機械学習アルゴリズムによる 耕作放棄の要因分析および予測モデルの開発

松井 孝典¹・宇賀田 徹²・町村 尚³

¹正会員 大阪大学助教 大学院工学研究科環境・エネルギー工学専攻

(〒565-0871 吹田市山田丘2-1)

E-mail: matsui @ see.eng.osaka-u.ac.jp

²非会員 大阪大学大学院生 大学院工学研究科環境・エネルギー工学専攻

³非会員 大阪大学准教授 大学院工学研究科環境・エネルギー工学専攻

日本では近年、耕作放棄地が増加傾向にあり、それに伴って農地での自然共生システムが崩れ、生物多様性や農地の持つ生態系サービスの低下などの様々な問題を引き起こしている。この問題に対して、耕作放棄が生じる要因を特定し、将来的な耕作放棄を予測することで、営農を持続するための農業計画の策定を支援する必要がある。このため、様々な地域や集落を対象とした要因の抽出や将来予測が行われてきているが、地域ごとに耕作放棄の特性が異なることから、すべてを説明する汎用モデルの構築が困難という問題が生じている。そこで本研究では、様々な地域に対して適用可能な、その地域固有の状況に応じた耕作放棄要因を特定し、かつ将来の耕作放棄量を予測するためのモデル生成プロセスを開発することを目的とする。モデル構造の特定には、要因の特定と予測が同時に可能な機械学習アルゴリズムである一般化線形モデル、ランダムフォレスト法、多変量適応型回帰スプライン法を適用し、モデル生成プロセスの最適化を図った。その結果、ランダムフォレスト法が最も精度よく耕作放棄量の予測とその要因の特定を行うことを示し、耕作放棄が進む地域や進展が遅い地域など様々な地域に適用が可能であることを示した。

Key Words : *abandoned agricultural land, machine learning, generalized linear models, random forest, multivariate adaptive regression splines*

1. はじめに

日本では農業従事者の高齢化や労働力不足、農作物の価格低迷による収益性の低下などを背景として、耕作放棄地が 1985 年以降増加し続けている¹⁾。しかし、農林水産省生物多様性戦略²⁾でも指摘されているように、農地は多くの生物にとって貴重な生育・生息環境を提供しており、それぞれ特有の生態系を形成・維持することで生物多様性の醸成に大きな役割を果たしてきた。同時に、農業の衰退はイノシシやシカ等による獣害の増加³⁾⁴⁾といった、生物から人間への直接的な被害も増大させることが知られている。このように、耕作放棄の進行は農業および農村の持つ多面的機能⁵⁾の低下を招くため、生態系サービスの供給能力を減少させ、自然共生システム自体の持続性に関わる諸問題を引き起こしている。

そのため、このような耕作放棄を抑制するためには、どのような要因が耕作放棄を促進するのかを特定した上で、将来の耕作放棄を予測しながら、優先度の高い対策を講じる必要がある。これを支援するべく、これまでも耕作放棄要因を分析する研究として、農家へのアンケ

ートや実地踏査によって耕作放棄の動機や傾向を分析した研究⁶⁾⁷⁾⁸⁾、特定集落の空間構造特性を分析することで耕作放棄を誘発する地理条件を分析した研究⁹⁾、農林業センサス¹⁰⁾等の公開されている統計情報に対して多変量解析¹¹⁾¹²⁾¹³⁾¹⁴⁾¹⁵⁾を行うことで耕作放棄の要因を分析した研究などがある。しかしながら、一般的に耕作放棄の要因は地域が固有にもつ農業の事情によって異なるため、すべての地域に適用可能なモデルで一元的に要因を説明することは困難を伴う。加えて、成長戦略の中で大規模な農業改革が宣言されたが、こうした改革は農業のもつ構造自体を大きく変化させるため、農業システムの非連続な移行も含めて精度を担保できる普遍的なモデルの生成はほぼ不可能であるといえる。一方で将来予測については、吉田ら¹⁶⁾は耕作放棄の要因分析を行ったうえでシステムダイナミクスにより演繹的に将来予測モデルを構築しているが、このモデル生成のプロセスは高度で緻密であるため、構造の変化や他地域への知識移転の都度にモデルを再生成することには大きな労力を伴う。

そこで、本研究では、耕作放棄対策の支援情報を提供するため、任意の地域で耕作放棄要因の特定と将来予測

が可能なモデルを出力できるモデル生成プロセス自体を開発することを目的とする。

2 モデル生成プロセスの構築方法

一般にモデル生成のプロセスは、(1) モデル生成用データの選定、(2) 応答変数および説明変数の設計、(3) モデル構造の選定、(4) モデルチューニングとバリデーションの順に行われる。本研究では、これらのプロセスを簡易な演算手順として定式化することで、そのプロセスに従えば様々な地域固有の状況に応じた耕作放棄要因の特定と耕作放棄量の精度よい将来予測ができるモデルの生成を可能とすることをめざす。このモデル生成プロセスの構築について、まず 2005 年の時点で耕作放棄地割合が 6 % と全国的に平均的な近畿地方において、最も耕作放棄が進展しており早急な対策が求められる奈良県をベースケースとして、上記の(1)から(4)の各サブプロセスについて、番号が対応する節に最適なモデル生成プロセスを探索した過程を示す。その後(5)節で、ベースケース以外の地域にその最適化したモデル生成プロセスを適用可能であるかを評価した方法を述べる。なお、本モデル生成プロセスは行政や民間、農業従事者など様々な農業関連主体に利用されることを想定し、特殊なソフトウェアや過度に高度なオペレーションを要求しないことが必要となるため、以降のデータ加工用の表計算ソフトウェアは広く利用されている Microsoft Excel 2010、モデル生成用ソフトウェアはオンライン上のフリーウェアである統計解析環境 R ver.3.02¹⁹⁾を用いて実行した。R 言語は、入手が容易な上、高度な解析性能を持つ様々なパッケージが公開されているスクリプト型の言語であるため、ソースコードさえあればどの計算環境でも同じ内容を実行できる。

(1) モデル生成用データの選定

モデル生成プロセスを簡便に定式化するためには、収集するデータが高コストな調査に依ったり秘匿情報に基づいたりするものではなく、経年的に収集する仕組みがある一般公開された情報であることが望ましい。そこで本研究では、モデル生成に用いるデータソースとして農林水産省による「農林業センサス—農業編—」¹⁹⁾を用いることとした。農業センサスとは、我が国農林業の生産構造、就業構造を明らかにするとともに、農山村の実態を総合的に把握し、農林行政の企画・立案・推進のための基礎資料を作成し、提供することを目的に 5 年ごとに行われる調査である。(1) 経営の態様、(2) 農業労働及び林業労働、(3) 耕地及びその他の土地、(4) 家畜及び蚕、(5) 農業用の機械及び施設、(6) 農業生産物、(7) 農作業、(8) 山林、(9) 育林及び素材生産、(10) その他農林業経営体の現況を把握するために必要な事項、といった 10 のカテゴリーからな

る 200 以上の項目から構成される統計表である。1960～2010 年の累年データが完備され、我が国の農業の構造を写像する上で大変貴重な基礎データである。このデータは通常の表計算ソフトウェアで可読なスプレッドシート形式で一般公開されているため、前述のデータの入手容易性基準は十分に満たせると考えられる。ただし、農業センサスは各農家に調査票形式で回答を依頼するものであり、規模や地域の特性により系統誤差を有することが想定され、全ての農地を網羅して検証しているわけではないため、誤差も含むことに十分に注意する必要がある。

ここで、農林業センサスでは、農業の主体は①販売農家、②自給的農家、③土地持ち非農家、④農家以外の農業経営体の 4 つに分類できる。このうち本研究では全体の農地面積のうち 90 % を占める①販売農家を分析対象とした。これは、②の自給的農家と③の土地持ち非農家では調査項目が少なく、④の農家以外の農業経営体は調査項目が少ない上に耕作放棄地面積が 2005 年時点で農業経営体全体の耕作放棄地面積の 1 % に満たないためである。

モデリングの空間解像度は、研究の最終段階では個々の農家スケールで耕作放棄の誘発要因を検知して、テーラーメード的に早期ケアを支援するためのモデル生成プロセスの開発をめざすが、現段階ではまずモデル生成プロセス自体を構造化することと主目的とするため、地域スケールでの地理特性をある程度反映し、統計データが安定する市区町村スケールを選定した。ベースケースの奈良県では村スケールを最小単位とする 153 個体となる。

(2) 応答変数および説明変数の設計

次に、このデータソースを基に耕作放棄量を表現する応答変数とその要因候補群となる説明変数の定義を示す。モデルの応答変数は市区町村の面積や農業規模に依存せず耕作放棄の進行度を表す指標として式(1)と(2)から算出される耕作放棄地割合 R を採用した。

$$R_i = \frac{ACL_i}{TCL_i} \quad (1)$$

$$TCL_i = CL_i + ACL_i \quad (2)$$

ここで、 R_i は地域 i の農家の耕作放棄地の割合、 ACL_i は地域 i の耕作放棄地の面積 [ha]、 TCL_i は地域 i の総農地面積 [ha]、 CL_i は地域 i の経営耕地面積 [ha] である。従って、応答変数 R_i は最小 0 から最大 1 までの割合の変数となる。

次に説明変数の設計では、多重共線性の対策として説明変数間の相関の発生を極力避けるため、ある説明変数を他のいくつかの変数の四則演算で求められないように加工した。説明変数も応答変数と同様に地域の面積や農業規模によらない変数とするために標準化するが、その際、各調査項目について、調査項目「主副業別農家数」を例とした表-1 にあるように、主に「経営耕地面積」、

表-1 説明変数の設計の例

第1クラス	経営耕地のある農家数			
	主業農家		準主業農家	
第2クラス		65歳未満が いる		65歳未満が いる
第3クラス		65歳未満の 農業専従者 がいる主業 農家率		65歳未満の 農業専従者 がいる準主業 農家率
変数名	主業農家率		準主業農家率	副業的農家率

または「販売農家数」、「経営耕地のある農家数」などを第1としたクラス分けをし、あるクラスの調査項目を1つ上のクラスの調査項目で除することで説明変数を作成することとした。

最後に、応答変数と説明変数を統合して1次データフレームを作成したのちに、欠損値を含む列ベクトルを削除する処理、およびVIF = 10を基準とした2説明変数間の多重共線性の総当たり判定を行い、多重共線性の疑いがある場合には代表性の高い変数のみを選択したモデリング用のデータフレームを作成した。このデータ加工のプロセスは統計解析で行う一般的な処理であり特殊性な作業が伴うものではないため、例えば表計算ソフトのマクロプログラミングなどによって必要に応じてアルゴリズム化が可能である。

(3) モデル構造の選定

ここからが2節で構築したモデリング用のデータフレームに対して、我が国の農業が持つ構造を抽出してR言語でモデリングをするプロセスとなる。通常、先行研究からの知見や専門家の知識などによって、例えば「耕作放棄が促進されるのは農業経営で最新の技術が実装されていないため、コストや品質等の比較優位が保てず、結果として耕作を放棄せざるを得ないのではないか」といった仮説を生成し、その検証のために回帰モデルでパラメトリックに検証を行い要因を特定し、そのモデルが有意な場合には、生成されたモデルを用いて将来予測を行うことが多い。その作法に従えば、今回の耕作放棄についてのモデルは式(1)のように応答変数は0-1の範囲でスケールされた変数であるため、一般的にロジスティック回帰分析が用いられる。例えば先行研究¹⁹⁾もその方式に従っている。

ここで本研究では「機械学習 (Machine Learning)」の技術を適用したモデル構造化の方略を考える。機械学習とはコンピュータに学習機能を持たせることを目的とした技術であり、近年の急速な計算資源の向上に伴って実利用への応用が盛んな分野である¹⁹⁾。仮説の候補を少数の線形結合子で表現してデータへのフィットの寄与を分析するというパラメトリック回帰のアプローチとは異なり、機械学習ではデータが持つ潜在的なモデル構造自体をアルゴリズムで帰納的に探索・識別し、要因の特定や将来予測を行うノンパラメトリック回帰手法である。

これまでに機械学習アルゴリズムには様々なものが

提案されているが、本研究で用いる機械学習アルゴリズムの選定基準は、①応答に対する説明変数の回帰構造の探索と評価ができる、②応答に関する予測精度の高いモデルの構築ができるという2点である¹⁹⁾。教師あり型のデータフレームに対する機械学習アルゴリズムには、代表的なものとしてSupport Vector Machine²⁰⁾やNeural Network²⁰⁾などが挙げられる。しかしこれらの手法は②の予測・識別精度の観点からは相当の性能を有する反面、カーネル法による高次元写像や多重の隠れ層の設定によってモデル設計を行うために①の要因の特定が明示的になされず、本研究の目的には適さない。

そこで本研究では、種々の機械学習アルゴリズムのうち、①と②を担保できる以下の3つをモデル構造の候補に選定した。候補の1番目は、一般化線形モデル (GLM: General Linear Model)²¹⁾である。GLMは応答変数、説明変数の質・量を問わず、リンク関数や誤差分布の形状をスイッチとして線形結合子による回帰を一般化したアルゴリズムである。特に誤差構造にbinomial, link関数にlogitが選定された場合には前述のロジスティック回帰モデルと等価になる。ここではノンパラメトリック回帰のアプローチでモデル構造ごと帰納推論するため、AICを目的関数としたstepwise法によるベストモデルの自動選定を行うこととした。Rのパッケージはglmを用いた。

第2のモデル候補は、多変量適応型回帰スプライン法 (MARS: Multivariate Adaptive Regression Splines)²²⁾である。回帰樹木 (CART: Classification And Regression Trees)²³⁾を耕作放棄の要因特定に利用した例もあるが、一般にCARTは精度が低いことが知られている。MARSではCARTのステップ関数近似を打ち切りベキ乗基底関数に変更し局所的に線形モデルをあてはめるものことで予測精度を高めたものである²⁴⁾。MARSのモデル式の構造は式(3)と(4)である。

$$h(P^D) = \beta_0 + \sum_{i=1}^I \beta_i H_i(P^D) \quad (3)$$

$$H_i(P^D) = \prod_{j=1}^{J_i} [s_{(j,i)}(p_{(j,i)} - c_{(j,i)})]^+ \quad (4)$$

この時、 P^D はデータフレームD内の説明変数、 h は説明変数 P^D を入力した際の最終的な予測値、 β_i は回帰係数、 I はモデル内の基底関数の総数、 H_i は*i*番目の基底関数、 J は基底関数 H_j の積因子の総数、 $s_{(j,i)}$ は基底関数 H_j の*j*番目の積因子の符号、 $p_{(j,i)}$ は既定関数 H_j の*j*番目の積因子に対応する説明変数、 $l_{(j,i)}$ は $p_{(j,i)}$ を指定するラベル、 $c_{(j,i)}$ は基底関数 H_j の*j*番目の積因子の分割点、 $[a]^+$ はMAX(a , 0)を表す。特に、式(4)にわかるように、GLMと比較して説明変数間の交互作用を樹木構造でティッピングポイントも含めて明示的に表現できるところに強みがある。なお、Rのパッケージにはearthを用いた。

最後に、第3の候補としてランダムフォレスト (RF: Random Forests)²⁹⁾を選定した。RFはCART型樹木を弱学習器としたアンサンブル学習型の機械学習アルゴリズムである。RFのアルゴリズムは以下と式(5)の通りである。

1. 2節で作成された P^D 個の説明変数を持つ N 個のサンプルからなるデータフレーム D に対して、ブートストラップにより B 組のサブサンプルを抽出してサブデータフレーム D^B を構築する
2. サブデータフレーム D^B ごとに回帰木 T^B を構築する。この際、回帰木を構築するための説明変数集合 P^B もランダムサンプリングによって決定する。ただし、 $P^B \leq P^D$ である
3. 生成された N 本の回帰木 T^B から得る予測値 h^B の期待値を最終的な予測値 h とする

$$h(P^D) = \sum_{b=1}^B h^B(P^B)/B \quad (5)$$

このように、RFはブートストラップに基づくため、外れ値に対して比較的頑健であり、環境学研究で扱うような非統制環境下の多ノイズ情報処理には有利と考えられる。またGLMやMARSと比較して多重共線性を回避できる利点もある。RのパッケージはrandomForestを用いた。

(4) モデルチューニングとバリデーション

そしてGLM、MARS、RFのそれぞれのモデル構造の候補について、2005年と2010年の農林業センサスから構成した2時点のデータフレームを用いてモデルチューニングとバリデーションを行い、その結果から耕作放棄の要因特定と耕作放棄量の予測を可能とするモデル生成プロセス用に最適な機械学習アルゴリズムを決定した。具体的には、ベースケースである奈良県のデータフレームに対して、以下のステップを行った。

- ステップ1:** GLM、MARS、RFの3つの機械学習アルゴリズムを2005年のデータフレームでアルゴリズムごとのパラメータのチューニングによりそれぞれ最適学習させる
- ステップ2:** 最適化後の3つのモデルに対して2010年のデータフレームの説明変数を入力し、2010年の耕作放棄地割合の予測値を出力する
- ステップ3:** その予測値と2010年のデータフレーム内の実測値に対して、相関係数、平均絶対残差によりモデルの性能を比較し、最適な機械学習アルゴリズムを選定する

ステップ1のチューニングについて、GLMでは応答変数が割合であることから応答変数の誤差の構造は二項分布(binomial)に固定した上で、link関数の候補としてlogit、probit、

complementary log-log、cauchitの4種のスイッチをチューニングパラメータとした。ベストモデルの探索方法はAICを目的関数とした変数増加型のstepwise法とした。このとき、2節のデータフレームの設計段階では2変数間の多重共線性のみを処理したことから、GLMにより多変量化した場合の説明変数の組み合わせでは多重共線性が生じている可能性がある。そこでstepwise法で選択されたモデル内の各説明変数それぞれについて、その他のモデル内変数を重回帰分析した際のVIFが10を超えた場合には、応答変数との相関係数を判断基準に変数を削除して再度ベストモデルの選択を行った。次にMARSのチューニングパラメータとして、式(4)、(5)で示す交互作用の最大次数 $J_{\max}$ 、前進ステップワイズ過程における基底関数の最大数 $L_{\max}$ を設定し、 $J_{\max}$ を1~5次、基底関数の最大数 $L_{\max}$ を10から100まで10ずつ変更させた場合のそれぞれのモデルを構築し、適合度を表すGCV (Generalized Cross-Validation) 基準で寄与率および選択された基底関数の数を比較することでチューニングした。なおGCVは以下の式(6)であらわされる。GCVの正則化項である ξ は、交互作用による過学習を回避するため、交互作用のみのときは2、それ以外は推奨値である3とした²⁹⁾。

$$GCV(I) = \frac{\sum_{n=1}^N (h_n - h(P^D)_n)^2 / N}{\{1 - I(\xi + 1)/N\}^2} \quad (6)$$

最後に、RFではブートストラップサンプリングによって発生させる樹木を500本、樹木1本に用いる説明変数の数を回帰モデルの推奨値である全説明変数の1/3、樹木生成の前進過程における停止条件は各終結ノードのサンプル数を5以下と固定した²⁹⁾。ここでRFはデータのノイズに対して頑健であるため、予測精度だけを目的とすれば全ての説明変数を用いてモデル生成を行っても構わない。しかし将来予測に利用するという観点では膨大な変数で構築されたモデルでは、シナリオ分析の際に説明変数の変動を表現する段階で操作性に困難が伴う。そこで、将来予測精度の向上に寄与している説明変数のみでモデルを構築するために、RFでモデル生成した際のGini係数の期待減少量を重要度基準として、重要度の降順の変数増加法で2005年に対する予測精度の包絡線を出力してスクリー基準から説明変数の採用数をチューニングした。MARSではGLMと同様に応答変数の誤差構造は二項分布(binomial)に従うとし、リンク関数は先行したGLMで最適とされたリンク関数を用いた。最後にステップ2、3では、ステップ1を通じてチューニングされた3つの最適モデルに対して、相関係数、平均絶対残差によって予測精度の評価を行い、最適な機械学習アルゴリズムを選定した。

(5) モデル生成プロセスの評価

これまでの1~4節では、奈良県をベースケースとして

最適なモデル生成プロセスを決定したが、1章で述べたように、モデル生成プロセスが満たさなければならない要求仕様は以下の2点である。

要求仕様1: 様々な地域に適用した際に、地域の状況に応じた耕作放棄の要因が特定されること

要求仕様2: その地域の状況に応じた要因を基に生成されたモデルで将来の耕作放棄量の予測の精度が高いこと

よって今回は、耕作放棄地割合 0.08 のベースケースである奈良県に加えて、高耕作放棄ケースとして 0.15 と全国で耕作放棄地割合が最上位である長崎県、低耕作放棄ケースとして 0.02 と全国平均 0.04 に比べても進行が遅く北海道の次に耕作放棄地割合が低い滋賀県を選定した。これらのケースに対してベースケースで最適化されたモデル生成プロセスを適用して、上記の要求仕様 1, 2 を満たしているかを評価した。要求仕様 1 は、3 つの地域で共通の要因に加えて地域に固有な要因を抽出できたかを評価した。要求仕様 2 は、4 節のモデルの予測精度の評価方法と同様の方法を用いた。

3. 結果と考察

(1) 生成されたデータフレームとモデルチューニング

まずベースケースである奈良県の 2005 年の農林業センサスに対してデータフレームを生成プロセスを適用した。まず奈良県の市区町村スケールでは欠損値処理を行った段階で 118 の説明変数を持った 130 個体が採用された。続いて 118 変数に対して $VIF = 10$ を基準とした多重共線性処理を行い、112 変数が採用された。以降のモデルチューニングと最適モデルの選定はこのデータフレームを用いて行った。ただし前述のように RF のみ多重共線性に対して頑健であるため 118 変数を全て使用した。

次に、このデータフレームに対して行ったモデルチューニングの結果を示す。まず GLM については、最適なリンク関数の候補である 4 種を用いてそれぞれモデルを構築した結果が表-2 である。この結果から、logit が complementaly log-log でもっとも予測精度が高くなっているのがわかる。しかしながら、耕作放棄の要因としてモデルに選択された説明変数は logit では 13 であるのに対し、complementaly log-log では 15 と多いため、少ない変数で高い予測精度を示した logit を最適なリンク関数と決定した。次に MARS では、GCV 基準によってモデル選択をして基底関数と交互作用の数についてのチューニング結果を表-3 示す。モデルの可読性と操作性を考慮した上で情報量の観点から GLM と差別化するためには、次数は 2 次以上で基底関数の数が少ないほうが望ましい。その意味では、2

表-2 一般化線形モデル(GLM)における
チューニングの結果

リンク関数 (選択された変数)	相関係数	平均誤差	平均絶対残差
logit (13 変数)	0.79	-0.002	0.022
probit(19 変数)	0.83	-0.002	0.021
complementary log-log (15 変数)	0.81	-0.003	0.023
cauchit(15 変数)	0.83	-0.001	0.021

表-3 多変量適応型回帰スプライン法(MARS)における
チューニングの結果

交互作用 (選択された基底関数)	相関係数	平均誤差	平均絶対誤差
1 次 (24 個)	0.69	0.000	0.027
2 次 (16 個)	0.61	0.004	0.030
3 次 (21 個)	0.73	0.005	0.025

図-1 ランダムフォレスト法(RF)におけるチューニングの結果

次と 3 次の交互作用モデルが候補となるが、相関係数や誤差ではそれほど大きな差は見られない。よって基底関数を少なくする意味で 2 次の交互作用までを許容する MARS を採用した。そして RF について、チューニングパラメータである投入する説明変数の数と、相関係数および平均絶対誤差の関係を示したものを図-1 に示す。モデル変数が 20 の付近で 2005 年の予測精度が収束していることがわかる。モデル変数が 18 で最も予測精度が高く、採用される変数も少ない。よって重要度が高い変数から順に 18 の説明変数を採用する場合が最適モデルであるとした。

(2) GLM, RF, MARS の予測精度の結果の比較

続いて、1 節で 2005 年のデータフレームに対して予測精度が最適にチューニングされた 3 つのモデルに対して、2010 年のデータフレームの説明変数を入力した際の予測値と実測値の比較を図-2 に示す。まず共変動性を表す相関係数について、GLM, MARS, RF のすべてのモデルについて相関係数は、一般に相関はあるといえる 0.4~0.7 程度に収まり予測値が実測値と同じ挙動を示すため、モデ

図-2 一般化線形モデル(GLM), 多変量適応型回帰スプライン法(MARS), ランダムフォレスト法(RF)の予測精度
(備考: プロットの濃淡はデータの密度を表す)

ルが意味を持つことがわかる。次に、予測値の精度を表す平均絶対誤差では、全体的に0.03程度となり、ベースケースである奈良県の耕作放棄地割合の平均値0.08に対して375%の誤差となり、一定の予測精度はあるがまだ改善の余地があることが示される。また3つのモデル間での性能を比較した場合、相関係数、平均絶対誤差の観点から耕作放棄地割合の予測には重要度の高い18の説明変数から生成したRFが最適であると判断した。しかしながら、RFは $R=0.71$ であることから実測値と類似する共変動を表現できているが、低耕作放棄地割合の領域で過大評価、高耕作放棄地割合で過小評価の傾向があり、これが平均絶対誤差0.03につながる。この結果から、局所域での予測精度の向上のみを目指すならば、例えば耕作放棄地割合0.10付近でデータをセグメンテーションするなど、耕作放棄地の進展度合によって別のモデルを構築する方略もあるが、逆にデータのセグメント間を超える急速な耕作放棄の進展が予測できなくなるなどの問題も生じる。さらに見方を変えれば、機械学習のもつ構造識別能力により、耕作放棄地割合0.10を閾値に耕作放棄を取り巻く状況が大きく変動して急速に促進される可能性のシグナルを帰納的に抽出されたとも解釈できる。

(3) モデル生成プロセスの他地域へ応用可能性の検証 a) 予測精度の検証

続いて、ベースケースである奈良県に対して実施したモデル生成プロセスの最適化の結果、すなわち「2005年のセル生成用データフレームで重要度が高い説明変数をスクリー基準で選択してRFでモデルを構築して2010年の予測値を出力する」というプロセスを、高耕作放棄ケースと低耕作放棄ケースに適用した結果を図-3に示す。

すべてのケースにおいて相関係数は0.7程度となり、予測値と実測値は共変動している。しかし平均絶対残差そのものは0.01~0.046までケースによって異なる。この精度の評価はケースごとの耕作放棄地割合の大きさにも依存する。上述のように耕作放棄地割合の平均値が0.08であるベースケースでは30%、0.15である高耕作放棄ケースでは30.7%の誤差である一方で、0.02である低耕作放棄ケースでは80%の誤差に相当し、中高耕作放棄ケースには予測精度が比較的良好であるが、低耕作放棄ケースに対しては感度が不足していることが示された。また、中低耕作放棄ケースとともに低域で過小評価、高域で過大評価となっている点が共通しており、耕作放棄ケース全域を通じてみた場合、S字状の耕作放棄地割合の曲線になっている

図-3 耕作放棄ケース別の耕作放棄量の予測精度 (備考: プロットの濃淡はデータの密度を表す)

可能性が示唆される。ただし、農業センサスの解説部分にも記載したように、データ自体に系統誤差が内包しているため、フィッティングを過度に期待したり、モデル出力の読み取りを過剰に行うことは避ける必要がある。

b) 地域固有の耕作放棄要因の抽出に関する検証

耕作放棄ケース別に、RF によって特定された耕作放棄要因に対して重要度が高い上位 10 変数を作付作物、農業収入、農地集約度、農地流動性、所有機械、労働力のカテゴリ別に整理したものを表-4 に示す。農林水産省の調査では^{28,29,30}、山間・中間・平地・都市域で共通する主要な耕作放棄要因として、「高齢化・労働力不足」、「地域内の担い手不足」、「農産物の価格低迷や収益性の低下」が挙げられている。これらの要因は、本分析でも全ケースを通じて農業専従者の不足や男性専従者の高齢化、豆や麦、果樹などの作付作物類というより詳細な要因で相対重要度と共に抽出できている。ベースケースと低耕作放棄地ケースの「稲以外の作物だけを作った田のあるの農家率」、高耕作放棄ケースと低耕作放棄ケースの「経営耕地 1ha あたりの世帯人数 [人/ha]」については、ケース間を超えて出現する耕作放棄要因であり、全国的な課題とともに、先行研究の詳細なヒアリングによる要因特定の知見⁹やモデル解析による結果^{7,13,14,30}、ともよく対応する。このことは、本研究でのモデル生成プロセスは大域での耕作放棄要因の動向を捕捉しながら、地域別の詳細な変数として表現できる可能性を示しており、政策上も地域の文脈に即したより優先的に対応のための施策を詳細に立案することを支援する機能を有するといえる。

またケース別にみた場合、高耕作放棄ケースである長崎県は離島、半島や中山間地域を多く有し、狭小な傾斜農地を中心に耕作放棄が進展しているが²⁹、今回の分析でも離島による農地の分散・小規模化に関連する「農地集約度」が最も重要として識別された。また長崎県はびわを名産とするが、これに応答して「果樹栽培の比率」が耕作放棄に影響を与える可能性も示した。これに対して「長崎県耕作放棄地解消 5 ヶ年計画」を通じて、農地の貸し手と借り手のマッチングによる利用集約や新規品目の導入が進められており、こうした具体的な取り組みの適合性がデータ科学からも支持できる点で政策評価にも有用であると考えられる。一方、低耕作放棄ケースの滋賀県は、近江米を代表に兼業農家による水稲を主力とした地域であり、農業機械の普及率が極めて高い特徴を持つ³⁰。これに対する要因の抽出では、作付作物の選択が重要と判定されており、低耕作放棄地帯であっても稲作・畑作などの穀物系を生産、販売する農家には耕作放棄に対しては個別のケアが必要であるという知識が得られる。トラクターの所有や田畑の経営などで要因群が抽出されており、ここでも地域固有の耕作放棄の文脈を推論できていることが伺える。

表-4 耕作放棄ケース別の耕作放棄要因

a) ベースケース（奈良県：耕作放棄地割合 0.08）

カテゴリ	変 数 名	相対重要度
作付作物	経営耕地（畑）のある農家率	100
	経営耕地（畑）の面積率	77
	経営耕地（田）の面積率	31
	販売目的でも類を作付（栽培）した農家率	29
労働力	農業専従者も準農業専従者もない農家率	29
	農業専従者および準農業専従者は女子だけいる農家率	27
作付作物	販売目的で豆類を作付（栽培）した農家率	18
所有機械	コンバインを所有した農家率	15
作付作物	稲以外の作物だけを作った田のある農家率	15
	何も作らなかった田の面積率	14

b) 高耕作放棄ケース（長崎県：耕作放棄地割合 0.15）

カテゴリ	変 数 名	相対重要度
農地集約度	経営耕地を持つ農家 1 戸あたりの経営耕地面積 [ha/戸]	100
労働力	男子の農業従事者に占める 65 歳以上の割合	74
	65 歳未満の農業従事者がいる準主業農家率	67
作付作物	販売目的で果樹を施設栽培した農家率	66
所有機械	トラクターを所有した農家率	65
労働力	男子の農業従事者の平均年齢 [歳]	60
農地流動性	貸付耕地のある農家率	51
労働力	経営耕地 1ha あたりの世帯員数 [人/ha]	50
	世帯員に占める農業経営者の割合	50
	世帯員に占める男子の割合	45

c) 低耕作放棄ケース（滋賀県：耕作放棄地割合 0.02）

カテゴリ	変 数 名	相対重要度
作付作物	稲を作った田の面積率	100
	販売目的で麦類を作付（栽培）した農家率	83
農業収入	農産物を販売した農家率	75
作付作物	稲以外の作物だけを作った田の面積率	66
	稲以外の作物だけを作った田のある農家率	54
所有機械	トラクターを所有した農家 1 戸あたりの所有台数 [台/戸]	46
農業収入	第 1 種兼業農家率	34
所有機械	経営耕地 (田または畑) 1ha あたりのトラクター所有台数 [台/ha]	34
労働力	経営耕地 1ha あたりの世帯員数 [人/ha]	27
	農業就業人口に占める基幹的農業従事者の割合	22

4. まとめと今後の課題

(I) まとめ

本研究では、機械学習アルゴリズムを比較した上で耕作放棄の要因分析と予測を同時に行えるモデルの簡易な構築プロセスを開発することを目的とした。機械学習アルゴリズムである、一般化線形モデル (GLM)、多変量適応型回帰スプライン法 (MARS)、ランダムフォレスト法 (RF) の中からモデル構築プロセスの開発において最適なアルゴリズムを決定するために、奈良県をテストケースとしてモデルチューニングや予測精度の検証を通じて RF を主としたモデル生成プロセスの最適化を行った。そしてそのモデル生成プロセスに従って他地域への適用可能性を評価することで、(I) モデル生成プロセスは中高耕作放棄

ケースでは比較的精度がよいが、低耕作放棄ケースでは改良の余地があること、(2) 地域固有の事情に応じた耕作放棄要因の抽出および共通する要因の特定ができる性能があり、より個別のケースに即した政策の立案や評価に寄与できる可能性があるということを明らかにした。

(2) 今後の課題

本研究では、一般モデルの構築を試みた先行研究³⁹⁾での問題を踏まえてモデル生成プロセス開発を拡張し、一定の成果を得たと考えている。このプロセスを良好化するための今後の課題として以下の点が挙げられる。

a) モデル構造の改良

第一に、本研究においては奈良県における耕作放棄を予測した結果から MARS をモデル構築プロセスに用いるアルゴリズムの候補から除外した。しかし、MARS は説明変数ごとに区分点を設定し、区間ごとの回帰係数を推定しており、さらに交互作用を取り入れることができるため、より詳細に耕作放棄要因の構造を把握することが期待できる。MARS の予測精度が低くなった原因を探り、チューニングを行い精度の向上を図る必要がある。MARS 法において予測精度が低くなった原因の 1 つとしては多重共線性の問題が考えられる。今回 MARS 法における多重共線性の処理は説明変数の設計段階のみで行ってしたが、実際に構築されたモデルでは選択された基底関数の間に多重共線性が発生しており、これが原因で予測精度が低下している可能性がある。MARS 法の多重共線性処理については現在、有効な方法が見つからないため、今後の最新の知見を待つ必要がある。また RF 法における変数選択の方法は変数選択を機械的に行うアルゴリズムにできていないため、このプロセスを利用する人によって RF 法の時のモデル変数の数が異なるという問題もあり、選択方法を自動化するアルゴリズムを考案する必要がある。さらに、RF 法は低耕作放棄ケースのように狭い範囲にモデル生成用サンプルの応答変数が偏って分布していると、複数の樹木により集団学習をして耕作放棄を予測した際、この狭い範囲にある値のノードにサンプルが振り分けられやすくなるため、その影響で応答変数の予測値がその帯域の周辺に集中してしまう問題がある。より正確な予測を行うにはモデル構築用サンプルの応答変数の分布の偏りを補正したモデルを開発する必要がある。

最後に、本研究において、用いるデータを市区町村スケールとしているが、より地域固有の事情が強く影響すると考えられる集落スケールでモデル生成プロセスを適用し、アルゴリズムの改良を図る。

b) データフレームの改良

本稿では 2010 年の耕作放棄面積割合を 2005 年の耕作放棄面積割合を引数にした関数で特定するというデータフレームでモデリングを行った。一方で、これまでの耕作

放棄が進展してきた履歴情報を付加することでより予測の精度が向上する可能性がある。農業センサスは時系列的に充実したデータであるため、2 時点間で全く同じ項目数の数値を入力しなければならないという機械学習モデルの仕様から、年次間での調査項目の変更が大きく課題となるが、例えば 2010 年から 2015 年までの耕作放棄面積割合の変化量が 2005 年から 2010 年までの変化量に回帰できる可能性は十分に期待できる。

c) 完全アルゴリズム化に対するデータ加工の問題

本研究の長期の目標は、全国のすべての市区町村に対して地理情報や機構情報など関連するすべての変数を利用して、帰納的に機械学習によりモデル構造ごと機械学習で推論させることとなる。研究の過程で全都道府県に対してモデル生成プロセスの実行を試みていたが、データベースへのスクレイピングの禁止規約や、一貫性がないファイル名で細分化したファイルセット、不要なセル結合や非表示設定などの理由によって、時間制約上実施が困難であった。特に近年、鯖江市が先進的な事例を示したように³⁹⁾、様々な公開情報を統合して新たな意味を生み出すための Linked Open Data の取り組みが世界的にも進展しており、公共データにおいてもデータの入手可能性や、可読性、加工の容易性は重要性を増していくことはほぼ間違いない。この問題は情報科学分野でも解決策が模索されており³⁹⁾、本稿のような事例を示しながら、データのコーディングの質の向上の必要性を示すことも重要な発展的課題であるといえる。

参考文献

- 1) 農林水産省：耕作放棄地対策の推進、<<http://www.maff.go.jp/j/nousin/tikei/houkiti/>>, 2013 年 8 月参照。
- 2) 農林水産省：農林水産省生物多様性国家戦略、<http://www.maff.go.jp/j/kanbo/kankyo/seisaku/s_senryaku/pdf/senryaku.pdf>, 2013 年 8 月参照。
- 3) 野元加奈, 高橋俊守, 小川澤正昭, 福村一成：栃木県茂木町の水田と畑地におけるイノシシ被害地点と周辺環境特性, 50 巻, 2 号, 129-135, 2010。
- 4) 奥村忠誠, 清水庸, 大政謙次：ニホンジカ (*Cervus nippon*) の分布拡大に影響を与える要因, 環境科学会誌, 22 巻, 6 号, 379-390, 2009。
- 5) 農林水産省：農業の多面的機能, <http://www.maff.go.jp/j/kokusai/kousyo/wto/w_14_kanren/pdf/tamenteki.pdf>, 2013 年 8 月参照。
- 6) 今井健, 奥村彰浩, 柳田洋吉：耕作放棄地の現状と課題-岐阜県可児市 A 地区の事例分析-, 岐阜大学農学部研究報告, 62 巻, 51-55, 1997。
- 7) 服部俊宏, 山路永司：農家条件からみた都市近郊の耕作放棄発生要因, 農村計画学会誌, 16 巻, 4 号, 325-333, 1998。
- 8) 野村亘, 村上暁信, 渡辺達三：都市近郊集団農地における耕作放棄に関する研究, 農村計画論文集, 22 巻, 91-96, 2003。
- 9) 鄭会勲, 淀川智之, 矢沢正士：耕作放棄地を有する農業集落の空間構造的特性-北海道の 2 町村を事例と

- して-, 農村計画論文集, 23 巻, 259-264, 2004.
- 10) 農林水産省: 農林業センサス, <<http://www.maff.go.jp/j/tokei/census/afc/>>, 2013 年 8 月参照.
 - 11) 吉田晋一, 佐藤豊信, 駄田井久: 中国地方を対象とした耕作放棄の要因分析-地域間の相違と要因間の関連に着目して-, 農村計画論文集, 23 巻, 277-282, 2004.
 - 12) 仙田徹志: 耕作放棄地の発生要因に関する計量分析, 農業経営研究, 36 巻, 1 号, 57-62, 1998.
 - 13) 服部俊宏, 山路永司: 都市近郊の耕作放棄地の地域分布とその発生要因, 農業土木学会論文集, 63 巻, 4 号, 513-520, 1995.
 - 14) 九鬼康彰, 高橋強: 数量化理論 I 類を用いた耕作放棄の発生要因分析-都市近郊農地の耕作放棄防止に関する研究 (I) -, 農業土木学会論文集, 65 巻, 5 号, 603-613, 1997.
 - 15) 稲葉弘道: 耕地面積と耕作放棄地の変化の要因分析, 千葉大学経済研究, 20 巻, 4 号, 725-752, 2006.
 - 16) 吉田晋一, 佐藤豊信, 駄田井久: 耕作放棄の要因分析と将来予測-システムダイナミックスを用いて-, 農林業問題研究, 41 巻, 1 号, 2005.
 - 17) GNU project: R language, <http://www.r-project.org/>, (2014.03.13 参照)
 - 18) Machine Learning, Kluwer Academic Publishers, Boston, 1986.
 - 19) Goto, M. & Matsubara, Y.: Evaluation of Ordinary ridge regression, Bulletin of Mathematical Statistics, vol.24, 1 53-173, 1979.
 - 20) Cortes, C.; Vapnik, V.: Support-vector networks. Machine Learning 20 (3), 273, 1995.
 - 21) McCulloch, Warren; Walter Pitts: A Logical Calculus of Ideas Immanent in Nervous Activity, Bulletin of Mathematical Biophysics 5 (4), 115-133, 1943.
 - 22) Nelder, John; Wedderburn, Robert: Generalized Linear Models, Journal of the Royal Statistical Society, Series A, 135 (3), 370-384, 1972.
 - 23) Friedman, J. H.: Multivariate adaptive regression spline (with discussion), Annals of Statistics, Vol.19, No.1, 1-141, 1991.
 - 24) Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J.: Classification and Regression Trees, Wadsworth, 1984.
 - 25) 下川敏雄, 杉本知之, 後藤昌司: 樹木構造接近法, 共立出版, pp. 53-83, 2013.
 - 26) Breiman, L.: Random Forests, Machine Learning, Vol. 45, No.1, 5-32, 2001.
 - 27) 杉本知之, 下川敏雄, 後藤昌司: 樹木構造接近法と最近の発展, 計算機統計学会, 18 巻, 2 号, 123-164, 2005.
 - 28) 農林水産省: 耕作放棄地の現状と課題 -, <www.maff.go.jp/j/nousin/tikei/houkiti/pdf/tebiki01.pdf>, 2014 年 6 月参照.
 - 29) 農林水産省: 経営する農地の拡大・縮小に関する意識・意向調査結果, <<http://www.maff.go.jp/j/finding/mind/pdf/20080226cyosa.pdf>>, 2014 年 2 月参照.
 - 30) 農林水産省: 食料・農業・農村及び水産資源の持続的利用に関する意識・意向調査, <<http://www.maff.go.jp/j/finding/mind/pdf/m230519.pdf>>, 2014 年 2 月参照.
 - 31) 森本健弘: 千葉県市川市柏井町四丁目における不耕作農地の形成と農業経営, 地理学評論, 66 巻, 9 号, 1993.
 - 32) 長崎県: 耕作放棄対策, <http://www.n-nourin.jp/~nouri/kousaku_gaiyo.html>, 2014 年 6 月参照.
 - 33) 滋賀県農業協同組合中央会: 滋賀県内の農業, <<http://www.jas.or.jp/agriculture/index.html>>, 2014 年 6 月参照.
 - 34) 宇賀田徹, 松井孝典, 町村尚: 奈良県における耕作放棄の要因分析および予測モデルの開発, 土木学会第 41 回環境システム研究論文発表会講演集, pp. 39-44, 2013.
 - 35) 鯖江市: データ シティ鯖江, <<http://www.city.sabae.fukui.jp/pageview.html?id=11552>>, 2014 年 6 月参照.
 - 36) 奥村晴彦: 「ネ申 Excel」問題, <<http://oku.edu.mie-u.ac.jp/~okumura/SSS2013.pdf>>, 2014 年 6 月参照.

(2014.07.18 受付)

A DEVELOPMENT OF FACTOR ANALYZING AND PREDICTING MODEL OF ABANDONED AGRICULTURAL LAND WITH MACHINE LEARNING ALGORITHMS

Takanori MATSUI, Tetsu UGATA and Takashi MACHIMURA

In Japan, cultivated land abandonment has been increasing rapidly in recent years and it causes degradation of biodiversity and decreases ecosystem services from agro ecosystem. Thus, it is required to determine the factors increasing cultivation abandonment and predict the future level of abandonment in regional scale in order to design strategy towards sustainable agriculture system under region-specific context. From this viewpoint, we tried to develop factor analysis and prediction model of abandonment cultivated land of Japan in regional scale. We used census of agriculture and forestry distributed by Ministry of Agriculture, Forestry and Fisheries of Japan. And to identify the structure of the cultivation abandonment from the census data, we employed three machine learning algorithms, General Linear Model (GLM), Multivariate Adaptive Regression Spline and Random Forests (RF). The result shows that the best machine learning algorithm to predict the cultivation abandonment in Japan was RF algorithm and it could detect major factors under regional specific situations.