

多目的最適化社会シミュレーションにおける解集合の形状生成要因特定と決定変数空間分析の手法の提案

○伊藤真人 後藤裕介（芝浦工業大学）

Proposal of Method for Identifying Factors Generating the Shape of Solution Sets and for Drawing Decision Variable Space in Solution Sets in Multi-Objective Optimization Social Simulation

*M.ito and Y.Goto (Shibaura Institute of Technology)

概要— 多目的最適化社会シミュレーションにおいて、出力されたパレート解集合に対して効果的な分析手法を明らかにすることは、研究者のみならず一般の利用者がシミュレーションの特性を理解するために有用である。本研究では目的関数空間上のパレート解集合に対しクラスター分析と決定木代理モデルを用いて解空間の形状を決定する要因を特定し、ヒートマップを用いて決定変数空間上のパレート解集合を描画する手法を提案する。

キーワード：多目的最適化社会シミュレーション，決定木代理モデル，ヒートマップ

1 はじめに

近年機械学習の実利用に向けて、解釈性を求める動きが広まっている¹⁾²⁾。そこでは信頼できるAI技術の開発の為に、ブラックボックスモデルを使用する際には人に理解できるような説明を行うことが推奨されている。これは機械学習だけでなく、コンピュータを用いた意思決定の場面で汎用的に求められるべきものであると考えられる。

他方、コンピュータを用いた意思決定支援として、社会シミュレーションが盛んに研究されており、本研究の主要な関連研究(以降先行研究と呼ぶ)である後藤³⁾は、その一環として多目的最適化社会シミュレーションによる解空間に基づく状況分類手法の提案を行っている。しかしそこで提案された手法は遺伝的アルゴリズムを用いた最適化や、エージェントベースシミュレーションという、内部処理を詳細に確認しづらい処理を含むため、解釈性の観点から、理解しやすい説明を行うことが必要であると考えた。

そこで本研究では解集合に対して説明を与える分析手法を提案する。これは、解集合の形状がどのような要因によって形成されるのか、解集合の中の施策はどのような内容になっているのかを分析できる。それにより直面している問題状況において、どのようなことを重視すべきなのか、どのような施策が望ましいのかを理解でき、多目的最適化社会シミュレーションの利用者を支援することができる。と考える。

1.1 社会シミュレーション

社会シミュレーションは、状況と施策をモデルの入力とし、分析者の関心がある評価指標を出力する。

本研究での状況とは、対象とするシステムの初期状態を説明するものであり、関与者による制御が不可能なものである。状況は複数の状況変数と呼ばれるパラメータにより定義される。社会シミュレーションでエージェントベース・モデルを用いる際には、状況変数に基づいて、モデル内のエージェント数や各エージェントの持つ属性値などが割り当てられる。例えば、都市を状況変数とすれば沖縄県那覇市と沖

縄県石垣市のように異なる2つの状況に関してシミュレーションが行われ、異なる結果が出力される。

他方施策とは対象とするシステムの状態に作用するものであり、モデル内のパラメータに影響をもち、関与者が制御可能な性質を持つ。施策は複数の施策変数と呼ばれるパラメータにより定義される。そして、施策変数に基づいてモデル内のエージェントがもつ属性値などに変化が現れる。例えば、給付金制度で、給付金の対象と金額を施策変数とすれば、その対象と金額に応じて住民の経済状況へ作用する。

この社会シミュレーションの分析方法として、シナリオ分析という手法が研究されており、主に個別の施策や個別の状況について、出現するパスの分析を行う。例えば、後藤・高橋⁴⁾では、不確実性の影響によるパスの変化を比較する可視化手法を提案している。そして、先行研究³⁾ではこれらの研究を踏まえて、多目的最適化を用いた状況分類手法を提案し、最適な施策の集合を使って状況を分類することを可能にした。しかしこの研究では、解集合の効果的な分析方法は提案されておらず、更には従来のシナリオ分析も適用し難いと考えた。そこで本研究では機械学習を用いて、分類ごとの特性については解集合の形状と、施策内容の効果的な分析手法を提案する。

1.2 分析手法について

機械学習に解釈性を与える研究として、現在では様々な手法が提案されている。有名なものは、線形モデルやルールモデルを用いて、任意の機械学習と局所的に近似できるモデルを作成し、説明を与えるLIME⁵⁾などである。しかし本研究で提案するのは、全体の傾向を捉えることができる大局的な説明を与える手法である。なぜなら、局所的な分類の境界を知るよりも、分類ごとの傾向を大まかに捉えるほうが解集合の分析を行いやすいためである。類似した手法としてDENG & Houtao⁶⁾の提案したInTreesがあり、これはランダムフォレストという複雑なモデルにルール抽出と単一の決定木モデルによる再現から説明を行っている。決定木モデルは単純な条件分岐を使って分類を行うモデルである為、直感的に理

解し易く可読性が高いモデルとして知られている。

本研究では分類の傾向と解集合の形状を分析する為、この手法にならない決定木モデルによる再現で説明を与える。具体的には、多目的最適化社会シミュレーションの後、目的関数空間上でのクラスター分析をした結果を被説明変数に、状況から抽出した端的な情報を説明変数として決定木を学習する。その結果得られた決定木を分類手法の説明とすることで、分類の根拠を発見することができる。それはつまり、目的関数空間上での解集合の形状について、形状を決定づける変数と値を発見できることを示している。なお、決定木を学習する際に、抽出する情報の細かさや、特性をうまく反映する様な情報の抽出に注意を払うことで決定木の精度を高めることができ、より信頼性の高い分類根拠を発見することができる。

つづいて施策内容の分析を行う為、決定変数空間上でのパレート解集合の可視化手法を提案する。本研究で使用するのは数値データの強弱を色で表し行列を可視化するヒートマップである。これは多くの変数の状態を一望できる情報密度の高い手法である。

この可視化手法は、最適な施策集合を分析できるため、最終的な施策の決定に役立つと考えられる。しかしこの手法には懸念もあり、値を色で表現する関係上、細かな差や絶対的な値を確認することが難しくなってしまう他、情報量が多すぎて認知能力の上限を超えてしまう可能性がある。そのため今回はそれらに注意しながら可視化手法を適用した。

2 提案手法

2.1 手法の概要

まず、本研究の提案手法の概要を Fig. 1 に示す。先行研究³⁾で提案されたものが、図中の左側の部分で、本研究で新たに追加したものが右側の部分である。

本研究でははじめに、先行研究³⁾を参考に状況群に対して多目的最適化を行う。その後2つの分析を行い、解集合の特徴を得る。1つ目の分析は、多目的最適化で出力された解集合について、目的関数空間上での距離を状況ごとに計算し、階層型クラスターリングを行う。そして出力される分類結果と状況を表

す変数から決定木を学習し、その決定木を分析する。その結果、解集合の形状を決める要因を特定できる。

2つ目の分析は決定変数空間における解集合をヒートマップにより可視化し分析を行う。それにより目的関数値に応じて解集合の中でどのように施策が変化するのか、また施策を決める決定変数の関係も分析することができる。

2.2 多目的最適化社会シミュレーション

この処理は基本的に先行研究³⁾に準拠した手法で行う。まず利用するシミュレーションのモデルは、状況と施策とシナリオの3つの入力を持ち、シミュレーション結果を出力する。次にその出力から目的関数値を算出し施策の評価を行う。そして、そのモデルについて、遺伝的アルゴリズムによる進化計算を利用することで状況ごとに最適な施策の集合を獲得する。このシナリオとは、前述の状況とは異なり、シミュレーション中の動作を指定するものである。本研究の適用事例では、失業や給与減少といった処理を行う対象などを指定している。

例えば、状況集合を C 、状況の数を CN 、施策集合を P とした時、状況 $c_i \in C | 1 \leq i \leq CN$ と施策 $p_j \in P | j \in N$ 、シナリオ y が存在するとする。シミュレーションモデル M は状況と施策、シナリオを入力として出力 $z = M(c_i, y, p_j)$ を得る、続いて出力から F_1, F_2 の2つの目的関数値を算出する。その後状況 c_i における施策 p_j を最適化し、施策 c_i に対応する解集合 O_i, D_i を得る。解の個数 $k \in N$ が共通の2種類の解集合があり、1つは目的関数空間上の解集合 $O_i = \{o_{i1}, \dots, o_{ik}\}$ 、もう一つは決定変数空間上の解集合 $D_i = \{d_{i1}, \dots, d_{ik}\}$ である。ここで、目的関数空間上の解 $o_{in} \in \mathbb{R}^2 (1 \leq n \leq k)$ は2次元のベクトルであり、決定変数空間上の解 $d_{in} \in P (1 \leq n \leq k)$ は施策集合に含まれる多次元のベクトルである。そして、この最適化をすべての状況について行い、 CN 個の解集合 $S = \{S_1, \dots, S_{CN}\}$ と $D = \{D_1, \dots, D_{CN}\}$ を獲得する。

2.3 クラスタ分析

今回は先行研究³⁾で行っていた距離計算のうち、目的関数空間における解集合に基づく距離計算法を用いたクラスタ分析で状況の分割を行う。

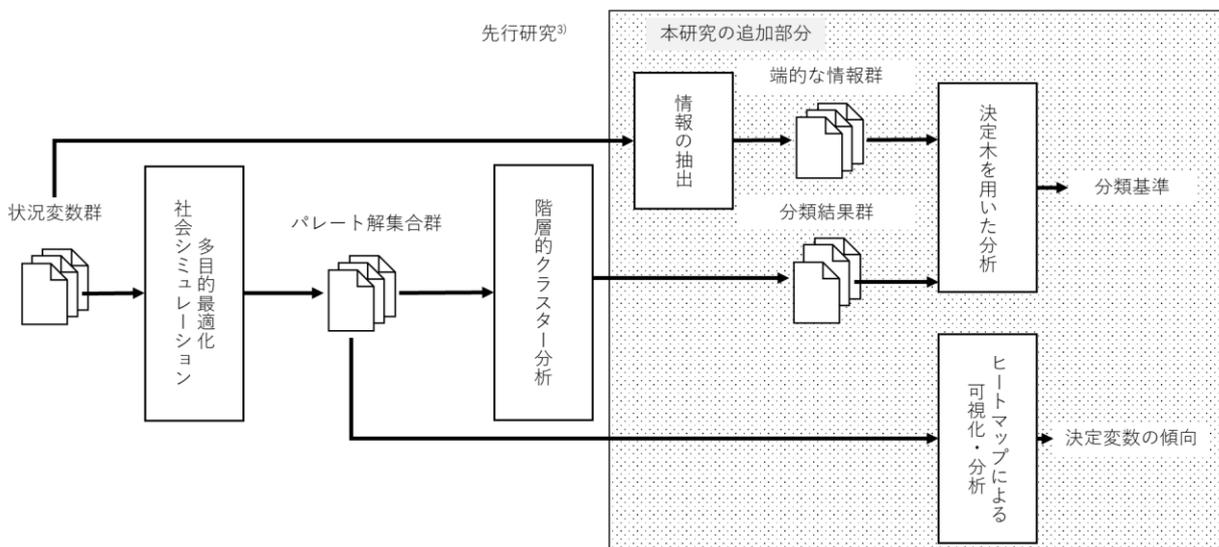


Fig. 1 提案手法の概要

2.3.1 距離計算法

状況 c_1, c_2 間の距離 $dist(c_1, c_2)$ は c_1 から c_2 への距離 $d(c_1, c_2)$ と c_2 から c_1 への距離 $d(c_2, c_1)$ の平均を取ることで求める。 $dist$ と d は解集合 O_1, O_2 を用いて、以下の式で表される。

$$dist(c_1, c_2) = \frac{d(c_1, c_2) + d(c_2, c_1)}{2}$$

$$d(c_1, c_2) = \frac{\sum_{o_{1i} \in O_1} \min_{o_{2j} \in O_2} (|o_{1i} - o_{2j}|)}{|O_1|}$$

2.3.2 クラスタ分析の手順

はじめに階層的クラスタリングによって状況をクラスタに分割し、すべての状況 c_i に対してクラスタごとに決められた番号を割り振る。このクラスタ一つ一つを類型と呼ぶ。続いてデンドログラムを描画し、適切なクラスタリングが行われていることを確認する。本研究における適切なクラスタリングとは、類型ごとの個体数の分散が小さく、極端に個体数の多いクラスタが無いものの事である。

また、類型ごとに解集合の傾向を確認するため、目的関数空間における解集合を描画する。この確認では、類型ごとに解集合たちがどの程度目的関数について最適化できているのかを確認する。本適用事例の場合では、2つの目的関数値が最小の点である(0,0)にどれだけ近づくことができているのかを目視で確認した。

2.4 情報の抽出

ここではシミュレーションの入力である状況変数を用いて、状況 c_i を端的に示す情報のベクトル $\mathbf{a}_i = (a_{i1}, \dots, a_{iN}) | N \in \mathbb{N}$ という説明変数を作成する。なお、エージェントの情報を全て記載するような状況変数は、データとして非常に大きく煩雑であるため、説明変数としていくつかの指標を設定し、状況変数からその指標値たちを求めることで情報を集約する。この説明変数の本体とも言える指標には、状況を表し解集合との関係が見込まれるものが適当である。また、決定木は複数の変数を組み合わせた判断が苦手な為、単体で情報量が少なく他の変数との比率などが意味を持つ変数は不適である。これらを踏まえた上で状況を端的に表す情報を、分析を行う当事者の関心に基いて作成する。

本研究で使用する指標は、エージェントごとの数値データの平均値、エージェントごとの特定の属性値の出現頻度、多目的最適化と関係のある指標の3種類である。一つ目と二つ目は抽出時に問題状況に対する特別な理解を必要としないため、状況変数から容易に抽出することが出来る。しかし最後の一つについては分析したい問題に応じて分析者が設定する必要があるため、抽出の難易度が高く、本研究でも抽出方法の一般化には至らなかった。そのため、参考程度にしかならないが、本適用事例では目的関数値やシナリオと関わりの深いエージェントの出現頻度などを利用した。

2.5 決定木を用いた分析

クラスタ分析の出力の類型と、状況を端的に表す変数を使い決定木モデルを学習し、その処理過程

を確認することで類型ごとに状況の傾向を分析する。分析の手順は、まず分析に適した決定木を作成し、つづいてその決定木の分類結果や精度、そして分類過程を観察することで、類型の説明を得る。

2.5.1 変数

類型の種類数を $TN \in \mathbb{N}$ とすると、被説明変数は、類型 $t \in \{1, \dots, TN\}$ である。説明変数には、2.4にて作成したベクトルを使用する。また、本手法では状況の予測が目的ではなく、状況分類を再現することが目的な為、過学習を恐れる必要はない。よってデータの訓練データとテストデータへの分割は行わない。

2.5.2 分析手法の詳細

本研究では、scikit learn によって提供されているpythonのDecision Trees⁸⁾を用いて決定木を実装した。このプログラムは描画が容易で、かつ可読性が高いため本研究に採用した。このモデルに状況 c_i を表す変数 \mathbf{a}_i を与えると、状況 c_i の類型を予測する。

続いてパラメータの調整方法である。決定木は木の複雑さを適切に設定しなければ、大きく複雑な木となり分析が困難になる。本手法では決定木のパラメータの内、alphaの値について主に調整を行いその事態を防ぐ。本手法で決定木に求められるのは、類型の傾向の説明をできる限り正確に、そして理解しやすい形で与えることである。そのため、決定木の正確性を表す精度と、決定木の理解しやすさを表す複雑度を重視して調整を行う。具体的には、複雑度を増やす事によって生まれる精度の増加が小さくなる付近のalpha値で作られた決定木をバランスの取れた決定木とし、それを分析に使用する。

実際に評価指標に使うのはmacro-F1値とノード数である。macro-F1値は決定木の精度を表す指標の一つで多クラス分類に適用できる指標である。ノード数は、分岐が増えるほど増加するため、決定木の複雑度を表す指標として適切であると考えられる。

このように調整を行った決定木から、分類過程を確認し、分析を行うことで類型の説明を与える。

また、決定木の学習結果を確認する際に、混同行列を参照するほか、割り当てられた葉ごとに目的関数空間上での解集合の形状も確認する。

2.6 ヒートマップによる可視化・分析

これまでの目的関数空間上での解集合を分析する手法とは異なり、ここでは決定変数空間上での解集合 D_i を分析する手法を提案する。決定変数空間上での解集合 D_i を分析するためには、多目的最適化時に利用した決定変数を個体ごとにグラフにして表示するのが望ましい。しかし、決定変数が複雑、つまり解 d_{in} の遺伝子(次元)数が多いと適切な表示法を考えなければならない。本研究で提案する表示法は、ヒートマップを利用したものである。このグラフは、縦横の軸と色を使って3種類の値を表現することができる。そのため、解集合 D_i を一望でき、解ごとの目的関数値も描画することで、遺伝子の相互作用や目的関数値に応じた変化も分析可能になる。

本研究では、Prike⁷⁾を参考に、解集合の遺伝子座について距離計算の後クラスタリングを行い、生成されたデンドログラムに則った順番で遺伝子座を並べヒートマップを表示する。それにより個別の状況 c_i

の解集合 D_i の分析時に、遺伝子同士の関係や遺伝子と目的関数値の関係を理解しやすくなる。しかし、本適用事例では扱わないが、状況同士の解集合を比較する際には、クラスタリングを使わずに遺伝子座を一定の規則で並べるほうが、同じ遺伝子座同士を比較しやすいという利点がある。

また、今回の適用事例は2つの目的関数値を最適化するため、目的関数空間ではパレート解は1つの曲線状に並んでいる。そのため、各個体を目的関数値の順番に並べることで可読性の高いグラフを描画できた。

なお、ヒートマップは情報密度が非常に高いため、表示する変数は絞らなければならない。適用事例では、解集合を通して変化の無い変数を非表示とした。

3 適用事例：コロナ禍における経済ショックに対する給付金施策

本研究では、提案手法の適用事例として、自治体における給付金施策の最適化と、その解集合の分類・分析を行う。自治体は政策立案時に類似した自治体の動向を調査することが知られているため、本手法の解集合分析手法を適用することで、状況に対する理解を深め、参照先の決定に活かすことができると考えられる。

3.1 シミュレーションの概要

使用するシミュレーションは基本的に先行研究³⁾の内容をそのまま踏襲しており、コロナ禍による経済ショックと、それに伴う給付金政策を行う。このシミュレーションでは、自治体の住民1人を1つのエージェントとし、自治体の状況を再現した空間を作る。そこにコロナ禍を模した失業と減給を実施し、その後給付金の支給を行う。そして最終的に実行結果を出力し、そこから2つの経済指標を算出する。

入力は、自治体の状況を再現するために使用する仮想合成人口個票データ、経済ショック内容を表す経済ショックシナリオ、給付金の支給先と支給金額を表す経済支援施策の3つである。このうち、仮想合成人口個票データと経済ショックシナリオは所与のものであり、経済支援施策のみを設計対象とする。

出力は、貧困世帯の減少を表す数値と適切な支給水準の可否を示す数値の2つの経済指標である。

3.2 仮想合成人口個票

モデルの入力である仮想合成人口個票データ⁹⁾は、住民一人一人の属性が記されたテーブルデータである。本研究では、このデータからいくつかの属性を抽出し、シミュレーションに利用する。

利用する属性はTable 1に示す通りで、今後のために詳細説明にて[]内にパラメータ値を示している。表に示す属性のうち、下の5つは支給対象の選択で利用するため、対応集合を表す記号を用意している。

複数の自治体について、2015年の仮想合成人口個票データを頂き実験を行った。その際、自治体の規模が一定の範囲内であった方が自治体同士の比較を行いやすいことから人口と市の種類を指定し、またこのシミュレーションでは職住一致が重要であることから、通勤時間の長い傾向にある都市圏の都府県を除外した。最終的な条件は全国の市の内、人口が2万人以上かつ、指定都市、中核市、施工時特例市以外かつ、東京、千葉、神奈川、埼玉、大阪、兵庫、奈良以外の県に属する自治体となった。その結果、自治体数は499個となり、それは2015年の全市790個の内、約6割となった。

3.3 シミュレーションの処理内容

シミュレーションのモデル M は、決定変数である経済支援策 p_j 、状況である市 c_i 、経済ショック y から実行結果 z を出力し、以下の式で表すことができる。

$$z = M(c_i, y, p_j)$$

そして出力 z を用いて後述の経済指標を計算する。

全体の処理の流れは、先程紹介した仮想合成人口個票データを用いて、実際の自治体を再現した空間を創造することからシミュレーションは開始する。その際、雇用形態が非就業者となっている住民については、産業分類と企業規模を無職[-1]とした。

続いて、予め設定しておいた経済ショックを与える。経済ショックは、産業・雇用形態・企業規模を使って労働者を指定し、その労働者たちの一定割合に対して失業または給与減少を与える処理である。実際の処理では、「宿泊業・飲食サービス業[130]、短時間労働者[20]、企業規模 10~99人[10]の50%に失業」のような処理を複数回行うことになる。この際、失業した住民の属性値のうち産業分類を失業者[-10]、雇用形態を無職[-1]企業規模を無職[-1]とする。

Table 1 仮想合成人口個票データから本研究で利用する属性

属性名	対応集合	種類数	詳細説明([]内の数字は、そのパラメータの値)
世帯人数			数値
年齢			数値
月収			数値
性別		2	男[0], 女[1]
家族類型	A ₁	9	単独世帯[0], 夫婦のみ[1], 夫婦と子ども[2], 男親と子ども[3], 女親と子ども[4], 夫婦と両親(夫の親)[50], 夫婦とひとり親[60], 夫婦・子どもと両親(夫の親)[70], 夫婦・子どもとひとり親(夫の親)[80]
世帯内役割	A ₂	8	単独世帯(男性)[0], 単独世帯(女性)[1], 夫・男親[10], 妻・女親[11], 子ども(女性)[20], 子ども(男性)[20], 親(男性)[30], 親(女性)[31]
産業分類	A ₃	20	日本標準産業分類による
雇用形態	A ₄	4	非就業者[-1], 一般労働者[10], 短時間労働者[20], 臨時労働者[30]
企業規模	A ₅	4	5~9人[5], 10~99人[10], 100~999人[100], 1000人以上[1000]

その後入力として与えた経済支援施策に則って給付金の支給を行う。

経済支援施策 $p_i = (Q, m)$ は給付の対象 Q と支給金額 m で構成される。給付の対象は、Table 1 にて対応集合を示した 5 つの属性によって指定する。各条件に対して合致する住民 A の部分集合を A_1, \dots, A_5 とすると、給付対象の住民の集合 Q は各条件に合致する住民積集合として以下のように定義できる。

$$Q = \bigcap_{k=1}^5 A_k$$

このとき、一律の金額 $m (m > 0)$ が Q に対して給付される。実際の処理では、「単独世帯[0]または女親と子ども[4]、全世帯内役割、宿泊業・飲食サービス業[130]または生活関連サービス業・娯楽業[140]または無職[-1]または失業者[-10]、全企業規模、雇用形態に対して 10 万の給付」のような処理が 1 度だけ行われる。なお、給付金額と支給対象には条件が存在し、その条件を満たす経済支援施策のみを実行する。支給金額は、全人口に対して 20,000 円ずつ支払う規模を上限とする。つまり、 $|\cdot|$ を集合の要素数とすると、以下の式で支給金額 m の条件が表される。

$$|Q| \times m \leq |A| \times 20,000$$

支給対象の条件については弱者保護の方針から優先条件が存在するが、詳細については割愛する。

そうして経済ショックと給付金の支給を行った後、実行結果 z を出力しシミュレーションは終了する。

3.4 経済指標

本適用事例では 2 つの経済指標 F_1, F_2 を算出する。この際、シミュレーションの不確実性を考慮するため、経済指標 F_1, F_2 は 2 回シミュレーションを実行して得られた指標 $f_{11}, f_{12}, f_{21}, f_{22}$ の平均を取る。

計算に使用する変数のうち、 $H_{N=n}^*$ のように上付きの*が付いているものは給付金支給後の変数、付いていないものは経済ショック前の変数である。

経済ショック前から給付金を支給した後の相対的貧困世帯数の変化を表す指標 f_1 は、世帯人数 n の相対的貧困世帯数 $H_{N=n}$ を主軸に、世帯人数 n の人口 $|B_n|$ 、総人口 $|A|$ を用いて世帯人数による重み付けを行い以下の式で表される。

$$f_{1k}(z_k) = \sum_{n=1}^4 \left(\frac{|B_n|}{|A|} \times \frac{H_{N=n}^*}{H_{N=n}} \right)$$

なお、世帯人数 5 人以上の世帯は数が少ないため、真の相対的貧困世帯数 $H_{N=n}^*$ から相対的貧困世帯数 $H_{N=n}$ に変更し便宜上 $H_{N=4} = \sum_{n=4}^{15} H_{N=n}^*$ としている。これは $|B_n|$ も同様である。また、相対的貧困の基準は、1 人世帯は 120,000 円、2 人は 170,000 円、3 人は 210,000 円、4 人では 210,000 円、5 人では 250,000 円のようにしている。

続いて失業や収入減少があった世帯の収入を減少分だけ保証することを目的とする f_2 は、 H をすべての世帯の集合として、世帯 h の世帯収入 I_h を用いて以下の式で表される。

$$f_{2k}(z_k) = \frac{\sum_{h \in H} |I_h - I_h^*|}{\sum_{h \in H} I_h}$$

これらの経済指標を用いて、最終的に多目的最適

化に利用する経済指標 F_1, F_2 を以下のようにして算出する。

$$F_m = \frac{\sum_{k=1}^2 f_{mk}(z_k)}{2}, (m = 1, 2)$$

多目的最適化では、この 2 つの経済指標が最小化するように学習を行っていくことになる。

4 適用結果

4.1 多目的最適化社会シミュレーション

本研究では、関連研究に習い、NSGA-II¹⁰⁾ を利用した 2 目的最適化を行った。施策を表す決定変数は、給付金の支給条件と金額を 2 進数バイナリ表現した 53bit の長さにてコーディングを行った。

予備実験を通じた確認を踏まえ、シミュレーションでは全ての状況について共通した設定を用い、最適化の遺伝的アルゴリズムは個体数 72、世代数 100、一様交差(交差確率 0.9)とした。また最適化に使用する経済指標は、ランダム性の影響を考え、乱数を変えて 2 回実行した際の平均とした。

以上のようにして、自治体ごとに解集合を求める計算を 3 回行った。この 3 つの解集合のうち、最も探索が進んだものをクラスター分析に利用する。

4.2 クラスタ分析

続いて、python にて SciPy から提供されている階層型クラスタリングを利用しクラスター分析を行った。この際、クラスター間の距離計算には群平均法を利用した。その結果を Fig. 2 に示す。

ここでは複数の自治体のパレート解集合を 1 つのグラフに表示している。各点が解を、縦軸と横軸が解の目的関数値を、各点の色が類型を表している。そして、左上のグラフにはすべての類型の解集合を、それ以外には類型ごとの解集合を表示している。また、グラフ上部のタイトルでは類型の番号と、括弧の中にクラスター内の自治体数を表している。

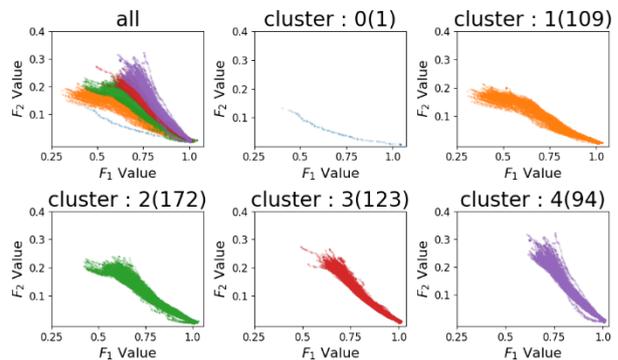


Fig. 2 クラスタ分析結果

このクラスター分析結果を見ると右下の F_1 が最大かつ F_2 が最小な点はほぼ全ての類型で共通しているため、理想的な点 $(0,0)$ との近さは、そのクラスターの解集合の傾きによって決まると言っても良い。ここから、自治体の分類は主にパレート解集合の傾きによって形成されていることが確認できた。この傾きはトレードオフ関係の強さによって決まる。トレードオフ関係とは目的関数の両立の難しさを表しており、強いと両立が難しく、傾きが急になる。今回の事例のトレードオフ関係は、貧困世帯の救済と

給与減少世帯の救済の両立の難しさを示している。

今回は、トレードオフ関係の弱いクラスタから cluster0,1,2,3,4 と類型名を命名した。

4.2.1 クラスタ分析の正当性確認

4.2.2 クラスタの分布

機械学習にかける際、クラスタごとのデータ数に極端な差があると、望み通りの予測が行えなくなる可能性がある。そのため、クラスタごとの自治体数を確認した。その結果は、cluster0 から順に 1, 109, 172, 123, 94 となった。cluster0 を実験から除けば、自治体数に大きな差はなくなり、問題なく機械学習が実行できると予想される。

4.3 情報の抽出

自治体の状態を表す情報の抽出は、Table 1 にある属性値などを使って作成した。

まず、安易に作成できる情報として給付金施策にも使用した下部 5 つの属性についての値ごとの人口比率または世帯比率と、自治体の総人口、総世帯数、平均年収を算出した。続いて問題状況に関わりのありそうな情報を作成したため Table 2 に示す。ここで書かれている相対的貧困世帯比率は、相対的貧困世帯数を総世帯数で割った数値である。

weak_worker や、poverty_fam などは、貧困世帯が目的関数に関わるため作成した。また、世代別の無職比率は、事前の調査で、貧困世帯を構成する住民の多くが無職であり、特に高齢者が多いことがわかってきたため、世代別に集計することで精度が向上するのでは無いかと考えて作成した。

なお、世帯人数別の相対的貧困世帯比率では、目的関数 1 を求めた時同様 4 人以上の貧困世帯数は 4 人の変数に計上した。

これらの情報を決定木の説明変数として利用する。

Table 2 問題状況に即した情報

変数名	詳細
weak_worker	短時間労働者と臨時労働者の人口比率の和
poverty_fam	相対的貧困世帯比率
poverty_fam2, 3, 4	世帯人数別の相対的貧困世帯比率
worker_age_unemp	生産年齢人口中の無職の人口比率
children_unemp	15 歳以下の無職の人口比率
senior_unemp	65 歳以上の無職の人口比率

4.4 決定木分析

本事例に対して提案する分析手法を適用することで、コロナ禍における給付金政策を実施する過程において、目的関数空間上におけるパレート解集合の形状を決める要因を類推することができる。ここで発見できる要因は自治体のパレート解の形状を直接決定づける物かは不明だが、全体の傾向をつかむ上では十分に有効であると考えられる。

4.4.1 決定木の作成

決定木には被説明変数として自治体の類型名を、説明変数として 4.3 で抽出した情報を使用する。続いてパラメータの調整を行う。まず評価指標である macro-F₁ 値とノード数について、パラメータ alpha を変えた際の変化を調査した。その結果をグラフにしたものがエラー! 参照元が見つかりません。である。

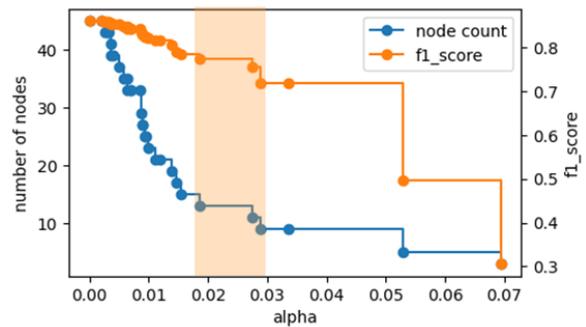


Fig. 3 alpha と評価指標の関係

ここから、alpha が 0.018 よりも小さくなると、急激にノード数の増加が始まるが、決定木の macro-F₁ 値は徐々にしか増加しないことがわかる。また、alpha が 0.03 以下の場合 macro-F₁ 値がかなり低く、ノード数をわずかに増やすだけでも macro-F₁ 値が大きく増加することがわかる。そのため色の塗った領域である $0.018 \leq \alpha \leq 0.02$ が精度と複雑さのバランスが取れた領域といえる。

続いて $\alpha \cong 0.0186$ の決定木を Fig. 4 に示す。この図は決定木の分類過程を表現しており、上から下に向かって分類が進む。各ノードにかかっている内容は、1 行目左に分割する変数名、右にしいい値を、2 行目はノードに含まれる個体数、3 行目はクラス番号ごとの個体数、4 行目にそのノードのクラスを記述しており、葉のノードは 1 行目がなく、4 行

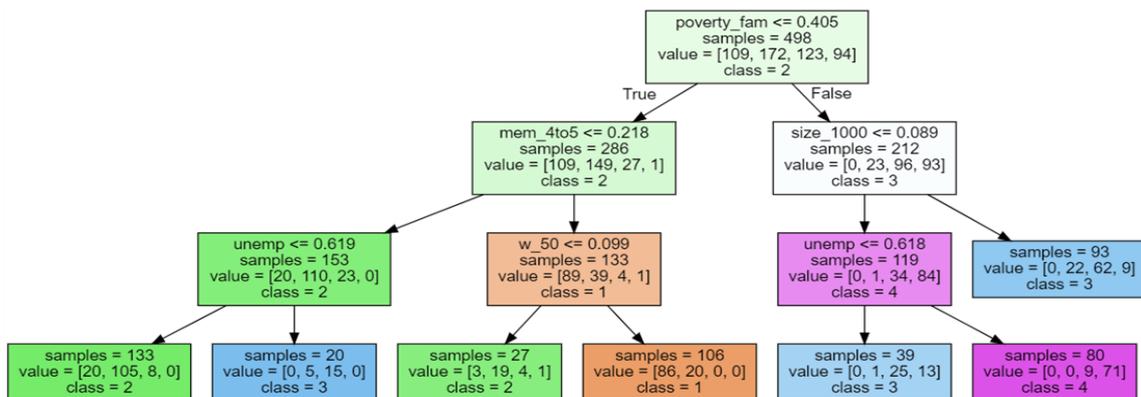


Fig. 4 作成した決定木

目は分類結果を示している。そしてノードの右下にあるノードは上のノードで変数の値がしきい値以下の個体が、右下にはしきい値より多い個体が含まれている。

4.4.2 結果

Fig. 4 の決定木の精度は macro-F₁ 値が 0.773 で、混同行列が Table 3 のようになった。混同行列とは、正解クラスと予測クラスを使って予測結果を表現する手法で、列が予測したクラスを、行が真のクラスを示しており、該当する個体がいくつ存在するかを書いている。また対角成分は正解した個体数である。

Table 3 混同行列

クラス	1	2	3	4
1	86	23	0	0
2	20	124	22	6
3	0	12	105	6
4	0	1	25	68

この macro-F₁ 値は、学習データに対する予測にしては高くないが、混同行列から、誤った予測は隣接したクラスに集中していることがわかる。また、目的関数空間上での解集合を確認したところ、予測に失敗した自治体はクラスタの境界に位置する自治体が多かった。これらの理由から解集合のトレードオフ関係の傾向は十分に再現できていると考えられる。

4.4.3 ルールの抽出

この決定木から抽出できるルールは以下の通り。

- 貧困世帯比率(poverty_fam)が 0.405 以上の場合はトレードオフ関係が強くなる。
- 4~5 人世帯比率(mem_4-5)が少ないとトレードオフ関係が強くなる。
- 従業員数 1000 人以上規模の企業に務める労働者人口(size_1000)が少ないと 8.9%以下の場合トレードオフ関係が強くなる。
- 無職比率(unemp)が 6 割以上の場合トレードオフ関係が強くなる。

世帯人数や企業規模に関するルールは因果関係が不明瞭だが、残りについてはシミュレーションの観点から納得できる説明であるように考えられる。貧困世帯が少ないと、そこまで費用をかけずに経済指標 F₁ を改善することができるため、トレードオフ関係が弱くなり、逆に貧困世帯が多いと経済指標 F₁ を改善するために必要な出費が多くなり、トレードオフ関係が強くなると予想できる。また、無職が多いとトレードオフ関係が強くなるというのも、貧困世帯救済に関係するためであると推測できる。その理由は実験に使用した仮想合成人口個票データでは、貧困世帯における無職の比率が高く、無職の多い家庭は貧困世帯になりやすい傾向があるためである。

4.4.4 予測に失敗した特殊な自治体

分類を間違えた自治体には、どの自治体とも形状が似ていない自治体も含まれており、自治体同士の距離を確認したところ、どの自治体とも距離が遠く、非常に孤立した自治体群であることがわかった。このことから、クラスタリングの段階で、孤立した自治体は除外したほうが良いと考えられる。

今回の特殊な自治体の半分以上は福島県沿岸部の自治体だった為、福島原発事故の影響が予想される。

4.5 ヒートマップによる可視化・分析

多目的最適化社会シミュレーションの事例では、決定変数空間の解集合を確認することで、問題状況に対する理解を深めることができる。

本研究では、ヒートマップによる可視化手法を適用するとともに、その解釈手順についても説明する。

4.5.1 決定変数

描画する解 d_{in} は、53bit 長の遺伝子であり、 j 番目の遺伝子座の値 g_{inj} は $\{0,1\}$ の内どちらかの値を取る。決定変数の遺伝子構成は家族類型 fam(9bit)、世帯内役割 role(8bit)、産業分類 ind(22bit)、雇用形態 emp(4bit)、企業規模 size(5bit)、支給金額 pay(5bit) である。各遺伝子は、Table 1 の属性の値と対応しており、例えば家族類型：単独世帯[0]の遺伝子が 1 となっていた場合、支給対象に単独世帯[0]が含まれることを意味している。また、Table 1 には書いていないが、決定変数 p_i には産業分類で無職と失業者を表す遺伝子と、企業規模で無職を表す遺伝子が追加されている。

この変数の内、支給金額以外は、遺伝的アルゴリズムの遺伝子の値をそのままヒートマップに利用するが、支給金額は、5bit の 2 進数表記となっているものを 10 進数に変換してヒートマップに利用する。

なお本適用事例では多目的最適化時の制約により値の変動が非常に少なく、状況 c_i によっては値が 1 度も変化しない遺伝子が存在する。そのような遺伝子については表示せずとも分析に大きく影響しない。そのため、情報量削減の一環で、ヒートマップには表示しないこととし、値が変わらない変数名のみを別途記載することとした。

4.5.2 描画

ヒートマップで描画するのは、先程説明した決定変数と、目的関数値である。色に関しては値が小さいほど薄い色を、大きいほど濃い色に設定した。目的関数値と支給金額は、値の範囲が一定ではないため、解集合 D_i の同じ遺伝子に対して正規化を行い、値が $[0,1]$ の範囲に含まれるように変換を行った。なお、目的関数は最小化するのが目標なので、薄い色の部分が優れた値ということになる。

4.5.3 クラスタリング

本手法では、遺伝子座同士について階層的クラスタリングを行い、距離の近い遺伝子座が近くに配置されるように並べ替えを行った。その際の距離計算では、全個体の同じ遺伝子座を配列とみなし、配列同士のユークリッド距離を用いた。

4.5.4 ヒートマップ

遺伝子のヒートマップの例を Fig. 5 に示す。

この図から、遺伝子にはそれぞれ傾向があり、およそ 4 種類の遺伝子が存在することがわかった。

- 目的関数 1 に類似した遺伝子
- 目的関数 2 に類似した遺伝子
- ランダムに値が切り替わる遺伝子
- 常に値が変わらない遺伝子(表示していない)

なお、目的関数に類似するとは、目的関数値が悪いときに支給対象に含むということである。

目的関数 1 に類似する遺伝子には世帯内役割を示

す role30,31 がある。これらが目的関数 1 に類似する理由は不明だが、彼らへの給付で貧困世帯の増加を抑えながらの失業世帯への給付が可能なのであろう。

また、目的関数 2 に類似した変数は数多くあり、給付過多を抑えながら貧困世帯に給付を行うにはこれらの給付が有効であると考えられる。産業分類では ind40, 160, 80, 50, 130, 90, 140 があり、貧困世帯にはこれらの産業の労働者が多い可能性がある。

ランダムに値が切り替わる遺伝子には、産業分類だと ind30, 110, 170, 20, 60, 120, 140, 70, 世帯内役割だと role30, 31 がある。これらは該当する人間が 1~2% 以下と少なく、変更しても目的関数に大きな影響を与えないことから、細かく変更することで最適化時にパレート解を増やしていると考えられる。

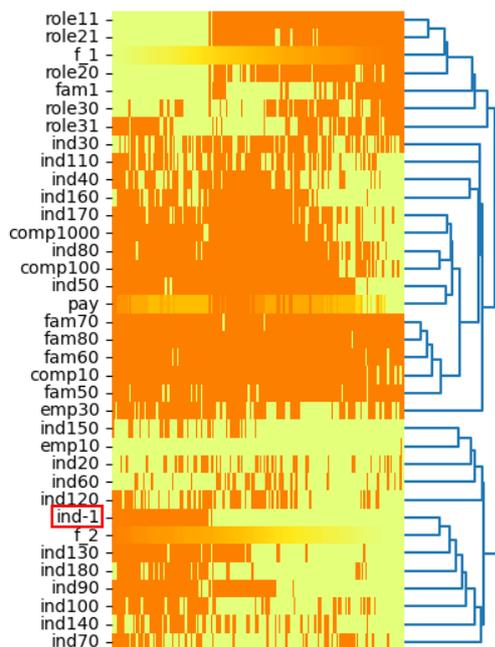


Fig. 5 山梨県中央市のヒートマップ

4.5.5 その他の知見

決定木による分析から解集合に無職比率が関係している事が判明している。そこで Fig. 5 にて赤枠にて囲った ind-1(無職)の値を観察すると、その値が 1 を取る左側 3 割と、右側 7 割で傾向解の傾向が大きく異なることがわかる。目的関数空間上の解集合でも、ind-1(無職)の値が変わる部分で、解集合の曲線が曲がるような変化が見られ、Fig. 2 の cluster2 を見ると、その傾向を確認しやすい。

これは貧困世帯に無職が多いことから、無職に給付を行う場合は目的関数 2 が非常に悪くなる代わりに目的関数 1 が大幅に向上するためと推察できる。

この分割は殆どの自治体で確認でき、この問題状況の特徴の一つであると考えられる。

5 考察

5.1 本手法の意義

今回提案したクラスター分析と決定木を用いた分析は、多目的最適化社会シミュレーションの分析を行う際に有用であることが適用事例からわかった。

従来は、解集合と関係の深いパラメータを発見する為にシミュレーションのパスや解集合の中身を細

かく観察する必要があったが、本研究では関係が予想される変数を多数用意し、決定木の説明変数として組み込むという容易な方法で、特徴的な変数を見出すことができた。適用事例では解空間の形状には貧困世帯や無職の比率が強く影響し、産業構造やその他の要因は影響が少ないことが判明した。この傾向は、クラスター分析にて分類を細かくしても大きく変わらなかった。しかし、あくまで今回の実験でそのような結果になったに過ぎず、目的関数や目的関数空間上での解集合の距離計算法を変更することで異なるクラスタを作り出し、異なる因子を発見することもできると考えられる。このようなフィードバックを行うことで、シミュレーションの妥当性や、実行承認性を高めることを期待できる点が本手法の利点の一つであると考えられる。

続いて決定変数空間の描画では、ヒートマップという情報密度の高い可視化手法を利用することで目的関数値と設計変数の関係を理解することができた。この分析を通して、設計変数がどちらの目的関数値に関わっているのか、また目的関数値との関係が少ないのかを判別することができた。

更に本適用事例では、決定木によって解集合に対する大まかな理解を得てからヒートマップを分析することで、解集合において特徴的な変化を発見でき、パレート解集合が大きく分けて 2 つの部分に分かれていることが確認できた。そこから、解集合を 2 つに分けて、より細かく効率的な分析を行うことが可能になると考えられる。これは 2 つの手法を組み合わせることで問題状況に対する新たな知見が得られたことを意味しており、本研究での提案手法の有用性を特に表す事例である。

5.2 本手法の限界

本手法は解集合の分析に状況変数を使っているため、適用範囲については状況変数が明確に記述された多目的最適化社会シミュレーションにのみ使用することが出来る。状況変数が明確に記述されているというのは、例えばエージェント一体一体の属性値が全て記述されている事や、エージェント作成時の属性の出現頻度が記述されている事を指す。この点さえ満たせば多くのシミュレーションに適用出来ると考えられるため、適用範囲は広いと予想される。しかし、本研究では一つの適用事例についてのみ研究を行っているため、本当に多くの事例に適用できるのかは未知数である。そこで、先程の適用範囲について問題なく適用することが出来るのか、他の事例についての分析の有効性はどの程度なのかを確かめる必要がある。

また、本手法はいくつかの問題も抱えており、一つは、クラスター分析結果の解釈についてである。本研究では、類型ごとの解集合の形状については目視で確認したのみである。この方法では、3 つ以上の目的関数を持つ最適化における解集合について分析することができないほか、形状の特徴をつかむ操作も分析者の直感に基づいており、汎用的な手法であるとは言い難い。そのため、解集合の形状を解釈する汎用的な手法を設計することも今後必要であるように考えられる。もう一つの問題は、本手法の適用前に、状況についての簡易的な特徴分析を行って

いない点である。本研究の適用事例では、500 近い市について最適化と分析を行ったが、最適化を行う前に各自治体についての集計などは行っておらず、どのような自治体が存在するのかわかっていない。その理由としては、500 近い自治体の内容は千差万別であり、どのような分特徴分析が効果的なのが不明だったためである。このような問題は、本研究の適用事例以外の事例でも見られると予想されるため、効果的な予備集計の手法を確立することには意義があると考えられる。

6 おわりに

本研究では、多目的最適化社会シミュレーションの解集合の分析手法として、クラスター分析と決定木を使った目的関数空間の解の形状分析と、ヒートマップを用いた決定変数空間の描画手法を提案した。

そして全国の499の市に、経済ショックに対する給付金政策の最適化を行い、本手法を適用した。

その結果、多目的最適化社会シミュレーションの解集合について、いくつかの知見を得ることができ、これが本手法の有効性を示していると考えられる。目的関数空間に対する分析では、解空間の形状と自治体の状態の関係を分析することができ、決定変数空間の分析では、施策内容と目的関数の関係と解集合の詳細な情報を知ることができた。

謝辞

本研究は JST 未来社会創造事業 JPMJMI20B3, 科研費 22K04586 の支援を受けたものである。

参考文献

- 1) 増井紀貞：機械学習の解釈性，人工知能，**33-5**，641/644 (2018)
- 2) 内閣府：人工知能技術戦略実行計画(案)
<https://www8.cao.go.jp/cstp/tyousakai/jinkochino/7kai/siryos.pdf>
- 3) 後藤，森田，白井，市川：多目的最適化社会シミュレーションにおける解集合に基づく状況分析手法の検討，第 27 回社会システム部会研究会，1/9 (2022)
- 4) 後藤，高橋：組織行動の持つ不確実性が行政評価制度の効果に与える影響の分析，経営情報学会誌，**18-2**，139/166 (2009)
- 5) M. T. Ribeiro, S. Singh, C. Guestrin: "Why Should I Trust You?" Explaining the Predictions of Any Classifier, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135/1144 (2016)
- 6) D. Houtao: Interpreting Tree Ensembles with inTrees, *International Journal of Data Science and Analytics*, **7-4**, 277/287(2019)
- 7) A. Pryke, S. Mostaghim, A. Nazemi: Heatmap Visualization of Population Based Multi Objective Algorithms, *Second International Conference on Evolutionary Multi-Criterion Optimization*, 361/375 (2007)
- 8) <https://scikit-learn.org/stable/modules/tree.html>
- 9) 原田，村田，柁井：家族類型と世帯内の役割を考慮した SA 法による大規模世帯の合成，計測自動制御学会論文集，**54-9**，705/717 (2018)

- 10) K. Deb, A. Pratap, S. Agarwal and T. Meyarivan: A Fast Elitist Multi-Objective Genetic Algorithm: NSGA-II, *IEEE Transactions on Evolutionary Computation*, **6-2**, 182/197 (2002)