変数（アンケート項目）との関係をモデル式で表し、モデル式の説明変数の係数の重みにより説明変数の重要度を評価する。

モデル式：$\mathrm{y} = \sum_{\mathrm{j}=1}^{\mathrm{Q}} \sum_{k=1}^{cj} a_{jk} x_{jk} + \varepsilon$

y：目的変数
Q：説明変数の数
cj：j 番目の説明変数のカテゴリ数
a_{jk}：説明変数 j の k 番目カテゴリのモデル式の係数（カテゴリーウェイト）
x_{jk}：j 番目説明変数 k 番目カテゴリのダミー変数
ε：誤差

各説明変数（アンケート項目）のカテゴリ（選択肢回答）に対応するダミー変数値を（x_{jk}）とした場合、モデル式の係数（a_{jk}）であるカテゴリウェイトの値により各アンケート項目の「住み心地」（y）の良し悪しに対する重要度を測る。

従って、説明変数（アンケート項目）の重要度を測るためにはカテゴリウェイト（a_{jk}）

の値を評価することが必要となる。正確には、ある説明変数 (アンケート項目) のカテゴリウェイトのうち最大値と最小値の差の絶対値の大きさ（レンジ）が、その説明変数の重要度に相当する。

以下に示す手法で、各アンケート項目のカテゴリウェイトのレンジを算出し重要度を評価する。

（[5] 参照。算出式は省略）

ここで、各個体（各アンケート回答者）の合成変量 ŷ をサンプルスコアと呼び、モデル式を使って個体ごとに算出可能である。群の特性を表す理論値として使用される。

(a) 群分けされた資料において各カテゴリの関係を明確にするために、群ができるだけ離れるようにカテゴリウェイト a_{jk} を設定する。a_{jk} はサンプルスコア ŷ と群との関係を示す相関比 η^2（2 群の離れ具合を示す指標）が最大になるように定められるため、相関比 η^2 を最大にするカテゴリウェイ a は η^2 を a で偏微分した微分方程式の解として求められる。

(b) s_b^2（群間変動）および $s_{\hat{y}}^2$（全体変動）は、各々ダミー変数データの全体変動行列 T とカテゴリウェイトの行ベクトル a' および列ベクトル a の積、群間変動行列 B とカテゴリウェイトの行ベクトル a' および列ベクトル a の積で求められる。

(c)（b）を（a）の偏微分方程式に代入して、カテゴリウェイト a を固有ベクトル、相関比 η^2 を固有値とした行列 $T^{-1}B$ の以下の固有方程式を得る。

$$\text{固有方程式:} T^{-1}Ba = \lambda a$$

(d) アンケート回答のダミー変数データ x_{jk} から行列 $T^{-1}B$ を求めて（c）の固有方程式に代入。固有値 λ（= 相関比 η^2）と固有ベクトル（= カテゴリウェイト a）を求める。

事前処理として以下を行う。

・相関比 η^2 は偏差（ダミー変数データと全体平均および群別平均との偏差）から構成されており、サンプルスコアの値は「差」の相対的な意味を表せばよい。従って今回は、各説明　変数の末尾のカテゴリウェイトを 0 に固定する。つまりモデル式から末尾カテゴリを除外する。

・相関比最大の条件からは、サンプルスコアの各項の値であるカテゴリウェイトの比しか求まらない（カテゴリウェイト値の大きさに任意性がある）ため、サンプルスコアの値を標準化する（サンプルスコア ŷ の分散を 1 とする）

(e)（d）の固有値および固有ベクトルを算出する手法としてプログラミング言語である Python の numpy ライブラリの linalg.eig 関数を使用する。また、求めた固有値および固有ベクトルよりカテゴリウェイト a を決定し、レンジを算出し見える化を行う。使用する機能として numpy ライブラリと併せて matplotlib ライブラリを使用する。

(f)（e）で求めたカテゴリウェイト a のレンジ値により、説明変数（アンケート項目）

の重要度を評価する

(2) 対象となるアンケート回答データと解析結果

分析対象のアンケート回答のデータ情報は有効回答者数(レコード数)1738 人、アンケート項目（説明変数）62 項目、項目選択数（カテゴリ数）424 個、である。アンケートのデータ情報を以下表 3 に示す。表 3 には、以降記載している各分析結果で説明変数を選択した結果も載せている。ここで、アンケート項目を説明変数、アンケート回答項目の選択項目をカテゴリと定義する。

表 3 分析対象のアンケート（選択式アンケート）回答結果の情報（分析前後）

	レコード数 （有効回答者数）	アンケート項目 （説明変数）	項目選択数 （カテゴリ数）
数量化Ⅱ類分析前	1738	62	424
数量化Ⅱ類分析後	1738	50	378
相関分析後	1738	41	313

本アンケート回答で、数量化 2 類で対象とする群は「住み心地が良い群」と「住み心地が悪い群」の 2 群である。下記図 3 に示すアンケート項目回答のうち、選択番号 1,2 を「住み心地が良い群」、選択番号 3,4,5,6 を「住み心地が悪い群」として群分けを行った。

Q1 あなたは、xxx町の住み心地についてどう思いますか?

1.とても住みよい 2.まぁまぁ住みよい 3.どちらとも言えない 4.やや住みにくい 5.とても住みにくい 6.未回答

図 3 自治体が実施した選択式アンケートの Q1(住み心地)

上記群分けに応じて、目的変数を「住み心地の良し悪し」(Q1 の回答結果）とした。このとき、

「住み心地」に寄与するアンケート項目（特徴量）を選択するために、目的変数に対する説明変数の重要度である説明変数のレンジを評価した。数量化Ⅱ類におけるレンジとは、2 群の相関比を最大にするようにカテゴリウェイト (説明変数の係数いわゆる重み) を算出し、アンケート項目のカテゴリウェイトの最大値から最小値を引いた値である。レンジの閾値を 0.15 として、値が閾値以下の説明変数を削除した（アンケート項目 (説明変数) は 12 個、項目選択数 (カテゴリ数) は 46 個削除)。その結果、アンケート回答のデータ情報は、アンケート項目 (説明変数) は 50 個、項目選択数 (カテゴリ数) は 378 個となった。(表 3 参照)

算出した各アンケート項目 (特徴量) のレンジに関して、python のグラフ描写ライブラリである matplotlib でその値の大きさが上位 21 を図 4 のグラフに示す。

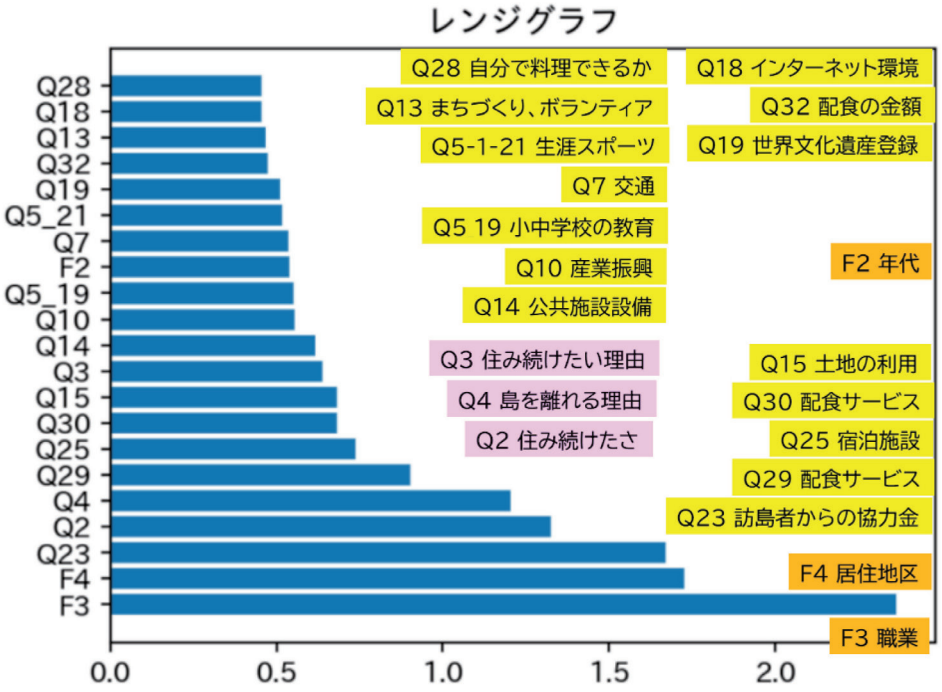


図 4 カテゴリウェイトのレンジ値ランキング上位 21

本プロセスの解析結果および評価内容を以下に纏める。

図 4 の通り、レンジランキング 1 位は職業（F3）、2 位は居住地区（F4）であり、住み心地を左右する上で職業、居住地区が重要であることがわかる。職業に関して「農業」のカテゴリウェイトの値が絶対値（マイナス）で最大で「住み心地」に影響している。また、居住地区によって「全員が住み心地が良い」「10% が住み心地が悪い」など意識が異なる。3 位は訪島者からの協力金（Q23）である。このポイントでアンケート回答の自由記述を眺めたところ、世界文化遺産に認定された環境を島としてどう利用するかが、住民の関心事の 1 部であることが認識できた。

特徴量選択の精度は最終的に主成分分析の累積寄与率にて評価できる。この寄与率に影響する要因として本プロセスでの特徴量選択結果があると考える。この特徴量選択結果の精度向上が課題である。課題のポイントを以下に示す。

(a) カテゴリウェイトが目的変数の「住み心地良い」に効いているカテゴリ検証

レンジにより目的変数に影響が高いアンケート項目は把握できたが、住み心地が良い／悪いのどちらに影響があるかは各々のカテゴリと目的変数のクロス分析が必要である。

(b) レンジの閾値のチューニングの試行錯誤が必要である。現状は正規分布の 3 σ外の比率を参考にして、これよりより低めに設定している。

3.1.2　相関分析（クラメール連関係数）

上記数量化Ⅱ類では目的変数の「住み心地」に対して重要度が高いアンケート項目（特

徴量）を選択した。さらに特徴量の説明力を上げて、後述の機械学習モデルにおける予測精度を高めるため多重共線性の排除を目的として、相関分析を活用して、相関の高い説明変数を集約した。相関分析の方法は質的変数間の相関の指標であるクラメール連関係数を利用した。クラメール連関係数の一般的、経験的な基準として非常に強い関連性があるとされる 0.5 以上の説明変数を削除した。表 4 にクラメール連関係数の強弱基準を示す。

表 4 クラメール連関係数の強弱基準

0.50 以上	非常に強い関連性がある
0.25 以上　0.50 未満	関連性がある
0.10 以上 0.25 未満	弱い関連性がある
0.10 未満	関連性がない

クラメール連関係数が 0.5 以上の 64 個の 2 つのカテゴリの組み合わせから、組み合わせの片方を無作為に削除する要領で、9 個の説明変数を削除した。この削除により相関分析後は、説明変数が 41 個、カテゴリ数が 313 個になった。（表 3 参照）

相関係数が高かった主な説明変数のペアは「生涯学習」と「生涯スポーツ」（0.806）、「文化・コミュニティ活動」と「生涯スポーツ」（0.680）、「歴史・文化や自然景観など、町の資源活用」と「緑地や海岸など、自然景観の保全」（0.656）、「文化財や町並みの保全」と「文化コミュニティ活動」（0.656）などであった。どちらを削除するかの基準としては、他の説明変数との相関を基準としたが、目的変数との相関を条件とした説明変数の削除基準の確立が今後の課題である。

因みに「生涯学習」と「生涯スポーツ」に関して目的変数との関係を見ると、二つとも目的変数「住み心地」との間のクラメール連関係数は 0.25 未満で、住み心地との関連性は少ない。娯楽施設などを作るときは、学習とスポーツ両方取り組む人が多く、学習とスポーツ両方を兼ね備えた施設が望ましいと思われる。

3.2　アンケート自由記述回答の解析（テキストマイニング）

主成分分析における「新住民価値」掘起しのためのアンケート文字回答の特徴語選択を行う。選択した特徴語を特徴量として選択・数量化し、3.1 で選択した重要度が高いアンケート項目からの特徴量とマージした両方の特徴量を選択する。アンケートから特徴量を抽出する上で、アンケート実施側が用意した尺度からだけでなく、アンケート自由記述回答のような回答者の視点による意見から特徴量を抽出することで、より回答者全体の意向を汲み取れると考える。そこで、主成分分析にかけるシビックプライド醸成に寄与する特徴量として、アンケート項目選択回答データと自由記述回答のデータの両方を用いる。その後「住み心地」を目的変数とした機械学習モデルに対して上記の両方の特徴量を入力し、機械学習を実行する。実行結果、最終的に主成分分析に入力する特徴量を最終選択する流れとなる。

以下に、テキストマイニングを適用したアンケート自由記述回答データからの特徴量抽

出の概要を示す。

アンケート自由記述回答データから抽出する特徴量は、回答者が記述した文書の単語で特徴的な語 (特徴語) とする。特徴語の抽出方法としては、回答者一人あたり 1 文書とする自由記述回答データを形態素解析し単語数を集計し、形態素解析・単語数集計データから TF-IDF 値を算出した後、TF-IDF の高いランキングで特徴語を選択して特徴量として数値化する。

記述データから特徴語を抜き出すには TF-IDF という指標を利用する。

TF-IDF（Term Frequency - Inverse Document Frequency）とは文書群について、単語がどのくらい特徴的かを表す指標であり、TF と IDF の積である。TF すなわち単語の出現頻度はそれぞれの文書中にその単語が出てくる頻度であり、IDF すなわち逆文書頻度は全体の文書のうち，その単語を含む文書の程度の逆数であり、いわばレア度である。TF-IDF の計算対象単語を t とすると以下の式になる。

$$\mathrm{TF} = \frac{\text{単語 t の出現回数}}{\text{文書内の総単語数}}$$

$$IDF = -\log\frac{\text{単語 t を含む文書の数}}{\text{総文書数}}$$

$$= \log\frac{\text{総文書数}}{\text{単語 t を含む文書の数}}$$

TF-IDF を計算し特徴語を抽出する前に回答記述データを形態素解析し、形態素を抽出する必要がある。形態素解析には KH coder という計量テキスト分析またはテキストマイニングのためのフリーソフトウェアを用いた。KH coder に全回答者の記述データを入力し、形態素解析した後、KH coder の機能で単語ごとに単語出現数を示した表を出力した。上記の TF-IDF の式に従い以下、KH coder で出力した表に対して図 5 のように TF-IDF を計算した。

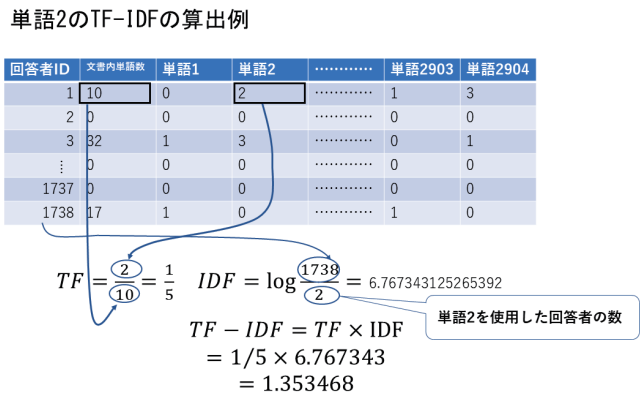


図 5 TF-IDF の計算例

回答者ごとの 1 文書に対して、形態素解析で抽出した単語ごとに TF-IDF の値を求める。例えば図 5 に示す回答者 ID が 1 の回答者の回答中に出現する「単語 2」に関して、TF は回答者 1 の回答文書内の「総単語数 (10)」に対する「単語 2 の出現回数 (2)」の割合となり、IDF は「総文書数 (1738)」を「単語 2 の出現回数 (2)」で割った値の対数値となる。

上記で算出した回答者ごと単語ごとの TF-IDF の値に関して、合計値と分散値を求め合計値のトップ 50、分散値のトップ 50 の単語を重複排除して 76 個の単語（特徴語）を選択した。選択語の一覧を図 6、選択方法を図 7 に示す。

ない.1	する	高血圧	回答	ない	円	タクシー	思う	糖尿	ある	塩	わかる	小児科	カロリー	減	なる	ほしい.1	
町	高齢	できる	いる	行く	子供	小値賀	人	増加	もっと	子ども	多い	前	人口	参加	制限	観光	
出来る	施設	町民	分かる	医療	行う	賃金	人材	進学	計画	減	保全	ナイ	流れ	中途半端			
しかた	DM	任せる	なんとも	環境	特に	身体	食事	IT	悪化	すべて	転勤	企業	誘致	葬儀	予備	納税	
ふるさと	船	用事	まかなう	クラウドファンディング				住宅	血糖	整備	訪れる	抜	余裕				

図 6 選択語一覧

回答者ID	文書内単語数	単語1	単語2	……	単語2903	単語2904
1	10	0	1.3534	……	0.1317	0.0439
2	0	0	0	……	0	0
3	32	0.8321	0.0201	……	0	1.1203
⋮	0	0	0	……	0	0
1737	0	0	0	……	0	0
1738	17	0.6274	0	……	0.2452	0

列ごとに合計、分散を算出し、降順に並べ替えする。

合計と分散の大きさ上位50語を特徴語とする。

和top50 and 分散top50 : 24個
和top50 or 分散top50 : 26 × 2 = 52 , 52 + 24 = 76個

特徴語 : 76単語

図 7 特徴語選択プロセス

数量化の手法としてダミー変数を適用する。つまり選択した 76 個の特徴語を説明変数（特徴語）として回答者ごとに、各特徴語のカラムに対して回答者の文書に各特徴語が出現する場合「1」、出現しない場合「0」を記録する。

選択した 76 個の特徴語の内容を見てみると、健康 / 医療に関する懸念や不安に関する語（回答）が多くみられた（「高血圧」「小児科」「糖尿」「カロリー」など）。また、交通や生活インフラに関する語（回答）も多かった（「タクシー」「施設」「環境」など）。生活するための交通手段や環境に関する意見や不満が散見される。特にタクシーが存在しないことへの言及回答が多かった。

特徴語選択の精度は最終的に主成分分析の累積寄与率にて評価できる。この寄与率に影響する要素として、特徴語の選択方法（基準、数、自治体を特徴付けない語の排除など）、数量化の方法などが想定できる。特に、今回の検証において主成分分析に入力する特徴量として、テキストマイニングで抽出した特徴語がほとんど選択されなかった、ダミー変数を活用したことで、特徴語の重みが低くなったことが考えられる。数量化の方法として、ダミー変数の代わりに TF-IDF 値を使うなどの方法の検討余地がある。

４．機械学習モデルの適用と精度評価（主成分分析のための特徴量選択）

主成分分析で導出する主成分に関して、「住み心地」を良くするための潜在的なテーマを導き出す目的であることから、主成分分析に入力する特徴量はアンケート項目「住み心地」の回答に対して強い影響度がある（重要度が大きい）変数になっていることが必要であると考える。重要度が大きい特徴量を選択するアプローチとして 3.1 および 3.2 で選択した特徴量を入力、目的変数を「住み心地の良し悪し」（Q1 の回答結果）として、「住み心地」を予測する機械学習モデルを構築・実行・精度評価を行う。本モデルの精度を高めた後、決定木のアンサンブル手法で用いられる feature importance 機能（scikit-learn ライブラリ）などを活用して、特徴量の重要度を評価し、主成分分析に入力する特徴量を選択［2 次選択］する。以下に実施したプロセスの概要を示す。

（a）学習および検証データの作成

3.1 および 3.2 で作成した、アンケート項目選択データと自由記述回答データを一つのデータに纏めて機械学習モデルに入力する特徴量［1 次選択］とした。特徴量は 312 変数でありダミー変数にて数量化した。特徴量の例を図 8 に示す。目的変数はアンケート Q1 の回答結果から、「住み心地が良い」：1、「住み心地が悪い」：0 の 2 値の数値データとしている。つまり 2 値分類の予測モデルを構築した。学習・検証データは、ホールドアウト法を適用して生成した。

	Q1	F2_15歳～19歳	F2_20歳代～30歳代	F2_40歳代～50歳代	F2_60歳代	F2_70歳代	F2_80歳代	F2_90歳以上	F2_無記載	F4_中村御	…	ふるさと	船	用事	まかなう	クラウドファンディング	自宅	血糖	抜	余裕	減
0	1	0	0	0	0	0	1	0	0	0	…	0	0	0	0	0	0	0	0	0	0
1	1	0	0	1	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	1	0	0	0	…	0	0	0	0	0	0	0	0	0	0
3	1	0	0	1	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0
4	1	0	1	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
1733	0	0	0	0	0	1	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0
1734	0	0	0	1	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0
1735	0	1	0	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0
1736	0	0	0	0	0	0	0	0	1	0	…	0	0	0	0	0	0	0	0	0	0
1737	0	0	1	0	0	0	0	0	0	0	…	0	0	0	0	0	0	0	0	0	0

1738 rows × 389 columns

図 8 入力した特徴量の例

（b）モデルの選択と精度評価

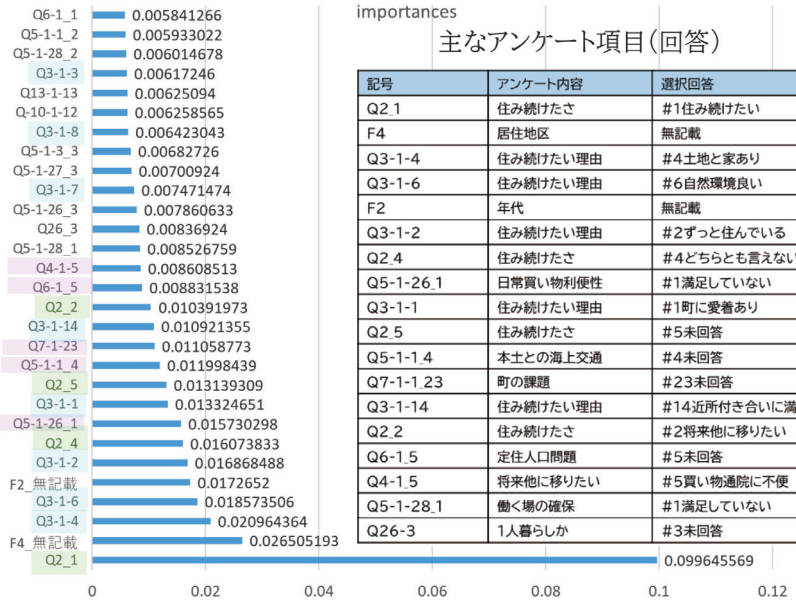
「住み心地」回答の予測を行うモデルを構築し、学習・予測・精度評価およびモデル精度の向上を行った。PyCaret という自動機械学習ライブラリを用いて、自動で機械学習モデルの精度評価を行った。PyCaret は、python のオープンソースのローコード機械学習ライブラリであり、仮説から考察までのサイクルタイムを短縮することを目的とする。また、主要な学習モデルの精度を比較、精度が高い順に並べてくれる（ランキングは pycaret が自動で出力）。さらに、accuracy や f 値、適合率などの指標の % が大きい順に自動でソートしてくれる。今回説明変数から「Q1(住み心地)」を分類する上で、PyCaret が自動選択した最も精度が高いモデルはランダムフォレストであり、accuracy（正解率)、適合率、再現率、f 値、AUC 共に 80% を超える精度を確認できた。図 9 に PyCaret が出力したモデルのランキング例を示す。

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
rf	Random Forest Classifier	0.8018	0.8628	0.9239	0.8056	0.8602	0.5247	0.5419	0.557
ada	Ada Boost Classifier	0.7961	0.8421	0.8814	0.8227	0.8502	0.5311	0.5377	0.224
et	Extra Trees Classifier	0.7952	0.8578	0.9027	0.8098	0.8532	0.5173	0.5281	0.562
lightgbm	Light Gradient Boosting Machine	0.7944	0.8544	0.8852	0.8184	0.8501	0.5239	0.5299	0.215
lr	Logistic Regression	0.7903	0.8290	0.8665	0.8247	0.8446	0.5220	0.5254	0.465
gbc	Gradient Boosting Classifier	0.7895	0.8451	0.8815	0.8147	0.8465	0.5128	0.5182	0.632
ridge	Ridge Classifier	0.7771	0.0000	0.8540	0.8164	0.8341	0.4938	0.4972	0.033
lda	Linear Discriminant Analysis	0.7688	0.8067	0.8465	0.8116	0.8277	0.4754	0.4794	0.143
svm	SVM - Linear Kernel	0.7624	0.0000	0.8364	0.8198	0.8206	0.4587	0.4782	0.064
knn	K Neighbors Classifier	0.7393	0.7669	0.8404	0.7822	0.8097	0.3972	0.4012	0.240
dt	Decision Tree Classifier	0.7196	0.6930	0.7769	0.7942	0.7850	0.3819	0.3831	0.048
dummy	Dummy Classifier	0.6595	0.5000	1.0000	0.6595	0.7948	0.0000	0.0000	0.022
nb	Naive Bayes	0.5535	0.7574	0.3916	0.8530	0.5346	0.2069	0.2678	0.030
qda	Quadratic Discriminant Analysis	0.3454	0.4914	0.0337	0.5848	0.0630	-0.0119	-0.0386	0.131

図 9 PyCaret が出力したモデルのランキング例

（c）特徴量の重要度評価と選択

ランダムフォレストの feature_importance 機能を使い、モデル学習・予測・精度評価後に（sklearn.ensemble.RandomForestClassifier における feature_importances_）各特徴量の重要度を評価した。feature_importance は、決定木のアンサンブル手法において特徴量の重要度が数値で表され、特徴量選択などに用いられる。「どの特徴量がどれくらい重要か」を実数で表したものである。本機能を使用して出力した特徴量の重要度が 0.004 以上の 49 個を主成分分析へ入力する特徴量として選択した［2 次選択］。主な重要度ランキングを図 10（数値 feature_importance の値）に示す。



主なアンケート項目（回答）

記号	アンケート内容	選択回答
Q2_1	住み続けたさ	#1住み続けたい
F4	居住地区	無記載
Q3-1-4	住み続けたい理由	#4土地と家あり
Q3-1-6	住み続けたい理由	#6自然環境良い
F2	年代	無記載
Q3-1-2	住み続けたい理由	#2ずっと住んでいる
Q2_4	住み続けたさ	#4どちらとも言えない
Q5-1-26_1	日常買い物利便性	#1満足していない
Q3-1-1	住み続けたい理由	#1町に愛着あり
Q2_5	住み続けたさ	#5未回答
Q5-1-1_4	本土との海上交通	#4未回答
Q7-1-1_23	町の課題	#23未回答
Q3-1-14	住み続けたい理由	#14近所付き合いに満足
Q2_2	住み続けたさ	#2将来他に移りたい
Q6-1_5	定住人口問題	#5未回答
Q4-1_5	将来他に移りたい	#5買い物通院に不便
Q5-1-28_1	働く場の確保	#1満足していない
Q26-3	1人暮らしか	#3未回答

図 10 主な重要度ランキング（feature_importance）

機械学習のモデルに入力した特徴量は 389 個(数量化Ⅱ類＋相関分析で選択した特徴量:313 個、テキストマイニングで選択した特徴量：76 個）であり、機械学習を実行し上記の重要度を評価した結果、主成分分析に入力する特徴量として 49 個のカテゴリ（アンケート項目回答）が含まれる説明変数（アンケート項目）を選択した。

「Q1 住み心地」の予測に対して、重要度が高いアンケート項目回答は、居住地区、年代の他に、「Q2 住み続けたさ」「Q3 住み続けたい理由」が上位を占めるのは想定通りであるが、その他「Q5-1-26_1」（日常の買い物が不便）、「Q5-1-28_1」（働く場の確保が不満）、「Q4-1_5」（買い物通院に不便）が上位にランキングされており、生活基盤（交通、経済的基盤）の充実が住民の住み心地に大きく影響していることが判る。また、アンケート回答無記載、未回答の項目の重要度が高い（F2 年代、F4 居住地区など）結果に関して、回答者の年代と居住地区を軸にした「住み心地」の住民意識を深掘りする必要がある。また、各特徴量の予測値への影響を測る PDP(Partial Dependent Plot：Scikit-learn のライブラリ) を適用した影響度の深掘りも試す余地があると考える。

５．新住民価値の掘り起こし（主成分分析）

(1) 主成分の算出

Python のライブラリである scikit-learn のクラス sklearn.decomposition.PCA() にて主成分分析を行った。第 1 ～ 10 主成分を算出した。

前述のように、主成分分析は軸にデータの分散を射影し、分散を最大化するように軸（主成分）を決定する。この分散の最大化問題は、式を変形していき、ラグランジュ未定乗数法などを経て、分散共分散行列の固有値問題に帰着する。固有ベクトルが主成分である。

58 種類の特徴量（機械学習の結果から一部年代に関する特徴量を追加)、1738 次元の標準化されたアンケートデータに対して、主成分分析を行った。第二主成分までの累積寄与度は 0.11560645 である。図 11 に算出した第 10 主成分までの累積寄与率を示す。

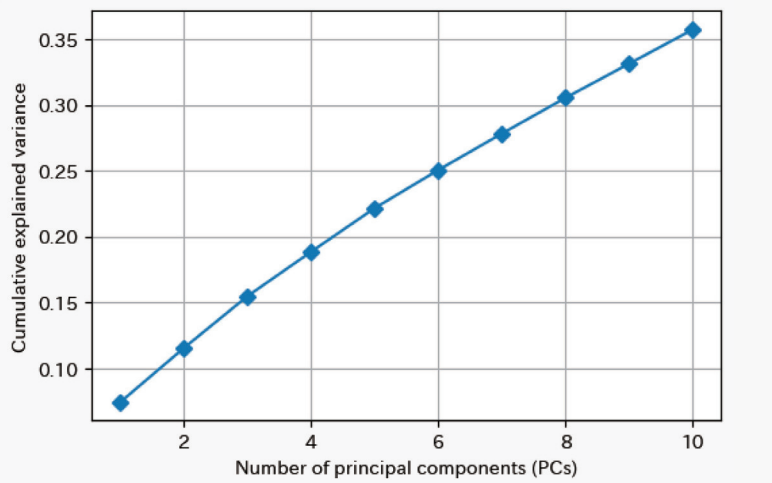


図 11 算出した第 1 ～ 10 主成分の累積寄与率

（2）算出主成分の意味付け

今回第 1、第 2 主成分の意味付けを行いその内容に応じて、本年度の自治体アンケートにアンケート項目を追加する。追加アンケート回答結果を確認した上で、今回の主成分分析により導出した潜在的な変数の意味付けの検証を実施する。

主成分の評価、意味づけの方法として、主成分負荷量と主成分得点に着目する。主成分負荷量は正負、絶対値の大きさによって主成分得点への影響度が異なる。上記分散共分散行列の固有ベクトルの値に相当する主成分負荷量を第 1 主成分 (pc1) および第 2 主成分 (pc2) の 2 次元座標にプロットした（図 12)。さらに本グラフに pc1、pc2 の意味付けを併せて示す。意味付けの内容は後述する。主成分得点は図 1 3 のように PC1 と PC2 の 2 次元グラフにプロットし、その点が目的変数のどの値に属しているかを表示させることで、各回答者が主成分の軸上で、どこに位置しているかを把握することができる。

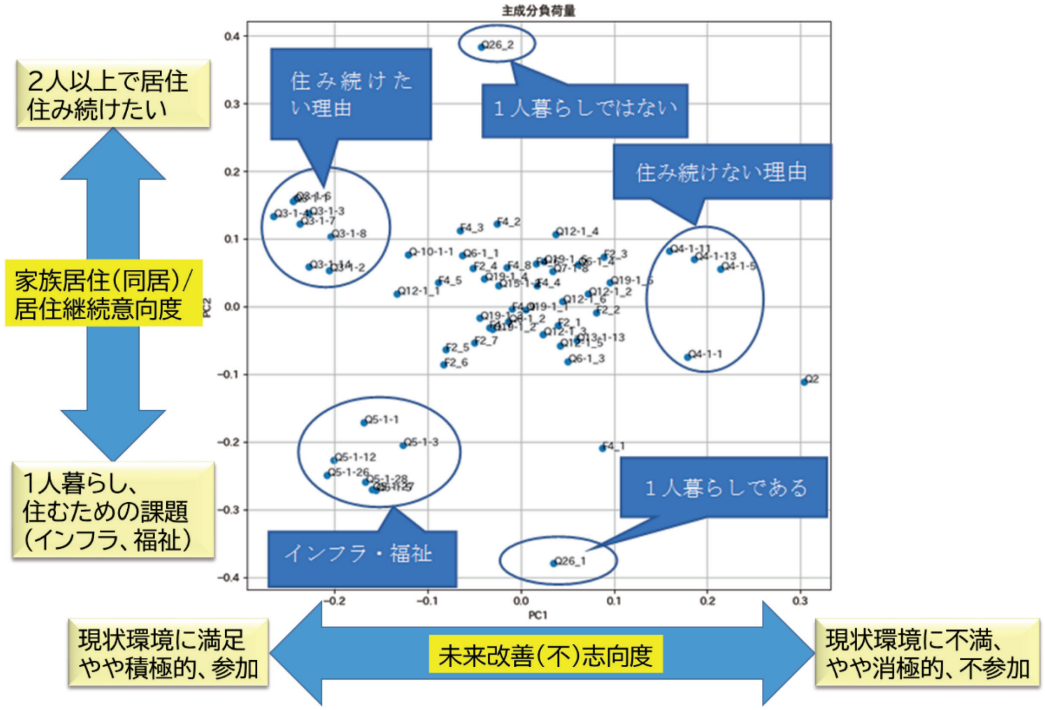


図 12 主成分得点の分布

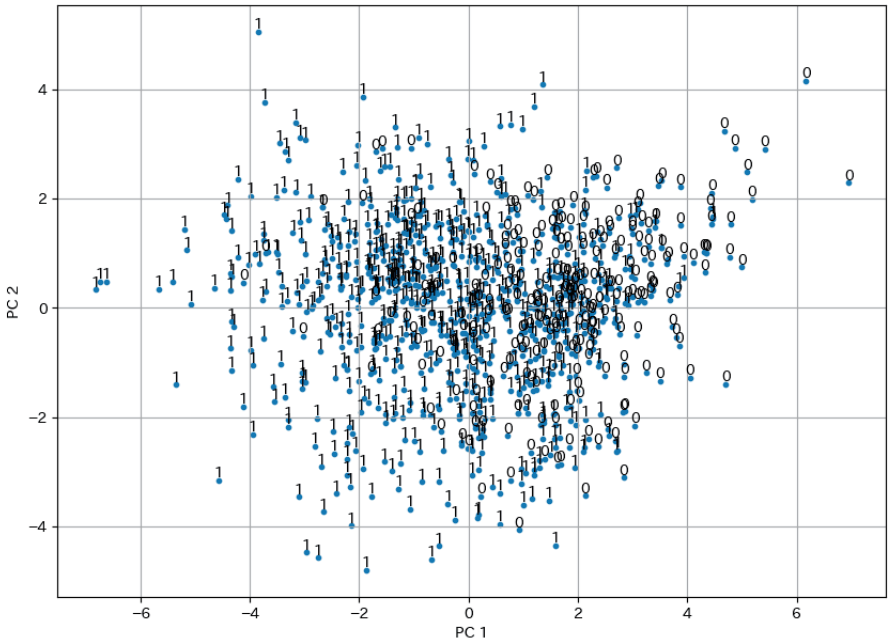


図 13 主成分得点をプロットした 2 次元グラフ

（a）第 1 主成分についての意味付け

①主成分負荷量の傾向

正の方向に住み続けない理由の項目が集まっている。住み続けない理由は「島外への買い物や通院に不便だから」、「昔からの風習・慣習が負担だから」、「福祉・保険・医療などの生活支援サービスに不満があるから」などであるが、やむを得ず小値賀町を離れるような項目は存在しなかった。また、主成分負荷量が 6 番目に大きい（絶対値は相対的に小さい）項目で「世界文化遺産登録の利活用などに期待することはあるか ?」というものに対して、「特に期待するものはない」の選択項目があった。

負の方向には住み続けたい理由の項目とインフラ・福祉に関する項目が集まっている。住み続けたい理由の項目はいずれも、やむを得ず住み続けるものではなく、何かしら住み続けたい因子があることが読み取れる。主成分得点を見ると、負の方向に住み心地が良いと答えた人が集まっていることがわかる。インフラ・福祉・イベント・行事なども負の方向に集まっていることから、「住み続けたい」かつ「住み心地が良い」と答えた人々がこれらに関心があることが分かる。

②意味付けのまとめ

第 1 主成分の意味：　未来改善 (不) 志向度

［主成分負荷量（正）の方向］

- 現状の生活環境に不満（インフラ・福祉などに不満あり、住み心地も悪く住み続けたくない）
- まちづくり参加意欲が小さく、やや消極的（まちづくり（インフラ・福祉）には関心がない）

・人口・就業・文化遺産活用に一部課題提起あり
・年代（10 ～ 30 代、40 ～ 50 代が主）、地域特性あり
［主成分負荷量（負）の方向］
・現状の生活環境に満足（町に愛着もあり、人間関係も良好で、住み心地も良く住み続けたい）
・まちづくり参加意欲が大きく、積極的（まちづくり（インフラ・福祉・コミュニティ・未来の生活環境）に関心があり、改善の期待または満足の意向）
・年代（60 ～ 90 代が主）、地域特性あり

③主成分の意味付けからの考察　（自治体施策の提案ポイント）
若い世代の積極性を提起するイベントなどの企画により、ベテラン世代との世代交代を図るきっかけづくりが重要と思われる。また、生活環境に関しては、医療の充実や交通・生活物資供給インフラの充実が急務であると思われる。

（b）第 2 主成分についての意味付け
①主成分負荷量の傾向
第 2 主成分はインフラ・福祉と一人暮らしである項目が負の方向に位置し、一人暮らしでない項目と住み続けたい項目が正の方向に位置している。

②意味付けのまとめ
第 2 主成分の意味：　家族と一緒に住み続けたい度
［主成分負荷量（正）の方向］
・一人暮らしでない
・住み続けたい意向
・年代（40 ～ 50 代が主）、地域特性あり
［主成分負荷量（負）の方向］
・一人暮らし
・インフラ・福祉が重要

③主成分の意味付けからの考察　（自治体施策の提案ポイント）
家族と一緒に生活する生活・経済的基盤の充実の優先度を上げることが必要と想定できる。また、1 人暮らしの生活に困らない（自治体サービス、医療ネットワーク）インフラや潤いをもたらす環境（趣味・憩い・文化的活動などのコミュニティ）も重要であると考えられる。

（3）分析精度向上のための対策
今回の主成分分析の課題として、第 2 主成分までの累積寄与率が低い（11.6%）ことが挙げられる。主成分の寄与率を向上させるには、特徴量の選択方法を改善する必要性があると考える。以下に、特徴量の選択方式の改善案を纏める。

（a）数量化 II 類：

・カテゴリウェイトの目的変数への効き方を評価した選択（住み心地が良いサンプルスコアへの寄与の度合いが高い説明変数を選択する）
・説明変数を選択するためのレンジ閾値のチューニング
・別モデルを用いた検証との比較評価を行う。例えば回帰分析モデルの回帰係数で評価する。

（b）相関分析：

相関分析では、相関が高い 2 変数を無作為に片方削除する方法ではなく、目的変数とのクラメール連関係数が高い変数を削除する。

（c）テキストマイニング：

・特徴語のデータの数量化において、ダミー変数を使わず TF-IDF の値を使用する。
・特徴語の選択方法の見直し（基準、数、自治体を特徴付けない語の排除）

（d）機械学習：

・検証データの分割法を交差検証法にする
・機械学習の不均衡データにより精度が低下する問題をランダムサンプリングなどの手法で改善する。
・重要度算定方法を変更する（PDP(Partial Dependent Plot) などの活用

６．新住民価値の検証

主成分分析で導出した主成分の意味付けの検証を行うにあたり、追加したアンケートの設問の概要を以下に示す。追加対象のアンケート設問は、自治体における 2022 年度の住民アンケートの一部として反映させて頂いた。本住民アンケート結果により導出主成分の妥当性検証を実施予定である。

（a）主成分の意味付けからの設問

・まちづくりへの参加意欲・自治体のまちづくりイベントに積極的 / 消極的とその理由
・住民自身のまちづくりイベントに関して参加する / 参加しないとその理由
・今後小値賀町が発展（人口増加、生活インフラ充実、産業振興）するためのテーマの問い

（b）主成分分析の時に明確になった寄与率が高いアンケート項目

買い物 / 交通の利便性改善、福祉対策、働く場の確保、まちづくりへの住民参加、
定住人口問題、世界文化遺産の利活用、生活コミュニティ　など

（c）機械学習にて重要度が高いアンケート項目

日常での買い物の利便性、本土との海上交通、定住人口問題、働く場の確保　など

7．おわりに

今回､201 8年度住民アンケート回答結果より､｢住み心地」を良くするための特徴量（アンケート項目回答）を入力し主成分分析を行い潜在的な住民価値に繋がる主成分を導出・意味付けを行い、その検証のためのアンケート設問追加を実施した。

各主成分の寄与率は低い値であり、寄与率向上の対策、対策後の主成分の意味付けと検証がさらに必要であると認識する。但し、アンケート回答からの潜在的な住民意識の見える化のためのプロセスのパターン例を構築することができたと考える。

今後の研究として、上記主成分導出の精度向上の他に、発掘した潜在的な変数の「住み心地」が良いことに対する重要度の評価、そして重要度が高い潜在的な変数が「正」になることに対して重要なアンケート項目を明確にすることを計画する。重要なアンケート項目から住民意識を高める居住環境条件を明確にし、価値がある自治体施策に繋げていくことを目的とする。

参考文献

[1] 読売広告都市生活研究局､｢シビックプライド - 都市のコミュニケーションをデザインする」.
[2] 井口達雄講義 , 数学解析第 1・講義ノート 5 平成 30 年度 5 月 : https://www.math.keio.ac.jp/~iguchi/Lectures/pdf/2019/Note_MA_5.pdf.
[3] 時弘哲治、｢微積分」、東京大学工学教程 .
[4] 菅　民郎 , 藤越 康祝 (著): 質的データの判別分析 数量化 2 類 .
[5] 佐藤浩輔：島根大学人間科学部 2019.07.13, 応用心理学研究Ⅰ , テキストマイニング講義資料 ,https://www.slideshare.net/cos039840935/ss-155407947.
[6] 相澤彰子、｢語と文書の共起に基づく特徴度の数量的表現について」、情報処理学会論文誌 , Vol.41,No.12, pp.3332-3343.
[7] Thomas M.Cover, Joy A.Thomas、｢情報理論 - 基礎と広がり -」.
[8] 涌井義之他、｢初歩からしっかり学ぶ　実習多変量解析入門」、技術評論社 (2011 年 12 月).
[9] 塚本邦尊他、｢東京大学のデータサイエンティスト育成講座」、マイナビ出版 (2019 年 3 月).
[10] 岡崎 直観 : 機械学習帳 , https://chokkan.github.io/mlnote/unsupervised/04pca2.html#equation-eq-pca-lambdas

山本　裕　東京国際工科専門職大学　工科学部　情報工学科　准教授
高田晃希　東京国際工科専門職大学　工科学部　情報工学科　２年生
黒羽　晟　東京国際工科専門職大学　工科学部　情報工学科　２年生
橋本沙也加　株式会社百代 代表取締役
橋本尚子　株式会社百代
岡田ゆかり　株式会社百代