

2024 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]

地方創生を推進する多角的複合指標の提案
—SVM に基づく主観的でない変数選択と
重み付け手法の検討—

衣川 凌太

(神戸大学国際人間科学部環境共生学科)

地方創生を推進する多角的複合指標の提案

—SVM に基づく主観的でない変数選択と重み付け手法の検討—

衣川凌太^{*1}

^{*1}：神戸大学国際人間科学部環境共生学科

1. はじめに

1.1. 研究の動機と目的

本研究は、筆者らが執筆した「独自指標作成による地方創生の方法論と兵庫県活性化の提案」を改良した研究である。過去の研究では、兵庫県が抱える複雑な課題をデータサイエンスの手法を用いて読み解き、分析結果より活性化のために必要な施策を提案することを目的としていた。今後の課題として独自指標の信憑性の評価を挙げていたが、執筆時点より改良を加えた後、実際に兵庫県統計課の課員の方から信憑性の評価を頂けた。しかし、筆者は、複合指標を作成する過程で複合指標を構成する指標（サブ指標）の選択やそれらを集約する際の重み付けにおいて主観的な判断が必要であることに、統計的な信憑性が欠けるのではないかと感じていた。これを踏まえ、統計技法に基づく複合指標のサブ指標選択と重み付けの手法について検討する共に、地方創生とデータサイエンスを組み合わせる際の主観性のあり方についても考えることを本研究の目的とする。

1.2. 複合指標に関する先行研究

筆者ら(2024)^[1]は、サブ指標選択の主観性排除のために、国・地方自治体が公開する資料を元にテキストマイニングを行うことで、地域課題や地方創生に関するキーワードを抽出し、サブ指標選択の参考とした。また、複合指標化する際に似た意味合いを持つサブ指標は除くべきだと一般的には考えられるため、主に重回帰分析の多重共線性の確認に用いられる VIF¹を活用して、サブ指標の削減を行った。これらの手法は主観性排除に寄与すると考えられるが、変数選択時の主観性は排除されておらず、あくまでもその後の主観性排除であるため十分な手法とは言えない。

ここで一度、地域研究における複合指標とはどのようなものか明らかにしたい。Kitchen, Lauriault, McArdle (2015)^[2]によると、都市に関する指標は、都市の様々な側面や部分がどのように展開して機能しているかを示すものだとして理解されているとして、中でも複合指標は、単一指標では明らかにできない様々な現象が相互に関連した多次元的な事柄を認識できるとしている。ただし、指標やダッシュボードによる新たな取り組みは、偶発性や欠点、政治性を認める必要があり、その有用性や価値を押し売りするものでもないとも言及しており、都市の歴史や地理などを踏まえることの必要性を強調している。Grupp, Schubert (2010)^[3]は、注意深く重み付けを行えば、複合指標は、政治分野において重要なトピックに注目させることができるため、価値ある貢献をしえると結論付けており、複合指標とレーダーチャート等の多次元的な表現を組み合わせることを推奨している。

複合指標の中でも、EU の IUS²で用いられる複合指標 SII が議論の対象として多く取り上げられ、サブ指標の選び方や重みを変更すると、SII のサマリースコアや国の順位が大きく変動することを示している。また、順位付けすること自体が不適切だと言及しているものもある。(Grupp, Mogee (2004)^[4], Edquist, Zabala-Iturriagoitia, Barbero, Zofío (2018)^[5])

¹ Variance Inflation Factor の略。一般的に、 $VIF \geq 10(R^2 \geq 0.95)$ が多重共線性の目安とされる。

² Innovation Union Scoreboard の略。EU メンバー国の研究・イノベーション能力を評価するスコアボード。SII は、Summary Innovation Index の略。

次に、複合指標作成の注意点について明らかにする。Cherchye, Moesen, Puyenbroeck (2004)^[6]は、複合指標を作成する過程で解決しなくてはならない主な問題を挙げており、指標を選ぶこと、どのような前処理（スケーリング、外れ値処理等）を適応するか決めること、複数の指標をどのように集約するか³の3つだとしている。Gruupp ら(2004)^[4]は、格差の大きさを隠さないようにするために、サブ指標を等間隔尺度でスケーリングすることを提案している。Wong (2009)^[7]は、例えば同じデータと分析手法を用いていたとしても、同地位で異なる空間単位もしくは地域系でまとめると、一貫した分析結果を導かないという可変単位地区問題(MAUP) が、地理学的研究だけでなく社会科学分野などにも影響を及ぼすと言及している。これに関しては、筆者ら(2024)^[1]の研究でも兵庫県を例に挙げて確認した。

1.3. 日本における複合指標の事例

日本の市区町村を対象として作成されている複合指標の例を5つ挙げる。

表 1 日本の市区町村を対象とした複合指標

指標名 ³	機関	特徴	改善すべき課題
地域評価指標	内閣府/内閣官房	・ 地域特性の見える化 ・ 7分野 22項目	・ 全てのデータが収集できない（地域依存）
地方創生 SDGs ローカル指標	内閣府	・ 日本の実情合わせた SDGs 指標	・ 市区町村データ欠損 ・ 等間隔尺度でない
地域指標分析ツール	環境省/(株)価値創造研究所	・ 政策立案に活用可 ・ 地域特徴を偏差値化	・ 単純平均で算出したスコアで順位付け助長
地域指標	農林水産省	・ 土地利用を類型化 ・ 安定的な指標	・ あくまでも農林対象 ・ 経済指標不使用
地域力指標	(公財)中部圏社会経済研究所	・ 2観点の指標 ・ SVR による変数選択	・ 順位ベースでの比較 ・ データ年度が不揃い

細かな違いはあるが、いずれの複合指標も統計データを活用した地方公共団体の課題把握や政策立案を推進するために作成されている。しかし、それらの用途で活用するためには、ある程度の信憑性が担保されている必要があるが、採用データや複合指標計算式、順位付けの助長等改善すべき課題が見受けられる。国は、地方公共団体での証拠に基づく政策立案（EBPM）の取り組みを推進しているが、全ての自治体でデータ分析基盤が整っているわけではない。ゆえに、手軽にデータ分析が行えるツールが必要になるが、既存の複合指標は適当とは断言しがたい。このままでは、EBPM に取り組める団体とそうでない団体が格差が広がるだろう。

ただ、参考になる部分も多い。政府機関主導で作成されている複合指標は、専門家の指導の下でサブ指標が選出されていることが多いため、主観的選択であるが比較的信頼でき、指標選択の参考になる。また、地域力指標では、鈴木、田辺（2016）^[8]の「感応度分析法」を参考に、サポートベクター回帰（SVR）を用いた主観的でないサブ指標選択をしている。更に、その過程で同時に重みも算出している。

³ 地域評価指標：<https://www.chisou.go.jp/sousei/about/chiikiapproach/datasheet/index.html>

地方創生 SDGs ローカル指標：https://www.chisou.go.jp/tiiki/kankyo/kaigi/sonota/sdgs_shihyou_risuto_2.pdf

地域指標分析ツール：<https://chiikijunkan.env.go.jp/manabu/shihyou-bunseki/>

地域指標：https://www.maff.go.jp/j/tokei/census/shuraku_data/2020/ia/index.html

地域力指標：<https://www.criser.jp/research/bunnseki/shihyo.html>

2.1. 研究方針

使用するデータは、**SSDSE-市区町村(SSDSE-A)**のみとする。理由は、欠損値の無い市区町村データであること、サブ指標選択の主観性をできる限り減らせること、たとえ恣意的なデータでなくとも統計技法に基づくサブ指標選択と重み付けが可能であるかの検証ができることの3つである。データ収集年度は2020年度で揃える。

統計技法に基づいて変数選択と重み付けを行う場合、目的変数が必要となる。そこで、本研究では国土交通省が公表している都市類型⁴を目的変数として利用する。

[illegible]

図 1 分析手順

SSDSE-A のデータを用いてデータセットを作成する。この段階で、外れ値の置換とスケーリングを行う。説明変数として用いるデータセットは、都市類型に合わせて同じ 863 市区に減らした。

都市類型表には、全 841 市区のデータが収録されているが、東京 23 区は特別区部として一つにまとめられていたので 23 区に展開した。なお、特別区部の都市類型は a 型であったため、23 区いずれの都市類型も a 型とした。ゆえに、全 863 市区のデータを収録している。また、2019 年 5 月 1 日より兵庫県の新潟市は丹波篠山市に市名を変更しているが、地域類型表での市名は篠山市のままであったため、地域類型は b 型そのままです市名を丹波篠山市に変更している。

4 パーソントリップ調査の円滑化のために、人口規模等から全 841 市区を a-j の 10 の都市類型に分類したもの。
(国土交通省 都市類型対応表: <https://www.mlit.go.jp/common/001241794.pdf>)

って、[a, b, c, d, e, f, g, h, i, j] = [9, 8, 7, 6, 5, 4, 3, 2, 1, 0] と置換した。

2) サポートベクターマシンを用いた変数選択

鈴木・田辺(2016)^[8]の研究を参考に、サポートベクターマシン(SVM)を用いた変数選択を行う。線形重回帰分析を用いた変数選択との違いは、目的変数と説明変数が非線形分布でも分析可能な点である。田辺、鈴木(2020)^[9]は、目的変数と非線形関係にあり、分布の対称性が高い連続変数を説明変数として用いて、線形重回帰分析を用いた変数減少法による変数選択と SVM の一種である SVR を用いた感度分析法による説明変数とを比較し、SVR と感度分析法の有用性、およびその解析による要因の信頼性を実証している。

なお、本研究で用いる目的変数は、離散変数であるため連続変数を対象としている SVR は利用できない。ゆえに、SVM で代用して変数選択を行う。具体的な分析手順は以下の通りである。

- (ア) 都市類型データを目的変数に、データセットを説明変数にして、SVM の最適なモデルを学習させる。この際、過学習を防ぐため、ホールドアウト法で学習データとテストデータに 7:3 の割合で分割する。また、ハイパーパラメータ(C, gamma, kernel)の最適化にはグリッドサーチを、モデルの汎化性能向上のためには 3 分割交差平均法を用いる。
- (イ) テストデータの説明変数を (ア) のモデルに当てはめて、都市類型を予測する。その予測値とテストデータの目的変数との二乗平均平方誤差 (RMSE) を求める。なお、1 回目に算出する RMSE は、基準値として最終的な変数選択の際に用いる。
- (ウ) データセットの説明変数の内、ある 1 つのサブ指標以外のサブ指標の値を全て平均値に置換し、そのデータセットを (ア) のモデルに当てはめて、都市類型を予測する。その予測値を説明変数、都市類型データを目的変数とする単回帰分析を行い、回帰直線の傾きを求め、それを重み (感応度) とする。
- (エ) (ウ) を全てのサブ指標について行い、重みの絶対値が最も低かったサブ指標を削除する。
- (オ) (ア) ~ (オ) を、説明変数が 1 つになるまで繰り返す。(モデル最適化は 1 回目のみ。)
- (カ) 変数を削減していく過程で RMSE は増加していくので、RMSE の基準値を超えない最大変数削減量を求め、残ったサブ指標を以後の分析で用いることとする。

変数選択終了後、重みの正負を確認して選択されたサブ指標の吟味を行う。比較対象として、回帰分析と重回帰分析の偏回帰係数を用いることとする。

3) 多重共線性の確認

似た意味合いを持つサブ指標を除くため、主に重回帰分析の多重共線性の確認に用いられる VIF を活用して、サブ指標の削減を行う。 $VIF = \frac{1}{1-R^2}$ で求められることから、重相関行列を用いて確認する。サブ指標削減基準は、 $VIF \geq 3 \Leftrightarrow R^2 \geq 0.67$ とする。

4) サブ指標の重み付け

3)までの処理を行ったデータセットを階層クラスター分析にかけることで要因分解する。要因分解には、主成分分析や因子分析等が主に使われるが、3)の処理でデータセットのサブ指標間の相関関係が低いを確認しているので、階層クラスター分析を用いることにした。

データセットをいくつかの要因に分解後、複合指標化してサマリースコアを算出できるようサブ指標の重み付けを行う。重み付けの手法として以下の 4 つを検証し、最適なものを選択する。

- (ア) 単純平均
- (イ) 19 のサブ指標全てを説明変数、地域類型を目的変数とした SVM での重みを利用
- (ウ) 要因ごとに、サブ指標を説明変数、地域類型を目的変数とした SVM での重みを利用

(エ) サブ指標の元データの分散を求め、分散の逆数を重み⁵として利用

ただし、(ア)と(エ)の場合は要因ごとのサマリースコアの最大値は変化しないが、(イ)と

(ウ)の場合は変動する。よって、サブ指標のスケーリング値がいずれも最大値であると仮定した

ときのサマリースコアをSVMでの重みを利用して求め、そのサマリースコアでSVMでの重みを除したものを実際の重みとして利用することとする。

要因*i* (市区町村数 *m*、サブ指標数 *n*) の複合指標 *CI* の計算式は、サブ指標を *I*、重みを *w* として、以下のよう示す。

$$CI_{im} = \begin{pmatrix} CI_{i1} \\ \vdots \\ CI_{im} \end{pmatrix} = \begin{pmatrix} I_{11} & \cdots & I_{1n} \\ \vdots & \ddots & \vdots \\ I_{m1} & \cdots & I_{mn} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix}$$

5) 地域型分類

4)で求めた重みを元に要因ごとに複合指標を作成する。その複合指標を階層クラスター分析にかけることで市区町村の地域型分類を行う。地域類型と比較することで指標の妥当性を評価する。

3. データセットの加工

使用したデータセットを、(表2)に示す。最終的には、★印がついているサブ指標を使用した。

原則として、2020年度のデータを使用する。ただし、2020年度のデータが存在せず、2020±1年度のデータが存在している場合、その次に新しい年度のデータを用いて一年あたり変化率⁶を算出し、変化率が5%以下であれば誤差とみなして使用しても良いこととする。また、一部サブ指標の計算過程では、データセットの都合上0で除算しなくてはならない場合があるが、その際の計算結果は0とする。

表 2 データセット一覧

ID	指標名	計算式	単位	出典
X_0	高齢者率	[65歳以上人口]/[総人口]*100	%	A-2024
X_1	後期高齢者率	[75歳以上人口]/[総人口]*100	%	A-2024
★X_2	従属人口比率	[(15歳未満人口)+[65歳以上人口]]/[総人口]*100	%	A-2024
X_3	外国人口比率	[外国人口]/[総人口]*100	%	A-2024
X_4	出生率	[出生数]/[総人口]*1000	%	A-2024/A-2022
X_5	死亡率	[死亡数]/[総人口]*1000	%	A-2024/A-2022
X_6	転入超過率	[転入者数 (日本人移動者)]/[転出者数 (日本人移動者)]*100	%	A-2022
★X_7	平均世帯人員数	[一般世帯人員数]/[一般世帯数]	人	A-2024
★X_8	核家族世帯比率	[核家族世帯数]/[世帯数]*100	%	A-2024
X_9	単独世帯比率	[単独世帯数]/[世帯数]*100	%	A-2024
X_10	高齢夫婦世帯比率	[65歳以上の世帯員がいる核家族世帯数]/[世帯数]*100	%	A-2024
X_11	婚姻率	[婚姻件数]/[総人口]*1000	%	A-2024/A-2022
X_12	離婚率	[離婚件数]/[総人口]*1000	%	A-2024/A-2022
X_13	可住地面積割合	[可住地面積]/[総面積 (北方地域及び竹島を除く)]	%	A-2022
X_14	可住地面積人口密度	[総人口]/[可住地面積]*100	人/km2	A-2024/A-2022
X_15	第一次産業事業所数割合	[事業所数 (民営) (農業、林業)]+[事業所数 (民営) (漁業)]/[事業所数 (民営)]*100	%	A-2024
X_16	第二次産業事業所数割合	[事業所数 (民営) (鉱業、採石業、砂利採取業)]+[事業所数 (民営) (建設業)]+[事業所数 (民営) (製造業)]/[事業所数 (民営)]*100	%	A-2024
★X_17	第三次産業事業所数割合	[[事業所数 (民営) (電気・ガス・熱供給・水道業)]+[事業所数 (民営) (情報通信業)]+[事業所数 (民営) (運輸業、郵便業)]+[事業所数 (民営) (卸売業、小売業)]+[事業所数 (民営) (金融業、保険業)]+[事業所数 (民営) (不動産業、物品賃貸業)]+[事業所数 (民営) (学術研究、専門・技術サービス業)]+[事業所数 (民営) (宿泊業、飲食サービス業)]+[事業所数 (民営) (生活関連サービス業、娯楽業)]+[事業所数 (民営) (教育、学習支援業)]+[事業所数 (民営) (医療、福祉)]+[事業所数 (民営) (複合サービス事業)]+[事業所数 (民営) (サービス業 (他に分類されないもの))]]/[事業所数 (民営)]*100	%	A-2024
X_18	第一次産業従業者数割合	[[従業者数 (民営) (農業、林業)]+[従業者数 (民営) (漁業)]/[従業者数 (民営)]*100	%	A-2024
X_19	第二次産業従業者数割合	[従業者数 (民営) (鉱業、採石業、砂利採取業)]+[従業者数 (民営) (建設業)]+[従業者数 (民営) (製造業)]/[従業者数 (民営)]*100	%	A-2024
X_20	第三次産業従業者数割合	[[従業者数 (民営) (電気・ガス・熱供給・水道業)]+[従業者数 (民営) (情報通信業)]+[従業者数 (民営) (運輸業、郵便業)]+[従業者数 (民営) (卸売業、小売業)]+[従業者数 (民営) (金融業、保険業)]+[従業者数 (民営) (不動産業、物品賃貸業)]+[従業者数 (民営) (学術研究、専門・技術サービス業)]+[従業者数 (民営) (宿泊業、飲食サービス業)]+[従業者数 (民営) (生活関連サービス業、娯楽業)]+[従業者数 (民営) (教育、学習支援業)]+[従業者数 (民営) (医療、福祉)]+[従業者数 (民営) (複合サービス事業)]+[従業者数 (民営) (サービス業 (他に分類されないもの))]]/[従業者数 (民営)]*100	%	A-2024
★X_21	第二次・第三次産業事業所割合	[第二次産業事業所割合]+[第三次事業所割合]	%	A-2024
X_22	第二次・第三次産業従業者割合	[第二次産業従業者割合]+[第三次従業者割合]	%	A-2024
X_23	経常収支比率	[経常収支比率 (市町村財政)]	%	A-2024
X_24	実質公債費比率	[実質公債費比率 (市町村財政)]	%	A-2024
★X_25	地方税比率	[地方税 (市区町村財政)]/[歳入決算額 (市区町村財政)]*100	%	A-2024
★X_26	民生費比率	[民生費 (市区町村財政)]/[歳出決算額 (市区町村財政)]*100	%	A-2024
X_27	土木費比率	[土木費 (市区町村財政)]/[歳出決算額 (市区町村財政)]*100	%	A-2024
★X_28	教育費比率	[教育費 (市区町村財政)]/[歳出決算額 (市区町村財政)]*100	%	A-2024
X_29	災害復旧費比率	[災害復旧費 (市区町村財政)]/[歳出決算額 (市区町村財政)]*100	%	A-2024
★X_30	幼稚園密度	[幼稚園数]/[可住地面積]*100	園/km2	A-2022
X_31	幼稚園及び保育所等数密度	[[幼稚園数]+[保育所等数 (基本表)]]/[可住地面積]*100	施設/km2	A-2023/A-2022
X_32	小学校密度	[小学校数]/[可住地面積]*100	校/km2	A-2022
X_33	中学校密度	[中学校数]/[可住地面積]*100	校/km2	A-2022
X_34	小中学校密度	[[小学校数]+[中学校数]]/[可住地面積]*100	校/km2	A-2022
X_35	高等学校密度	[高等学校数]/[可住地面積]*100	校/km2	A-2022
X_36	小学校一校当たり児童数	[小学校児童数]/[小学校数]	人	A-2024
X_37	中学校一校当たり生徒数	[中学校生徒数]/[中学校数]	人	A-2024
X_38	小中学校一校当たり生徒数	[(小学校児童数)+[中学校生徒数]]/([小学校数]+[中学校数])	人	A-2024
★X_39	高等学校一校当たり生徒数	[高等学校生徒数]/[高等学校数]	人	A-2024
X_40	小学校教員一人当たり児童数	[小学校児童数]/[小学校教員数]	人	A-2024
★X_41	中学校教員一人当たり生徒数	[中学校生徒数]/[中学校教員数]	人	A-2024
X_42	小中学校教員一人当たり生徒数	[(小学校児童数)+[中学校生徒数]]/([小学校教員数]+[中学校教員数])	人	A-2024
X_43	就業者比率	[就業者数]/[総人口]*100	%	A-2024
★X_44	女性就業者比率	[就業者数 (女)]/[総人口 (女)]*100	%	A-2024
X_45	完全失業率	[完全失業者数]/([総人口]-[非労働力人口])*100	%	A-2024
★X_46	女性完全失業率	[完全失業者数 (女)]/([総人口 (女)]-[非労働力人口 (女)])*100	%	A-2024
X_47	非労働力人口比率	[非労働力人口]/[総人口]*100	%	A-2024
X_48	労働率	[就業者数]/([総人口]-[非労働力人口])*100	%	A-2024
★X_49	公共施設数密度	[(公民館数)+[図書館数]]/[可住地面積]*100	施設/km2	A-2024/A-2022
X_50	人口一人当たり公共施設数	[(公民館数)+[図書館数]]/[総人口]	施設	A-2024
X_51	小売店数密度	[小売店数]/[可住地面積]*100	事業所/km2	A-2024/A-2022
X_52	飲食店数密度	[飲食店数]/[可住地面積]*100	事業所/km2	A-2024/A-2022
X_53	大型小売店数密度	[大型小売店数]/[可住地面積]*100	事業所/km2	A-2024/A-2022
X_54	人口一人当たり小売店数	[小売店数]/[総人口]	事業所	A-2024
★X_55	人口一人当たり飲食店数	[飲食店数]/[総人口]	事業所	A-2024
★X_56	人口一人当たり大型小売店数	[大型小売店数]/[総人口]	事業所	A-2024
★X_57	医療施設数密度	[[一般病院数]+[一般診療所数]+[歯科診療所数]]/[可住地面積]*100	施設/km2	A-2024/A-2022
X_58	人口一人当たり医療施設数	[[一般病院数]+[一般診療所数]+[歯科診療所数]]/[総人口]	施設	A-2024/A-2023
X_59	人口一人当たり医師数	[医師数]/[総人口]	人	A-2024/A-2023
★X_60	人口一人当たり医療従事者数	[[医師数]+[歯科医師数]+[薬剤師数]]/[総人口]	人	A-2024/A-2023

⁵ サブ指標数 *m* のデータセットにおいて、あるサブ指標のスケーリング前の分散を *V_i* とすると、 $w_i = \frac{1/V_i}{\sum_{i=1}^m 1/V_i} (i = 0, 1, \dots, m)$

⁶ 市区町村数 *m* の指標において、2020±1年度のデータを *A_i*、2020±1-*n*年度のデータを *B_i* とすると一年あたり変化率は、 $\frac{1}{mn} \sum_i^m \frac{A_i - B_i}{B_i} (i = 0, 1, \dots, m)$

4. 分析結果

4.1. データセットの作成

四分位数を用いて限界値を設定し、外れ値の除外と置換を行った。上限値： $Q3 + 1.5 * IQR$ 、下限値： $Q1 - 1.5 * IQR$ と設定⁷し、上限値以上もしくは下限値以下に該当するデータはそれぞれの上限值で置換する。

スケーリングでは、SVM で分析可能な形にするため正規化を行う。ただし、そのまま正規化を行うと、間隔尺度が各サブ指標の最大値と最小値のレンジによって異なってしまう。よって、先に標準化を行い、共通の最大値と最小値を設定した上で正規化を行う。各サブ指標の標準化得点の最大値の平均が2.689、最小値の平均が-2.400であったので、最大値：2.5、最小値：-2.5と設定した。

あるデータ x_{mn} （m：市区町村, n：サブ指標）の正規化得点 x'_{mn} の計算式を次に示す。

$$x'_{mn} = \begin{cases} 0, & (x_{mn} < -2.5) \\ \frac{x_{mn} + 2.5}{5}, & (|x_{mn}| \leq 2.5) \\ 1, & (x_{mn} > 2.5) \end{cases}$$

4.2. サポートベクターマシンを用いた変数選択

モデルの学習の結果、最適なハイパーパラメータは、 $C=20$, $\gamma=0.14$, $\text{kernel}='rbf'$ となった。また、基準となる RMSE の値は、 $\text{RMSE}=1.739$ となった。(図 2) は、削減した変数の数と、RMSE の推移を示したグラフである。この結果、RMSE の基準値を超えない最大変数削減量は 31 変数で、 $\text{RMSE}=1.729$ となった。

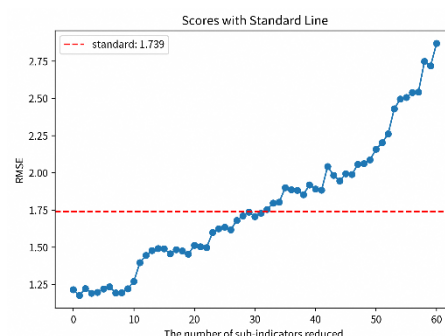


図 2 変数削減数と RMSE の推移

残った サブ指標 (30)	X_2, X_4, X_5, X_7, X_8, X_9, X_17, X_21, X_25, X_26, X_28, X_30, X_38, X_39, X_41, X_42, X_44, X_45, X_46, X_47, X_48, X_49, X_53, X_54, X_55, X_56, X_57, X_58, X_59, X_60
---------------------	---

次に、残ったサブ指標の重みの正負を確認し、妥当な指標であるかを吟味する。比較対象として、各サブ指標を説明変数、地域類型を目的変数とした単回帰分析と重回帰分析の回帰係数の正負を用いる。

表 3 削除対象になったサブ指標

指標名 ID	出生率 X_4	小中学校教員一人 当たり生徒数 X_42	労働率 X_48	大型小売店数 密度 X_53	人口一人当たり 小売店数 X_54
SVM_重み	-0.16026	0	0.151083	0	0.189255
単回帰_回帰係数	0.297969	-0.65006	-0.63803	0.738776	-0.63481
重回帰_偏回帰係数	0.017962	-0.03972	-0.03939	0.045605	-0.03784

(表 3) より、X_42 と X_53 は SVM で算出された重みが 0 であったため削除、X_4, X_48, X_54 は、回帰係数と SVM で算出された重みの正負が一致しなかった、かつ SVM で算出された重みで四分位数を用いて限界値（上限：0.15, 下限：-0.09）を算出し、当該指標の重みが外れ値⁸であると認めら

⁷ Q1：第 1 四分位数、Q3：第 3 四分位数、 $IQR = Q3 - Q1$ ：四分位範囲

⁸ (外れ値と認められるほど SVM で算出された重みの絶対値が大き) = (大きな影響を与える可能性が高い)、と判断したため削除した。

れたため削除した。なお、正負が一致しなかったが、外れ値だとは認められなかったサブ指標は、注意すべき指標とした。

4.3. 多重共線性の確認

(図3)は、(4.3.)で選択したサブ指標を用いた重相関行列を示している。注意すべき指標、かつサブ指標削減基準を満たした指標(X_5, X_9, X_38, X_45, X_47, X_59)は削除した。

ここまでのサブ指標選択・削除を通して残った指標は、以下の19指標である。これらのサブ指標を用いて分析を進める。

残った サブ指標 (19)	X_2, X_7, X_8, X_17, X_21, X_25, X_26, X_28, X_30, X_39, X_41, X_44, X_46, X_49, X_55, X_56, X_57, X_58, X_60
---------------------	---

4.4. サブ指標の重み付け

(4.3.)で完成したデータセットを階層クラスター分析にかけた結果が(図4)である。

ここで、地方創生において重要な要因にはどのようなものがあるだろうか。そこで、筆者は、内閣官房・内閣府総合サイト「地方創生」にて公開されている、「ひと・まち・しごと創生総合戦略⁹」にて、4つの基本目標を参考にすることとした。よって、19のサブ指標を4つの要因に分解した(表4)。

表4 サブ指標の要因分解

要因とサブ指標	特徴	関連する基本目標
地方都市性 X_7, X_39, X_41, X_44	地方部の特徴に関連(農家が多いため女性就業者比率が高い、単位当たりの生徒数が少ない、平均世帯人数が多い)	地方とのつながりを築き、地方への新しい人の流れをつくる
生活利便性 X_8, X_28, X_46, X_49, X_56	家族での暮らしやすさに関連(家族暮らし主体で、暮らしに便利な大型小売店や公共施設、暮らしに必要な教育や女性の就業率というサブ指標が集約)	ひとが集う、安心して暮らすこと ができる魅力的な地域をつくる
都市中心性 X_17, X_55, X_58, X_60	都市の中心性を示す(東京・大阪などの中心都市のスコアが高い、サービス業に関するサブ指標が集約)	稼ぐ地域をつくるとともに、安心して働けるようにする
持続可能性 X_2, X_21, X_25, X_26, X_30, X_57	持続可能性に関連(子供・高齢者、財源に関するサブ指標が集約)	結婚・出産・子育ての希望をかなえる

次に、要因ごとに複合指標化するために重み付け手法の検討をする。検討を容易にするために、現在の市区町村データセットを都道府県ごとで平均を取り、都道府県データセットに変換する。そして、重み付け手法ごとに複合指標を4つずつ作成し、それを元に都道府県の分類をクラスター分析によって行う。いずれの手法でも都道府県を4つの分類に分け、ヒートマップで示したのが(図5)である。

ヒートマップによる可視化、要因ごとのサマリースコアより、手法(ア)と(ウ)を組み合わせることでさらに良い重み付けができるのではないかと考えた。よって、ここでは主観性に頼り、適当な重み配分を検討した結果、手法(ア)の重み:手法(ウ)の重み=8:2の比率で配分することで、もっともらしい分類ができると結論付けた(図6)。この重みを元の1741市区町村データに当てはめる。

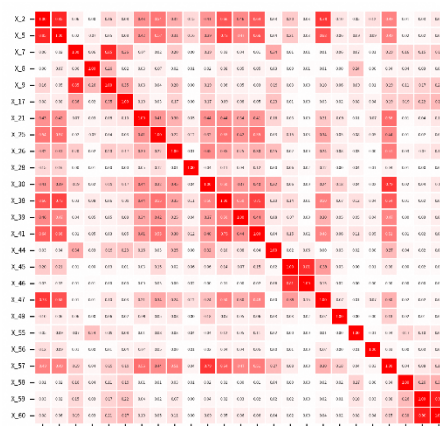


図3 サブ指標の重相関行列

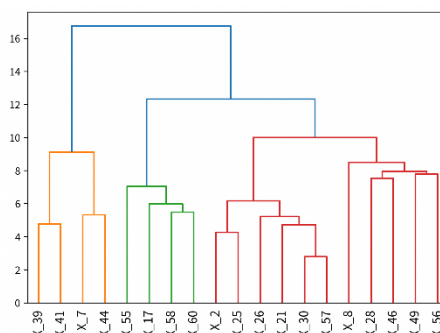


図4 要因分解デンドログラム

⁹ 内閣官房・内閣府総合サイト 地方創生「まち・ひと・しごと創生「長期ビジョン」」「総合戦略」「基本方針」: https://www.chisou.go.jp/sousei/mahishi_index.html

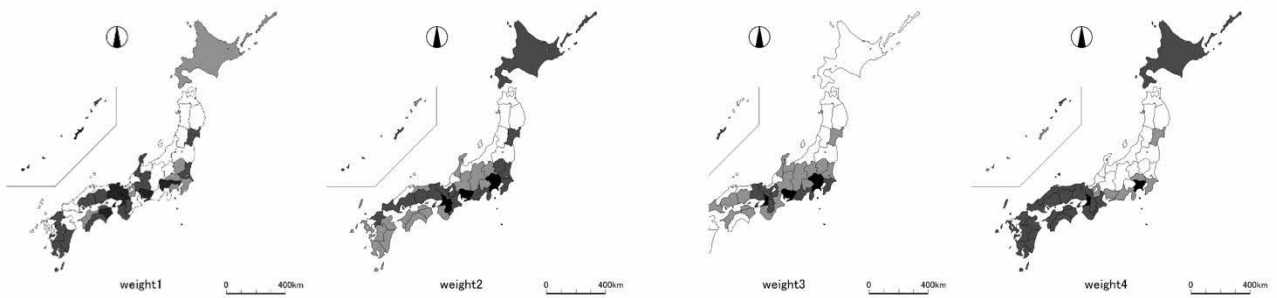


図 5 重み付け手法別地域分類ヒートマップ ※左より手法 (ア) (イ) (ウ) (エ)

4.5. 地域型の分類

(図 6) の分布の確認や地域別要因ごとのサマリースコアを求めることで、その特徴から分類に名称を付け、その結果をレーダーチャートにまとめた (図 7)。目的変数としていた地域分類と本研究の複合指標を用いて市区町村を 10 分類したもの (図 8) を比べると、人口が集中している市街についての分類はもちろん、目的変数ではカバーしきれていなかった地方部の詳細な地域分類も実現していることが分かる。

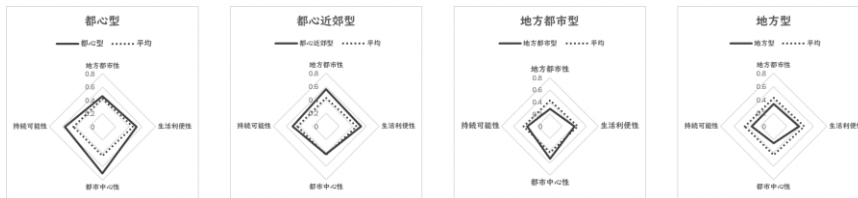


図 7 地域分類別サマリースコアレーダーチャート

5. 結論

本研究では、先行研究や既存の指標を参考に、課題点を解決した市区町村評価のための複合指標を作成した。具体的には、市区町村単位でありながらデータの欠損・年度的大幅なずれ・外れ値を無くし、等間隔尺度でのスケーリングを実施、極力主観性を省いた統計技法に基づく変数選択・重み付け・地域分類を行った。それでいて、筆者ら(2024)^[1]が過去に作成した独自指標と同様に、レーダーチャート (図 7) や棒グラフ (図 9) での容易な分析、要因分解 (図 10) での詳細な分析も実現している。ゆえに第一の研究目的は果たした。本研究を通して第二の目的であった主観性のあり方について考えたい。複合指標作成の過程で主観性を除く行為は、統計的には正しい。しかし、あくまでも地方創生のツールの 1 つであるならば、その地域に寄り添った主観性が必要だと考える。ゆえに、重み付けに際しては、実情を反映することができるならば主観性を用いた独自性があるといいと結論付けたい。

本研究の限界は、使用データが限られていることである。今後の展望としては、更に多くの指標を交えた上で統計的な変数選択法を追求したい。

6. 参考文献

- [1] 衣川遼太、中島筑士、嵯峨佑季、丸山晃平、斎藤真一「独自指標作成による地方創生の方法論と兵庫県活性化の提案：統計データ分析コンペティション 2023 受賞論文 高校生部門 統計活用奨励賞」、統計、75 巻、5 号、pp.77-82 (2024)。
- [2] Kitchin, R., Lauriault, T. P., McArdle, G. : "Knowing and governing cities through urban indicators, city benchmarking and real-time dashboards", Regional Studies, Regional Science, 2 巻、1 号、pp.6-28 (2015)。
- [3] Grupp, H., Schubert, T. : "Review and new evidence on composite innovation indicators for evaluating national performance", Research Policy, 39 巻、1 号、pp.67-78 (2010)。
- [4] Grupp, H., Mogee, M. E. : "Indicators for national science and technology policy: how robust are composite indicators?", Research Policy, 33 巻、9 号、pp.1373-1384 (2004)。
- [5] Edquist, C., Zabala-Iturrigogitia, J. M., Barbero, J., Zofio, J. L. : "On the meaning of innovation performance: Is the synthetic indicator of the innovation Union Scoreboard flawed?", Research Evaluation, 27 巻、3 号、pp.196-211 (2018)。
- [6] Cherehye, L., Moesen, W., van Puyenbroeck, T. : "Legitimizing diverse, yet comparable: on synthesizing social inclusion performance in the EU", Journal of Common Market Studies, 42 巻、5 号、pp.919-955 (2004)。
- [7] D. W. Wong : "Modifiable Areal Unit Problem", International Encyclopedia of Human Geography, pp.169-174 (2009)。
- [8] 鈴木孝弘、田辺和俊 : "幸福度の都道府県間格差の統計分析", 東洋大学紀要 自然科学編, 60 巻、pp.93-112 (2016)。
- [9] 田辺和俊、鈴木孝弘 : "サポートベクター回帰における変数選択の有用性の検証：都道府県別全死因死亡率の影響要因の分析", 統計数理, 68 巻、1 号、pp.175-192 (2020)。

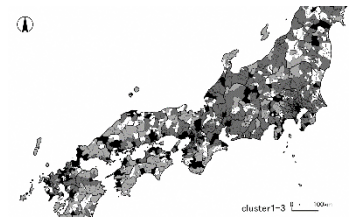


図 6 地域 4 分類

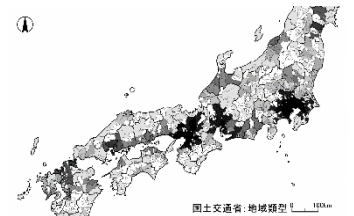


図 8 地域 10 分類

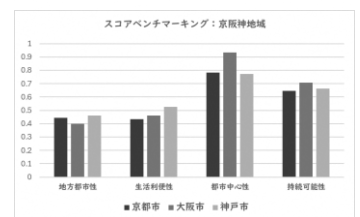


図 9 サマリースコア棒グラフ

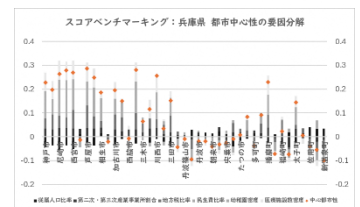


図 10 要因分解