ロジスティック回帰モデル

「Island Scrub Jay」という鳥類の生息地予測

KIM SUYEON

CHAPTER

分析背景

「Island Scrub Jay」は、サンタクルーズ島にのみ生息する鳥類であり、その生息地の変化に対して脆弱である。気候変動や生息地の破壊といった脅威が存在する中で、本種の生息地選好を理解し、予測することは、保全戦略を策定する上で不可欠である。

分析目標

GOAL１．「Island Scrub Jay」が好む地形の特性を分析し、理解する。

GOAL２．一般化線形モデルを用いて「Island Scrub Jay」の存在有無を予測する。

Dataset

2008年秋にサンタクルーズ島の307地点で収集された「Island Scrub Jay」の観測データ

-データサイズ: 5625

- 欠損値（NA）を含む行数: 5322

- 反応変数：「Island Scrub Jay」の存在有無（1： 存在、0： 不在）

- 説明変数 ：

　　1）x, y座標：空間的な位置情報

　　2）elev：調査地点の海抜高度

　　3）forest：調査地点における森林の割合

　　4 ）chap：調査地点におけるchaparral (低木地帯）の割合

反応変数のNA処理

	isj	x座標	y 座標	elev	forest	chap
NA 行	5318	0	0	2838	2838	2838

- 反応変数isjに欠損値を含むデータは合計5318 件であり、データセット全体の95%がNAである。このNAは、該当地点で「Island Scrub Jay」のデータが収集されなかったことを意味する。
- このデータセットに関する追加情報が得られないため、isjの欠損値を補完する方法は使用しない。したがって、isjがNAであるデータ（行）を活用しない。

説明変数のNA処理

	isj	x座標	y 座標	elev	forest	chap
NA 行	0	0	0	4	4	4

反応変数isjに欠損値が存在するデータを削除した後、再度欠損データ（行）の数を確認すると、上記の表の通りである。

- 変数elevのデータが存在しない場合、森林割合や低木地帯割合のデータも存在しない。

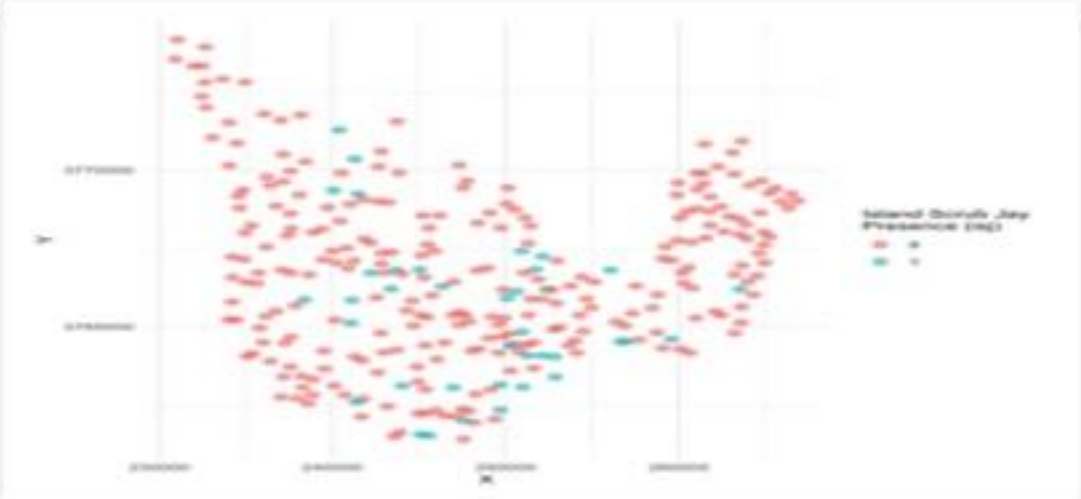
- 合計307件のデータのうち、約1%に欠損値が含まれているため、これらのデータを使用しない。

1）303件のデータを使って、「Island Scrub Jay」の存在有無を予測するモデルを作る。
2）この予測モデルを2484件のデータに適用して、欠損値のisjを予測する。

*2484件のデータは、反応変数isjにのみ欠損値（NA）が存在し、説明変数には欠損がない。

03. データ前処理

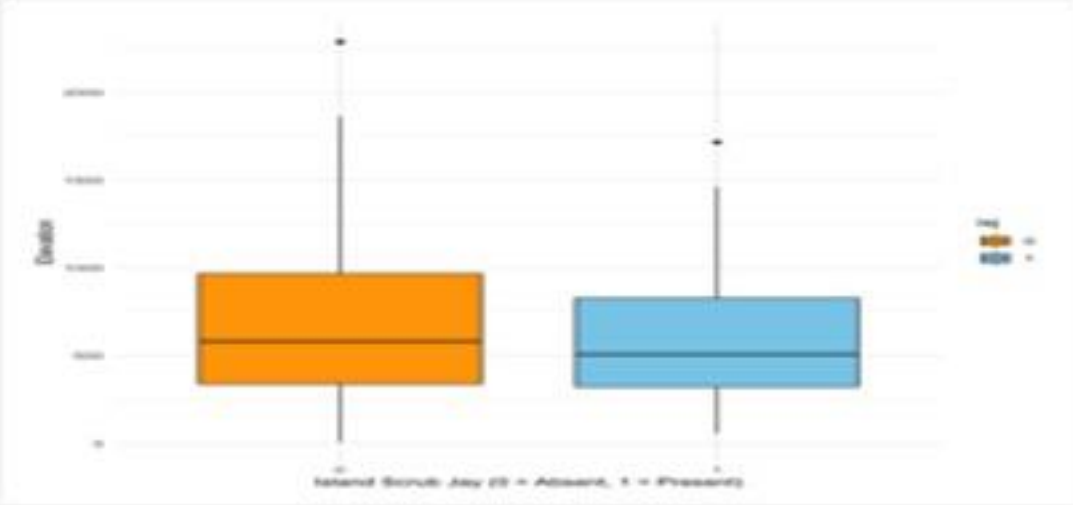
x, y座標



📌 [「Isalnd Scrub Jay」が存在しない地域(isj = 0)]
- 全体的に均等に分布しており、特定の位置に集中していない。

📌 [「Island Scrub Jay」が存在する地域(isj = 1)]
- 特定の位置に集中して観察される傾向もあり、特にX座標が約240000～250000、Y座標が 3765000～3770000の範囲で密集している。

elev

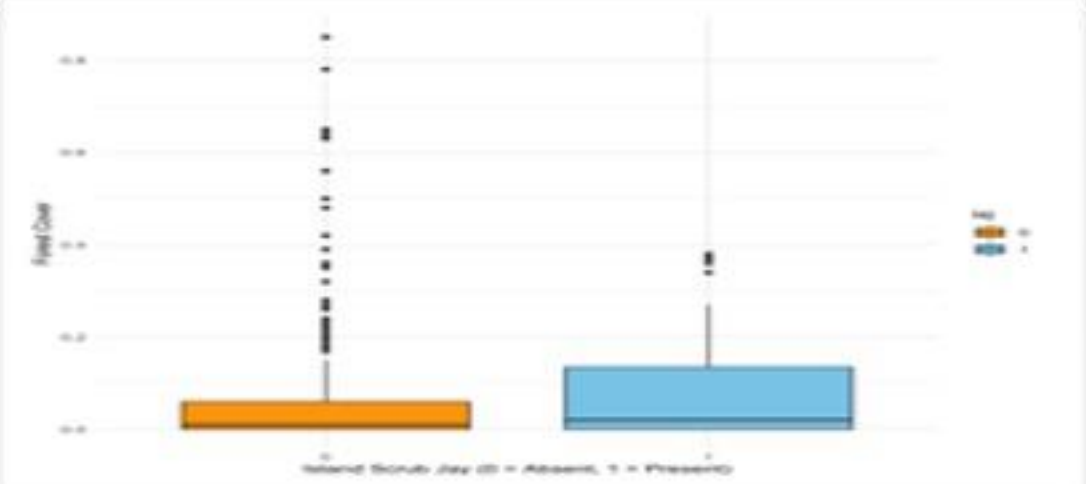


📌 [「Isalnd Scrub Jay」が存在しない地域(isj = 0)]
- 標高の分布が広く、一部に外れ値が存在する。

📌 [「Island Scrub Jay」が存在する地域(isj = 1)]
- 存在しない地域よりも分布の幅が狭い。

03. データ前処理

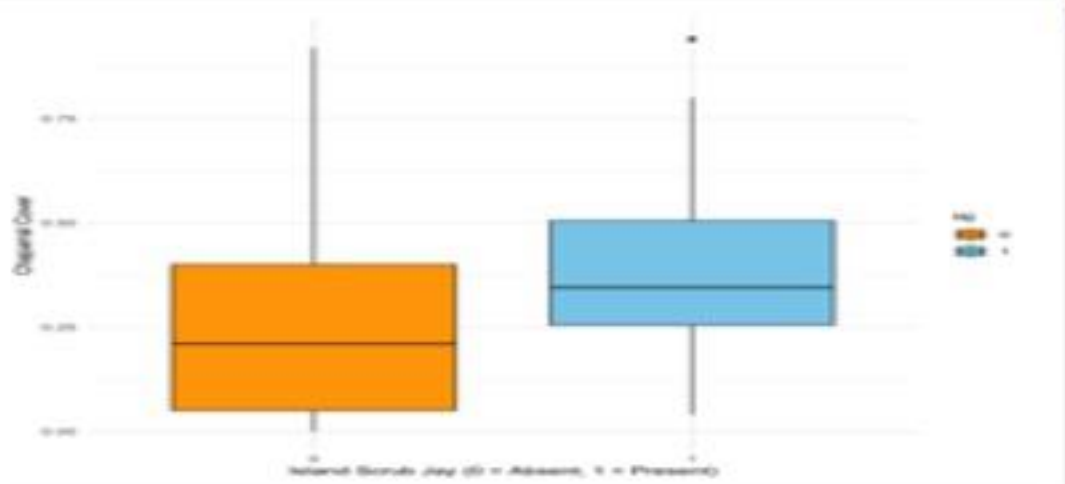
forest



📌 **[「Isalnd Scrub Jay」が存在しない地域(isj = 0)]**
- 森林割合はほとんど0〜0.1に分布し、中央値は約0.05と非常に低い。
- 森林割合はほとんど0〜0.1に分布し、中央値は約0.05と非常に低い。

📌 **[「Island Scrub Jay」が存在する地域(isj = 1)]**
- 森林割合の分布がより広い。
- 外れ値（Outlier）はほとんど見られない。

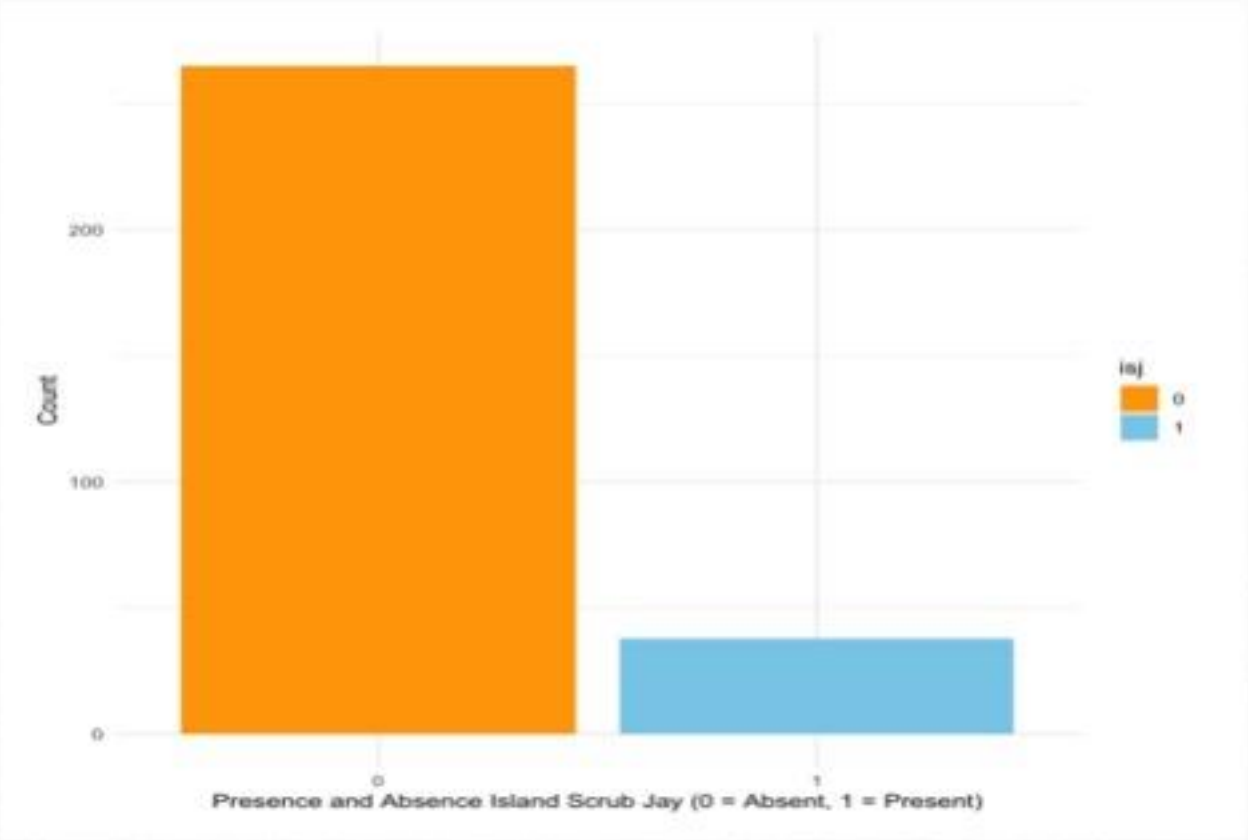
chap



📌 **[「Isalnd Scrub Jay」が存在しない地域(isj = 0)]**
- ほとんどのデータが約0.1〜0.45の範囲に分布。
- 低木地帯割合が低い地域では存在しない。

📌 **[「Island Scrub Jay」が存在する地域(isj = 1)]**
- 中央値は、存在しない地域よりも高い。
- 低木地帯割合が約0.6〜0.7の高い値でも観察される。
- 外れ値が1つ存在し、低木地帯割合が非常に高い地域でも「Isalnd Scrub Jay」が生息する可能性があることを示唆している。

反応変数isj



「Island Scrub Jay」が存在しないデータの方が、存在するデータよりも多い。

-> 使用するデータ303件のうち、isj=1のデータは38件のみであり、データが偏っていることがわかる。反応変数（isj）は二値データであるため、プロビット回帰とロジスティック回帰の両方が使用可能。

Logistic Regression

1）反応変数が1または0の二値データであるため、「Binary Regression」を使用する。

- プロビット回帰も「Binary Regression」の一種だが、ロジスティック回帰の方が変数の解釈が容易であるため、こちらを選択する。

2）「Isalnd Scrub Jay」が好む地域の特徴を把握するため、仮説検定を活用する。

- 仮説検定を通じて、説明変数が反応変数に対して有意な影響を与えるかどうかを確認するため、このモデルを使う。

分析フロー

1. **新しい変数の定義**
   Rawデータセットに多項式（Polynomial）や交互作用（Interaction）を追加 し、新しい変数を作成する。

2. **データセット分割**
   TrainデータとTestデータを7:3の割合で分割する。

3. **分類のThresholdの定義**
   モデルの予測値に対する適切な分類のThresholdを定義する。

4. **最適な説明変数の組み合わせを選択**
   既存の説明変数と新たに定義した変数の中から、誤分類率（Misclassification Rate）が最小となる組み合わせを探索する。

5. **仮説検定**
   仮説検定を行い、反応変数に有意な影響を与える説明変数を特定する。

6. **最適なモデルの可視化**
   選択したモデルを、先に説明した 2484件のデータセットに適用する。

04. モデリング＆予測

Step1 新しい変数の定義

Rawデータに多項式（Polynomial）や交互作用（Interaction）を追加し、既存の変数を活用して以下の15個の新しい変数を定義する。

変数	x2	y2	elev2	forest2	chap2
定義	x^2	y^2	$elev^2$	$forest^2$	$chap^2$
変数	x.y	x.elev	x.forest	x.chap	y.elev
定義	x*y	x*elev	x*forest	x*chap	y*elev
変数	y.forest	y.chap	elev.forest	elev.chap	forest.chap
定義	y*forest	y*chap	elev*forest	elev*chap	forest*chap

Step2 Train, test dataset

TrainデータとTestデータを7:3の割合で分割する。

Step3 分類のThresholdの定義する

-「Island Scrub Jay」が存在する地域をより多く検出することを重視するため、分類のThresholdを0.5ではなく0.3に調整する。

-303件のデータのうち、isj = 1（「Isalnd Scrub Jay」が存在するデータ）は38件のみであり、データが偏っていることが分かる。

-この調整により、存在しない場所を誤って"存在する"と判定するリスク（False Positive）が増加するが、見逃していた生息地を特定できるメリットの方が重要であると判断し、閾値を0.3に設定する。

Step4 最適な説明変数の組み合わせ

誤分類率を最小にする説明変数の組み合わせを特定した結果、誤分類率は0.120879となった。
以下の5つの組み合わせが、最も低い誤分類率を示した。

Best Features1: ['x', 'forest', 'chap', 'forest2', 'x.y', 'y.chap', 'elev.chap', 'forest.chap']
Best Features2: ['x', 'forest', 'chap', 'elev2', 'x.y', 'x.forest', 'y.chap', 'elev.forest', 'elev.chap']
Best Features3: ['y', 'elev', 'forest', 'chap', 'elev2', 'forest2', 'y.forest', 'y.chap', 'elev.chap', 'forest.chap']
Best Features4: ['y', 'elev', 'forest', 'chap', 'elev2', 'forest2', 'x.elev', 'y.forest', 'y.chap', 'elev.chap', 'forest.chap']
Best Features5: ['y', 'elev', 'forest', 'chap', 'y2', 'elev2', 'forest2', 'x.chap', 'y.forest', 'y.chap', 'elev.chap', 'forest.chap']

この組み合わせの中で、AICが最も低い変数の組み合わせは「Best Features1」である。

Step5 仮説検定

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.6723     0.6347  -5.786  7.2e-09 ***
x             142.3436    42.4995   3.349 0.000810 ***
forest          3.1322     1.1592   2.702 0.006894 **
chap        -1714.9275   573.0856  -2.992 0.002768 **
forest2        -2.4274     1.0427  -2.328 0.019912 *
x.y          -142.2103    42.4085  -3.353 0.000798 ***
y.chap       1715.9381   573.2942   2.993 0.002761 **
elev.chap      -0.5189     0.3824  -1.357 0.174807
forest.chap    -0.8115     0.4921  -1.649 0.099122 .
---
```

有意な変数に対する仮説検定をする。

H0:変数elev.chapは反応変数に有意な影響を与えない。
H1: 変数 elev.chapは反応変数に有意な影響を与える。

H0:変数forest.chapは反応変数に有意な影響を与えない。
H1:変数forest.chapは反応変数に有意な影響を与える。

➔ 有意水準0.05において、p値が有意水準より大きいため、elev.chapとforest.chapは反応変数に有意な影響を与えるとは言えない。

=> elev.chap、forest.chap以外の説明変数は反応変数に有意な影響を与えることが確認された。

Step6 最適なモデル

logit(P)=−3.6723+142.3436·x+3.1322·forest-1714.9275·chap-2.4274forest2-142.2103·(x·y)+1715.9381·(y·chap)-0.5189·(elev·chap)-0.8115·(forest·chap)

1) すべての説明変数が0のとき、「Island Scrub Jay」が存在する確率は存在しない確率の約2.5%である。

2) 東へ向かうほど、「Island Scrub Jay」はより多く生息する傾向がある。しかし、北東部では、生息確率が低くなる可能性がある。

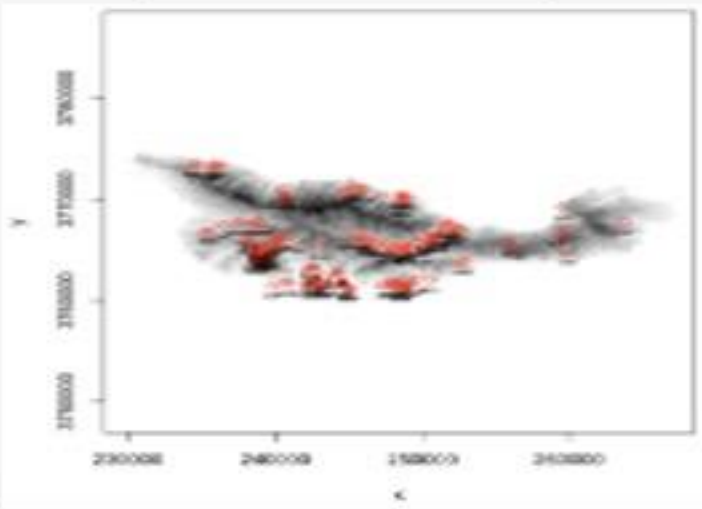
4) 森林の面積が増加するほど、生息確率は上昇する。しかし、適度な森林面積が必要であり、過度に多いとむしろ不利になる可能性がある。また、 低木地帯地域内には不利になる可能性がある。

5) 単独の低木地帯地域では生息が困難であり、特定の条件（北部地域など）でのみ例外的に生息する可能性がある。

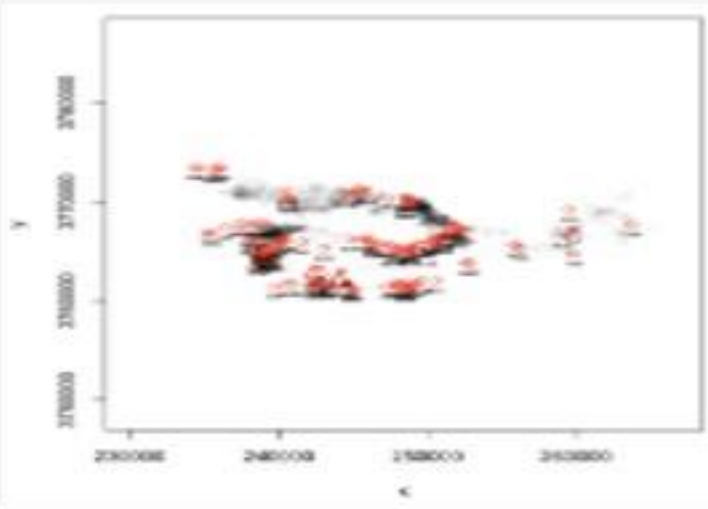
6) 標高が高い低木地帯地域は、「Island Scrub Jay」の生息に非常に不利である。

2484件のデータセットに適用

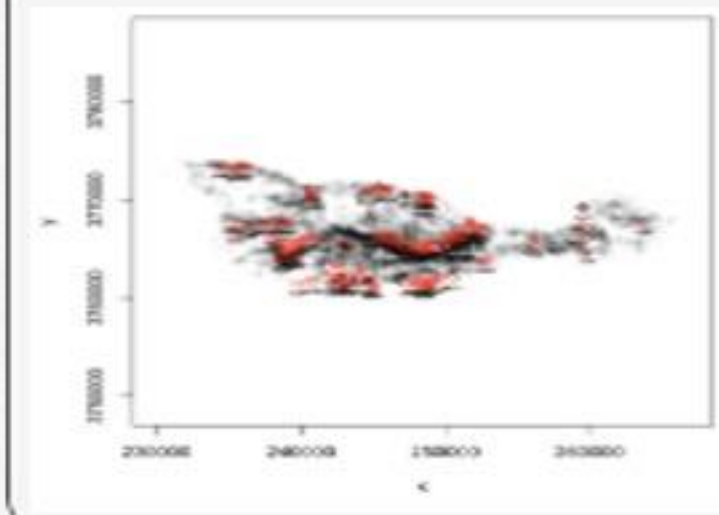
標高



森林割合



低木地帯割合



*反応変数の値がNAではないデータにおける予測確率を示すグラフを作成し、適合したモデルが 「Island Scrub Jay」が存在すると判断した地点のみoマークを付ける。

分析結果

- 説明変数elev.chap、forest.chapを除いたすべての説明変数は、反応変数に有意な影響を及ぼす。

- 標高がそれほど高くなく、森林と低木地帯の割合が適度に調和している場合、「Island Scrub Jay」が生息する確率は高くなる。その際、北東へ向かうほど生息確率は低くなる。

プロジェクト意義

1)「Island Scrub Jay」の生息地特性の定量的理解
- 説明変数ごとの影響を数値で評価し、「Island Scrub Jay」が好む環境の要因を明確にした。

2) 生息地の予測モデルの構築
- ロジスティック回帰モデルを用いて、「Island Scrub Jay」の生息確率を予測するモデルを構築した。
- 未観測の地域でも、生息する可能性を推定できた。

3) 生態系保全や管理への応用
- 「Island Scrub Jay」の生息地を科学的に特定し、保全活動の優先順位をつける指標 になる。
- 将来的に気候変動や森林破壊が生息地に与える影響をシミュレーションする基盤となる。

4) データ中心のアプローチの適用
- 環境データの活用と統計モデルを組み合わせることで、生態学研究の新たなアプローチを示した。
- 単なる観察結果ではなく、予測可能なモデルを活用することで、より客観的な保全計画の策定が可能になった。