O5　機械学習アルゴリズムによる耕作放棄の要因分析・予測モデルの開発

～ 耕作放棄対策支援を目的として ～

Developing a diagnostic and predicting model of cultivated land abandonment

with machine learning algorithms for the countermeasure

指導教員　町村尚准教授・地球循環共生工学領域

28E11009　宇賀田徹（Tetsu UGATA）

Abstract: In Japan, cultivated land abandonment has been increasing in recent years, upsetting natural symbiotic systems in farmlands. Thus, it is required to determine the factors of abandonment and forecast future abandonment. In this study, we developed a process to produce diagnostic and predicting model of abandonment. We employed three machine learning algorithms, which are Generalized Linear Model (GLM), Multivariate Adaptive Regression Spline (MARS), and Random Forests (RF). The result shows that the accuracy of GLM decreases according to the factors of abandonment in samples. Therefore we concluded the best machine learning algorithm for diagnostic and predicting model is RF in this problem.

Keywords: cultivated land abandonment, multivariate adaptive regression splines, random forests

１．　序論

日本では耕作放棄が増加しているため，耕作放棄圧力が高い地域を特定し，その促進要因の緩和策を支援する技術の開発が必要である．これに対して先行研究では様々な耕作放棄要因の特定や予測モデルの開発を試みてきたが，これらは特定地域を対象としているものが多い．そこで本研究では，耕作放棄の文脈が異なる地域別の要因分析と耕作放棄対策の効果予測を可能にすべく，耕作放棄の要因分析と予測を同時に行えるモデルの構築プロセスを開発すること目的とした．

２．　モデル構築プロセス

モデル構築用のデータは全て市区町村単位で農林水産省が集計した農林業センサスから収集した．また，モデル構築プロセスに用いる機械学習アルゴリズムを決定するときの分析対象地は関西で最も耕作放棄が進んでいる奈良県をケースとした．次に，応答変数は耕作放棄地面積を農地面積で標準化した耕作放棄地面積率を採用した．説明変数は収集したデータを経営耕地面積や販売農家数などの要因別の母数で標準化し，欠損値および VIF 基準によるデータの加工を行った．また，耕作放棄の要因が明示された上で，十分な予測確度が期待できるという要求から，モデル構築プロセスに用いる機械学習アルゴリズムの候補には一般化線形モデル，多変量適応型回帰スプライン法，ランダムフォレスト法の 3 つのアルゴリズムを選定した．その中から 2005 年のデータで構築したモデルによる 2010 年の予測確度の比較を相関係数と平均絶対残差（残差）を基準に行い，最適アルゴリズムを決定した．

（１）　一般化線形モデル（GLM: Generalized Linear Model）

応答変数は比率のため，誤差構造は二項分布に従うと仮定した．リンク関数は各リンク関数で構築したモデルの予測確度を比較し，最も予測確度が高かった logit を最適なリンク関数と判断した．ベストモデルの選択は AIC 基準のステップワイズ法を用い，その結果，13 の変数が選択された．

（２）　多変量適応型回帰スプライン法（MARS: Multivariate Adaptive Regression Spline）

応答変数の誤差構造とリンク関数は GLM と同様に二項分布および logit とした．説明変数の選択には主効果モデル，2 次，3 次の交互作用モデルを採用した．

（３）　ランダムフォレスト法（RF: Random Forests）

500 回のブートストラップを行い，樹木 1 本に用いる説明変数は総数の 1/3，終結ノードを構成する標本数は 5 以下とした．また，各変数の重要度 [1)] を計算し，重要な変数から順に加えたモデルを構築し，モデル変数と予測確度の挙動からスクリー基準の変数選択を行い，18 変数が採択された．

３．　結果

GLM，MARS，RF の予測確度を比較した結果，GLM と RF の残差はどちらも 0.03 と MARS のみが予測確度が低かったため，本研究で採用するアルゴリズムには不適切と判断した．次に 2005 年のデータで構築した GLM，RF による 2010 年の耕作放棄予測値と実測値を比較した結果を**表 1** と**図 1 (a)**に示す．結果から、RF の方が相関係数は高く，外れ値に対して頑健である．次に， MARS を除く 2 つのアルゴリズムを用いたモデル構築プロセスを耕作放棄地面積率が高い長崎県と低い滋賀県に適用し，モデルの予測確度を比較した（**表 1**，**図 1 (b)**，**(c)**）．GLM は長崎県において残差が大きくなっており，地域により構築されるモデルの予測確度に差があるが，RF では大きく差はないため，耕作放棄の文脈に依存せず予測が可能である．さらに，要因分析においても明確なモデル式はないが，変数の重要度から要因が特定できる．以上から，本研究のモデル構築プロセスにおいて最適なアルゴリズムは RF と判断した．

４．　考察

地域によって GLM の予測確度が異なる理由は，奈良県，長崎県，滋賀県ではモデルに採用された変数の数がそれぞれ 13，38，28 であり，RF でもそれぞれ 18，28，25 と長崎県，滋賀県では選択される変数が多く，このように耕作放棄要因が多く存在する地域では，モデルが過学習を起こすためだと考えられる．ゆえに説明変数の特徴ごとにクラスタリングするなど，耕作放棄要因が似ている標本ごとにモデル構築を行うことで過学習を回避できる可能性がある．

RF は滋賀県での将来予測において，予測値の分布範囲が実測値の分布範囲に比べて狭くなっている（**図 1 (c)**）．この原因は滋賀県では，2005 年の耕作放棄地面積率が 0 から 0.15 まで分布しているが 0.02 以下に全標本の 49 % が偏って分布しているため，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，これらの小さな値のノードに標本が振り分けられやすくなり，その影響で全樹木による予測値の平均が実測値より小さくなるためだと考えられる．ゆえに，この問題を回避するためにはモデル構築用標本において，耕作放棄地面積率の分布の偏りを補正したモデル開発を行う必要がある．

表 1 各県での耕作放棄予測の結果

県名	アルゴリズム	相関係数	平均絶対残差
奈良県	GLM	0.64	0.030
	RF	0.71	0.030
長崎県	GLM	0.50	0.149
	RF	0.69	0.046
滋賀県	GLM	0.36	0.019
	RF	0.61	0.015

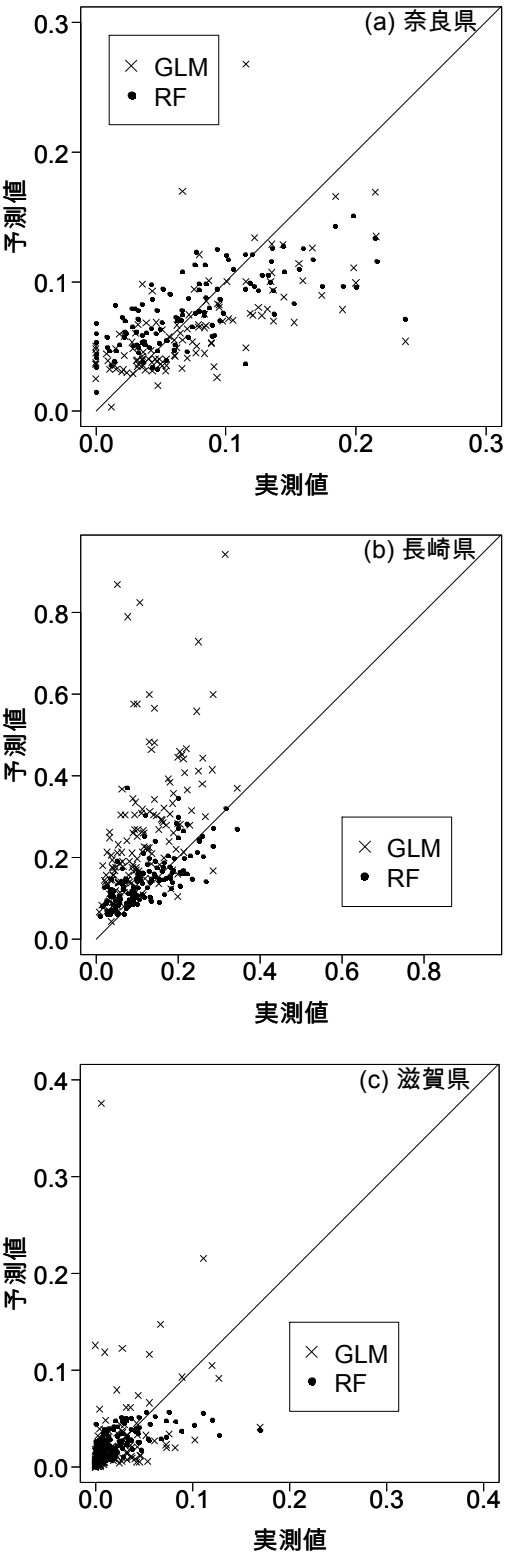

図 1 耕作放棄の将来予測の結果

参考文献

1) Breiman : Random Forests, Machine Learning, Vol. 45 (1), pp. 5-32, 2001.