2023 年度　統計データ分析コンペティション

審査員奨励賞　[大学生・一般の部]

市区町村ごとの失業率の要因分析

富張 聡祥（東京大学理学部情報科学科）

市区町村ごとの失業率の要因分析

富張　聡祥[*1]

*1：東京大学理学部情報科学科

１．研究のテーマと目的

失業率は、経済の健全性を示す重要な指標の一つであり、経済の現状や将来の方向性を示す。失業率の正確な分析は、経済の状態の理解や将来の予測に寄与するだけでなく、効果的な政策立案のために必要な情報を与える。

近年、機械学習手法が多様な領域での予測や分析に用いられるようになっており、失業率に関してもこれらの先進的な手法を駆使した研究が増え、高い予測精度が達成されている[2,3]。一方、機械学習手法の多くには分析や予測に至るまでの過程が不透明で、得られた結果の背後にある理由や要因が不明確であるという欠点がある[1]。このようなブラックボックス的な特性を持つ機械学習手法を用いた失業率の分析は、政策立案や実務的な意思決定において重要となる説明性や透明性を欠いてしまう可能性がある。

そこで、本研究では、機械学習手法の中でも説明性が高いとされる手法を用いて市区町村ごとの失業率の分析を行うことを目的とする。失業率の要因や地域ごとの特徴を明らかにすることで、具体的な施策の提案や、より的確な政策の方針の指針を与えることを目指す。

２．研究の方法と手順

2.1 重回帰と Elastic Net 回帰

まず、市区町村ごとの説明変数$(x_{i,1},\dots,x_{i,d})$と被説明変数である失業率y_iのデータに対して線形モデル

$$y_i = \sum_{k=1}^{d} \beta_k x_{i,k} + \beta_0 + \epsilon_i$$

が成り立つことを仮定し、古典的な統計分析手法である重回帰と過学習を防ぐために正則化項を加えた Elastic Net 回帰を行った。損失関数としては平均二乗誤差を用いた。このとき、Elastic Net 回帰の損失関数は正則化の強さを表すパラメータλと L1 正則化と L2 正則化の割合を決めるパラメータ$\alpha \in [0,1]$を用いて

$$L = \sum_{i=1}^{n}\left(\sum_{k=1}^{d} \beta_k x_{i,k} + \beta_0 - y_i\right)^2 + \lambda\alpha \sum_{k=1}^{d} |\beta_k| + \frac{\lambda(1-\alpha)}{2}\sum_{k=1}^{d} \beta_k^2$$

と表される。これらのハイパーパラメータは訓練データを用いた5分割クロスバリデーションによって決定した。そして、回帰式の係数β_kの値によって、各説明変数の失業率への寄与度や重要度を評価した。

2.2 ランダムフォレスト回帰と SHAP 値

次に、各市区町村のデータにランダムフォレスト回帰を適用した。ランダムフォレスト回帰は、決定木ベースのアンサンブル学習アルゴリズムで、特定の関数形を前提とせずに動作し、過学習を防ぐ特徴を持つ。このアルゴリズムでは、ノードtでクラスkが正しく分類される確率を$\mathrm{p}(\mathrm{k} \mid \mathrm{t})$とするとき、ジニ係数と呼ばれる不純度

$$\sum_{k' \neq k} \mathrm{p}(\mathrm{k} \mid \mathrm{t})\mathrm{p}(\mathrm{k}' \mid \mathrm{t}) = \sum_{k} \mathrm{p}(\mathrm{k} \mid \mathrm{t})\big(1 - \mathrm{p}(\mathrm{k} \mid \mathrm{t})\big) = 1 - \sum_{k} \mathrm{p}^2(\mathrm{k} \mid \mathrm{t})$$

を最も大きく減少させるように決定木を学習する。ジニ係数を減少させる説明変数ほど重要であると考えることで、説明変数の影響を推定することができ、失業率の予測に対する説明変数の重要度を評価した。ただ

し、この評価方法では説明変数の影響が正の方向か負の方向かは評価できない。

そこで、SHAP 値を用いて説明変数の影響の方向性も含めた評価を行なった。SHAP 値は協力ゲーム理論の Shapley 値を基にし、各特徴量のモデル予測への寄与率を表す。入力データをx、モデルをf、特徴量の数をM、xを簡略化したベクトルを$x' \in \{0,1\}^M$、fをxに対して局所的に近似したモデルをf_xとするとき、特徴量iの寄与率ϕ_iは

$$\phi_i(f,x) = \sum_{s \subseteq x'} \frac{|s|!\,(M-|s|-1)!}{M!}[f_x(s) - f_x(s \setminus i)]$$

と表される。SHAP値はLocal Accuracy、Missingness、Consistencyの性質を満たし、モデルの予測結果の説明に適している。

2.3 主成分分析と k-平均法

最後に、主成分分析と k-平均法によって、データの次元削減とクラスタリングを行い、変数間の関係の分析と可視化を行なった。

主成分分析はデータの情報を最大限保持したまま次元を縮約する方法であり、分散が最大となる新しい軸にデータを射影する。本研究では 17 次元のデータに対して主成分分析を適用し、得られる主成分の上位３つを利用した。

k-平均法はクラスタ中心の更新とデータ点のクラス分類を繰り返すことでクラスタリングを行うアルゴリズムである。この方法には初期値への依存が大きいという短所がある。本研究ではクラスタの数は 3 とし、17 次元のデータに適用した。外れ値の影響を抑えるために、主成分分析の第三主成分までの各主成分について、上下 5%の値を持つ市区町村は外れ値と見なし、取り除いた上でクラスタリングを行なった。

３．データセットの加工

表 1 に示す様にデータを加工し、17 の変数を得た。最低賃金以外の変数は教育用標準データセット（SSDSE-A）[1]から、最低賃金は厚生労働省のホームページ[2]から全国 1739 の市区町村ごとのデータ[3]を取得した。

データの前処理として、すべての変数を平均が０、分散が１になるように標準化して使用した。これらの説明変数は機械学習モデルの解釈性を高めるために、多重共線性を考慮して選択した。被説明変数である失業率を除いた変数の VIF（分散拡大因子）は表 1 に示す様にいずれも 5 未満の値となったため、説明変数の多重共線性は深刻ではないと判断した。

表 1 変数の一覧

変数名	説明	単位	平均	標準偏差	VIF（標準化後）
人口密度（可住面積当たり）	総人口 / 可住面積	人 / ha	13.82	26.53	1.327
高齢者率	65才以上人口 / 総人口	%	34.74	1.354	4.501
外国人人口	外国人人口 /総人口	%	1.354	1.362	1.507
出生率	出生数 / 総人口	%	0.574	0.253	4.714
転入超過率（日本人移動者）	（転入者数-転出者数）/ 総人口	%	-0.3218	0.8586	1.914
転出者率	転出者数 / 総人口	%	3.476	1.638	2.898
婚姻率	婚姻数 / 総人口	件 / 千人	3.35	1.58	3.950
可住面積率	可住面積 / 総面積	%	48.99	20.43	2.074
事業所数（民営）	事業所数 / 総人口	千所 / 人	49.00	20.44	1.469
地方税（市町村財政）	地方税 / 総人口	円 / 千人	148.1	134.5	1.895

1 SSDSE：教育用標準データセット，独立行政法人統計センター

2 地域別最低賃金の全国一覧，厚生労働省 https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/koyou_roudou/roudoukijun/minimumichiran/index.html

3 福島県葛尾村と福島県双葉町のデータは欠損値を含むため削除した。

学校数	（義務教育学校数+高等学校数)/ 総人口	校 / 千人	0.0755	0.276	1.180
労働力人口	(就業者数+完全失業者数) / 総人口	%	50.42	4.537	2.610
第１次産業就業者割合	第１次産業就業者 / 就業者数	%	10.27	9.854	3.566
第２次産業就業者割合	第２次産業就業者 / 就業者数	%	24.70	8.118	2.370
図書館数	図書館数 / 総人口	館 / 千人	0.069	0.199	1.157
最低賃金	都道府県ごと	円	853.3	59.8	1.987
失業率	就業者数 /（就業者数+完全失業者数)	%	3.645	1.151	-

４．データ分析の結果

4.1 重回帰と Elastic Net 回帰の結果

重回帰と Elastic Net 回帰の決定係数を表２、回帰係数を図１に示す。Elastic Net 回帰のハイパーパラメータは、チューニングの結果、正則化の強さを表すパラメータλは 0.10、L1 正則化と L2 正則化の割合を決めるパラメータαは 0.5 を用いた。テストデータに対して、Elastic Net 回帰の方が重回帰よりも決定係数が大きく、Elastic Net 回帰の正則化が過学習を効果的に抑制できていることが示唆される。回帰係数を比較すると、値の大きさの順番が二つの回帰で類似していた。Elastic Net 回帰では外国人人口、婚姻率、地方税の係数の値が 0 となり、正則化の効果が確認された。

4.2 ランダムフォレスト回帰の結果

ランダムフォレスト回帰の決定係数を表２、ジニ係数と SHAP 値に基づいて算出した説明変数の重要度を図２、説明変数ごとの SHAP 値を図３に示す。この回帰モデルは今回検証したモデルの中では訓練データ、テストデータ共に最も決定係数の値が大きくなった。しかし、訓練データとテストデータの決定係数に差があり、過学習が起こっていた。ジニ係数と SHAP 値に基づく説明変数の重要度は、おおむね一致していた。図３は各説明変数の SHAP 値をプロットしたもので、説明変数の値と SHAP 値の関係性を読み取ることができる。例えば、最上段の労働力人口は変数の値が大きい点（赤色に近い点）ほど SHAP 値の値が小さくなっている傾向が見られ、失業率と負の関係にあることが示唆される。

表 2 回帰モデルの決定係数

手法	訓練データ	テストデータ
重回帰	0.3558	0.2767
Elastic Net回帰	0.3131	0.3351
ランダムフォレスト回帰	0.9159	0.4754

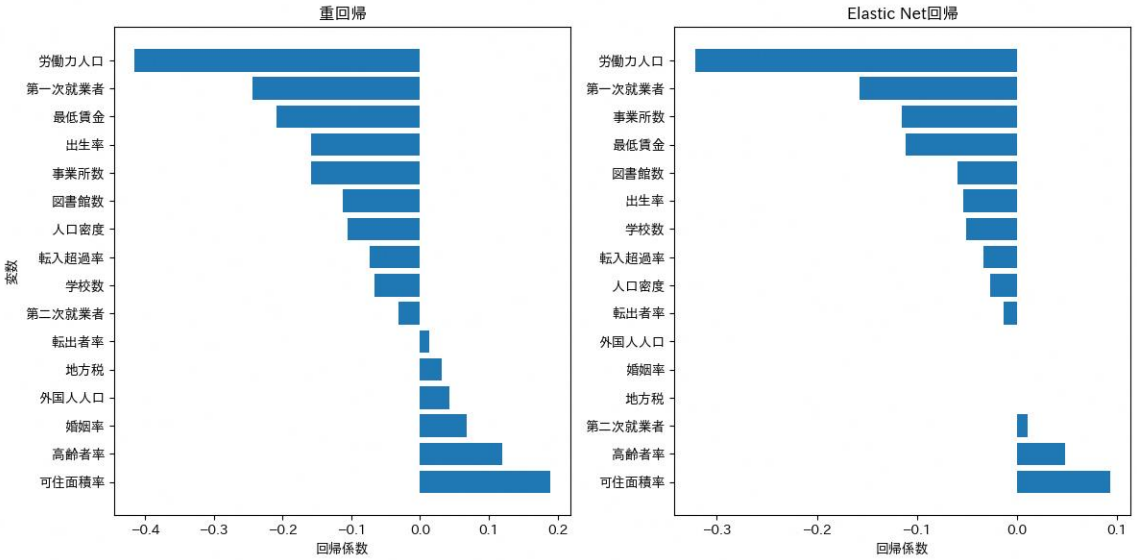

図 1 重回帰と Elastic Net 回帰の回帰係数

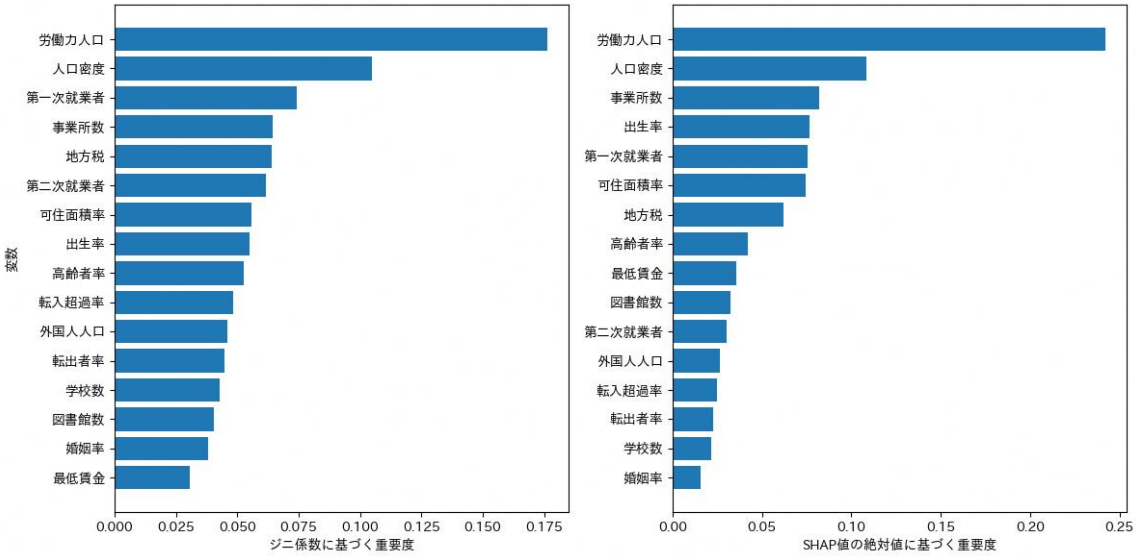

図 2 ランダムフォレスト回帰における説明変数の重要度

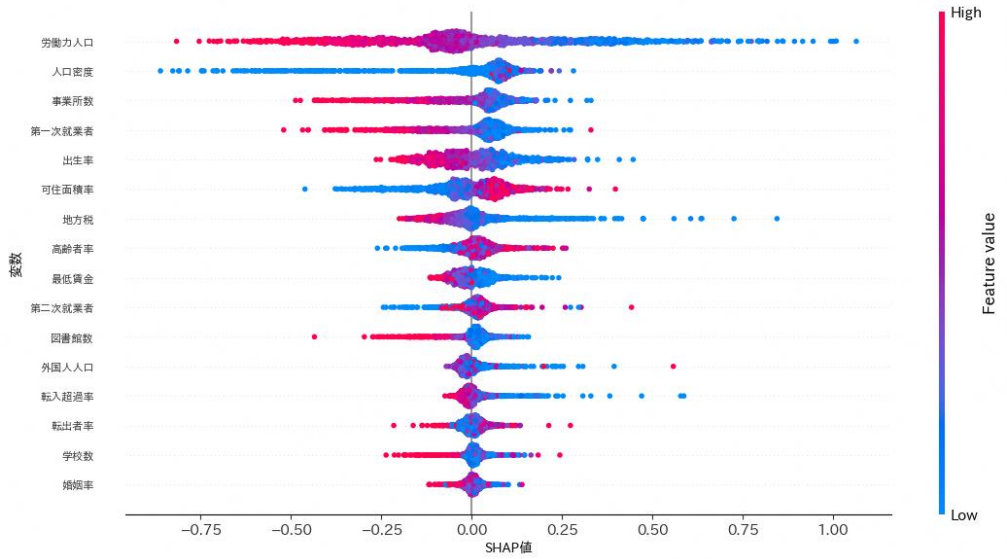

図 3 ランダムフォレスト回帰の SHAP 値

4.3 主成分分析とk-平均クラスタリングの結果

主成分分析とk-平均クラスタリングの結果は表３、表４、図４の通りである。主成分分析の第三主成分までの累積寄与率は約0.50となった。k-平均クラスタリングでは、各クラスの要素数の差が大きくなく、均等なクラスタリングが行われた。図４には、同じ散布図を異なる角度から４枚表示しており、第一主成分の方向で大きくクラスが分かれていることが分かる。このことから、主成分分析の結果とクラスタリングの結果が整合していることが確認された。

表 3 主成分分析の結果

	第一主成分	第二主成分	第三主成分
寄与率	0.2392	0.1736	0.0904
人口密度	0.3581	-0.1059	-0.3661
高齢者率	-0.4185	-0.0085	-0.0806
外国人人口	0.2083	-0.0035	-0.0819
出生率	0.3188	0.3227	0.2044
転入超過率	0.0985	-0.2836	-0.1680
転出者率	0.1916	0.4041	-0.1905
婚姻率	0.3293	0.3367	0.1230
可住面積率	0.3569	-0.1583	-0.0139
事業所数	-0.1768	0.1820	-0.3197
地方税	0.1595	0.3503	0.1806
学校数	-0.0430	0.1931	-0.1643
労働力人口	-0.1450	0.3410	0.1453
第１次産業就業者	-0.3169	0.2096	-0.1816
第２次産業就業者	0.0266	-0.1304	0.5898
図書館数	-0.0505	0.1778	-0.0675
最低賃金	0.2788	-0.1161	-0.3535
失業率	0.0894	-0.2942	0.1935

表 4 k-平均クラスタリングの結果

		クラスタ0	クラスタ1	クラスタ2
市区町村数		544	339	382
主成分の平均値 (PC1, PC2, PC3)		(-1.441, 0.186, -0.219)	(1.739, -0.522, 0.190)	(0.084, -0.406, 0.780)
市区町村例と主成分値		青森県野辺地町 (-2.51, -0.79, 0.34) 長野県小川村 (-2.52, 1.00, 0.73) 和歌山県海南市 (-0.18, 0.07, -0.60)	福島県本宮市 (2.66, -1.37, -0.45) 茨城県日立市 (2.91, -0.93, -1.12) 富山県南砺市 (3.05, 0.45, 0.52)	栃木県那珂川町 (-0.87, -0.82, 1.27) 埼玉県八潮市 (-0.28, -0.40, 1.25) 京都府舞鶴市 (-0.53, -0.99, 0.95)
平均値	人口密度	-0.412	0.333	-0.243
	高齢者率	0.657	-0.923	-0.200
	外国人人口	-0.388	-0.009	0.280
	出生率	-0.352	0.626	-0.084
	転入超過率	-0.322	0.524	0.039
	転出者率	-0.197	0.121	-0.425
	婚姻率	-0.335	0.433	-0.085
	可住面積率	-0.597	0.881	-0.039
	事業所数	0.238	-0.571	-0.112
	地方税	-0.191	-0.010	0.025
	学校数	0.011	-0.148	-0.089
	労働力人口	0.196	-0.561	0.174
	第１次産業就業者	0.554	-0.719	-0.413
	第２次産業就業者	-0.210	-0.124	0.846
	図書館数	0.015	-0.194	-0.102
	最低賃金	-0.513	0.198	0.045
	失業率	-0.052	0.283	0.030

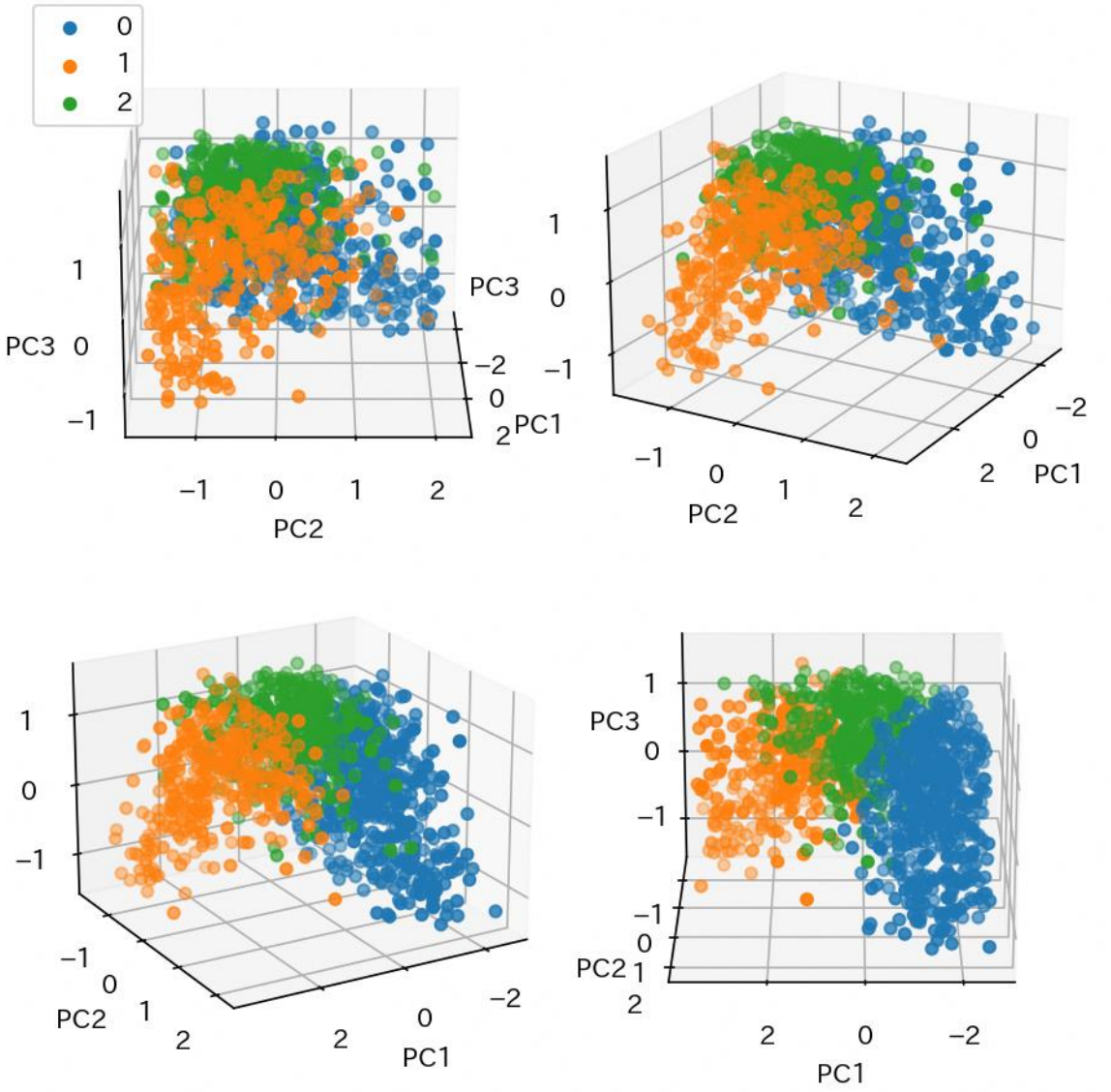

図 4 主成分分析と k-平均クラスタリングの結果

５．結果の解釈

5.1 重回帰、Elastic Net 回帰、ランダムフォレスト回帰の結果に対する解釈

重回帰、Elastic Net 回帰、ランダムフォレスト回帰の結果から失業率と説明変数の関係を評価できる。これらの回帰では、一貫して、労働力人口、事業所数、最低賃金、図書館数、出生率が失業率と負の相関、高齢者率が失業率と正の相関を持っていた。これらの結果は、都市部や雇用機会の豊富な地域では失業率が低いことを表していると考えられる。特に、労働力人口は失業率の予測に対する寄与がどの回帰でも最も大きいことが示唆され、労働力人口と失業率との負の相関があると考えられる。また、第一次就業者数の負の相関と可重面積率の正の相関も全ての回帰で一貫して見られた。このことからは、自然環境が豊かで第一次産業が盛んな地域では失業者が少ない傾向にあることが推察される。

ランダムフォレスト回帰の結果からは重回帰や Elastic Net 回帰の線形モデルでは捉えきれない関係性が推察された。人口密度と失業率との正の相関、地方税、転入超過率、外国人人口と失業率との負の相関がランダムフォレスト回帰の結果のみに表れていた。SHAP 値のプロットを見ると、これらの相関のばらつきが大きいため、これらの変数は他の変数との相互作用を持つのではないかと考えられる。人口密度はジニ係数、SHAP

値からの算出方法のどちらでも労働力人口に次いで重要度が高いため、失業率の予測に大きく寄与したことが示唆される。このようなランダムフォレスト回帰の表現力の高さが決定係数の高さにも反映されていると考えられる。

5.2 主成分分析、k-平均クラスタリングの結果に対する解釈

表４や図４から k-平均クラスタリングでは第一主成分の方向で大きくクラスが分かれていることが分かり、主成分分析の結果と整合的であった。クラス 0 は第一主成分の値が大きく、クラス 1 は値が小さく、クラス 2 はその中間の値を取る。表３から第一主成分は人口密度や出生率、婚姻率の成分が正の方向に大きく、高齢者率や第一次産業就業者数の成分が負の方向に大きいことが分かる。このことから第一主成分は都市化の度合いを反映していると考えられ、クラス 0 は農村地域、クラス 1 は都市部、クラス 2 はその中間の特性をもつ地域として解釈できる。この解釈は表４にある、各クラスの高齢化率や可住面積率、第一次産業就業者の平均値からも妥当である。実際、表４中の例にある、クラス 0 の青森県野辺地町は豪雪地帯の過疎地域、クラス 1 の茨城県日立市は工業都市として知られる都市部、クラス 2 の京都府舞鶴市は都市と農村の特性を併せ持つ地域である。

ただし、失業率の大小についてはこのクラスタ分類で明確な違いを特定することは難しい。クラス 1 では失業率の平均値がやや高いが、これは都市部の失業率がやや高い傾向にあるためだと考えられる。失業率は今回の主成分分析やクラスタリングでは明確に現れない、他の多様な要因を持っていると考えられる。

5.3 結論と今後の課題

本研究では、市区町村ごとの失業率に影響する要因を、古典的な重回帰分析と機械学習技術を活用して分析した。古典的な回帰分析とランダムフォレスト回帰の双方からの一貫した結果として、労働力人口や第一次産業の活動が失業率に影響を与えることが確認された。ランダムフォレスト回帰モデルの結果からは、人口密度、地方税、転入超過率、外国人人口といった要因が他の要因と組み合わさって重要な要因となることも明らかになった。

主成分分析と k-平均法によるクラスタリングの結果では、市区町村の特性が都市部であるかどうかという軸に大きく現れることが分かった。しかし、失業率自体は明確に現れなかったことから、失業率が複数の要因に影響を受ける複雑な指標であると示唆される。これらの結果を受けて、失業率の低下策としては、単一の対策ではなく、各市区町村の独自の状況に対応する多面的なアプローチが効果的だと考えられる。

一方、機械学習手法の解釈性の問題や、ランダムフォレスト回帰の過学習、k-平均法の初期値依存性といった分析自体の限界が確認された。今後の研究では、説明変数の精緻な選択や、モデルの複雑性の増加に伴う予測精度の向上と解釈の困難性とのトレードオフ、さらに統計的因果推論や時系列データを用いた将来予測の検討が課題である。

参考文献

(1) Burkart, Nadia, and Marco F. Huber. :“A survey on the explainability of supervised machine learning.”、*Journal of Artificial Intelligence Research*、70、pp.245-317 (2021).

(2) Mittal,Mamta, et al :“Monitoring the impact of economic crisis on crime in India using machine learning.”、*Computational Economics*、53、pp.1467-1485 (2019).

(3) Yurtsever, Mustafa. :“Unemployment rate forecasting: LSTM-GRU hybrid approach.”、*Journal for*

Labour Market Research.、57、pp.1-9 (2023).