Multivariate Analysis I

HES 505 Fall 2022: Session 19

Matt Williamson

Objectives

By the end of today you should be able to:

- Recognize the link between regression analysis and overlay analysis
- Generate spatial predictions based on regression analysis
- Extend logistic regression to presence-only data models

Estimating favorability

$$F(\mathbf{s}) = \prod_{M=1}^{m} X_m(\mathbf{s})$$

- Treat $F(\mathbf{s})$ as binary
- Then $F(\mathbf{s}) = 1$ if all inputs ($X_m(\mathbf{s})$) are suitable
- Then $F(\mathbf{s}) = 0$ if not

Estimating favorability

$$F(\mathbf{s}) = f(w_1 X_1(\mathbf{s}), w_2 X_2(\mathbf{s}), w_3 X_3(\mathbf{s}), \ldots, w_m X_m(\mathbf{s}))$$

- $F(\mathbf{s})$ does not have to be binary (could be ordinal or continuous)
- $X_m(\mathbf{s})$ could also be extended beyond simply 'suitable/not suitable'
- Adding weights allows incorporation of relative importance
- Other functions for combining inputs ($X_m(\mathbf{s})$)

Weighted Linear Combinations

$$F(\mathbf{s}) = \frac{\sum_{i=1}^{m} w_i X_i(\mathbf{s})}{\sum_{i=1}^{m} w_i}$$

- $F(s)$ is now an index based of the values of $X_m(\mathbf{s})$
- w_i can incorporate weights of evidence, uncertainty, or different participant preferences
- Dividing by $\sum_{i=1}^{m} w_i$ normalizes by the sum of weights

Model-driven overlay

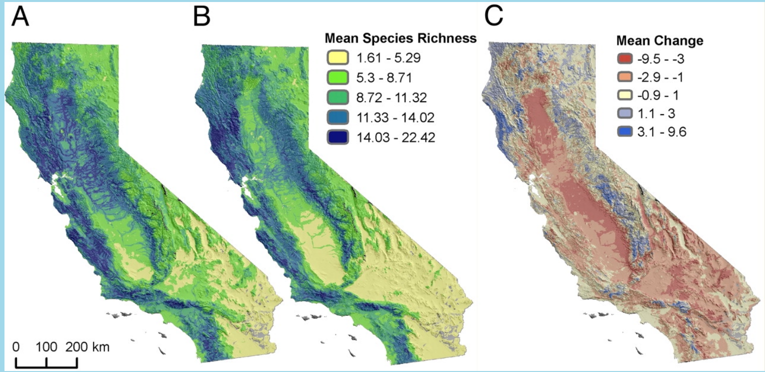
$$F(\mathbf{s}) = w_0 + \sum_{i=1}^{m} w_i X_i(\mathbf{s}) + \epsilon$$

- If we estimate w_i using data, we specify F(s) as the outcome of regression
- When F(s) is binary → logistic regression
- When F(s) is continuous → linear (gamma) regression
- When F(s) is discrete → Poisson regression
- Assumptions about ϵ matter!!

Logistic Regression and Distribution Models

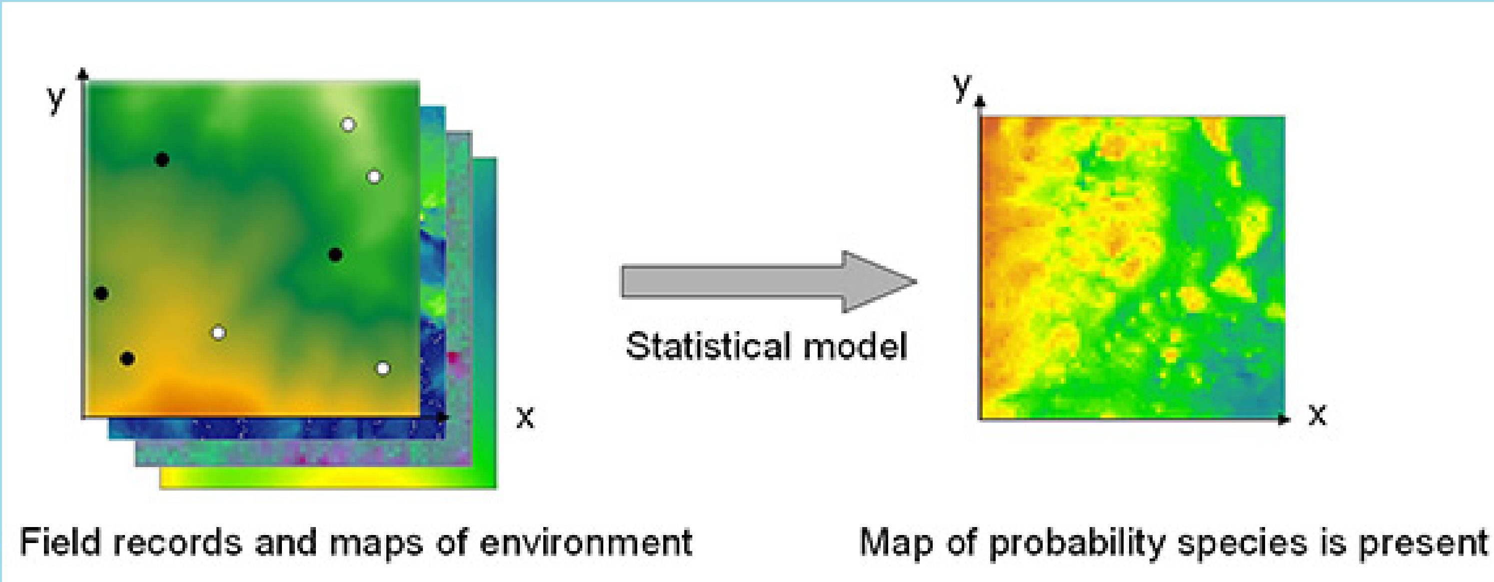
Why do we create distribution models?

- To identify important correlations between predictors and the occurrence of an event
- Generate maps of the ‘range’ or ‘niche’ of events
- Understand spatial patterns of event co-occurrence
- Forecast changes in event distributions



From Wiens et al. 2009

General analysis situation



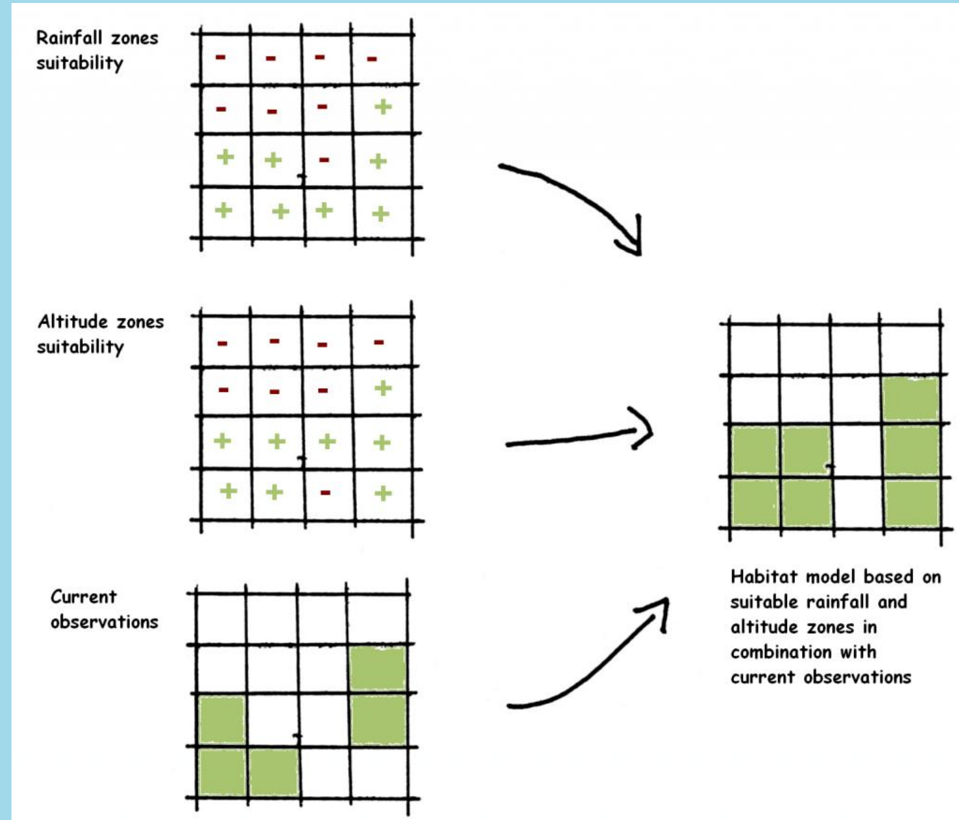
From Long

- Spatially referenced locations of events (**y**) sampled from the study extent
- A matrix of predictors (**X**) that can be assigned to each event based on spatial location

> **Goal**: Estimate the probability of occurrence of events across unsampled regions of the study area based on correlations with predictors

Modeling Presence-Absence Data

- Random or systematic sample of the study region
- The presence (or absence) of the event is recorded for each point
- Hypothesized predictors of occurrence are measured (or extracted) at each point



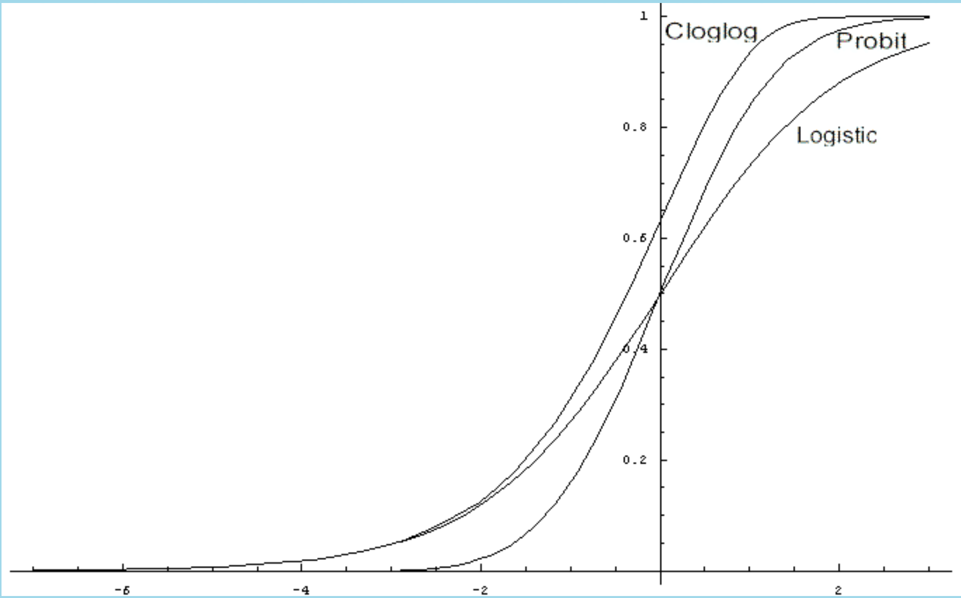
From By Ragnvald - Own work, CC BY-SA 3.0

Logistic regression

- We can model favorability as the **probability** of occurrence using a logistic regression
- A *link* function maps the linear predictor ($\mathbf{x_i}'\beta + \alpha$) onto the support (0-1) for probabilities
- Estimates of β can then be used to generate 'wall-to-wall' spatial predictions

$$y_i \sim \mathrm{Bern}(p_i)$$

$$\mathrm{link}(p_i) = \mathbf{x_i}'\beta + \alpha$$



From Mendoza

An Example

Inputs from the `dismo` package

An Example

The sample data

```
head(pres.abs)
```

An Example

Building our dataframe

```
pts.df <- terra::extract(pred.stack, vect(pres.abs), df=TRUE)
head(pts.df)
```

An Example

Building our dataframe

```
pts.df[,2:7] <- scale(pts.df[,2:7])
summary(pts.df)
```

An Example

Looking at correlations

```
pairs(pts.df[,2:7])
```

An Example

Looking at correlations

```
corrplot(cor(pts.df[,2:7]), method = "number")
```

An Example

Fitting some models

```
pts.df <- cbind(pts.df, pres.abs$y)
colnames(pts.df)[8] <- "y"
logistic.global <- glm(y~., family=binomial(link="logit"), data=pts.df[,2:8
logistic.simple <- glm(y ~ MeanAnnTemp + TotalPrecip, family=binomial(link=
logistic.rich <- glm(y ~ MeanAnnTemp + PrecipWetQuarter + PrecipDryQuarter,
```

An Example

Checking out the results

```
summary(logistic.global)
```

An Example

Checking out the results

```
summary(logistic.simple)
```

An Example

Checking out the results

```
summary(logistic.rich)
```

An Example

Comparing models

```
AIC(logistic.global, logistic.simple, logistic.rich)
```

An Example

Generating predictions

```
preds <- predict(object=pred.stack, model=logistic.simple)
plot(preds)
plot(pres.pts$geometry, add=TRUE, pch=3, col="blue")
plot(abs.pts$geometry, add=TRUE, pch ="-", col="red")
```

An Example

Generating predictions

```r
preds <- predict(object=pred.stack, model=logistic.simple, type="response")
plot(preds)
plot(pres.pts$geometry, add=TRUE, pch=3, col="blue")
plot(abs.pts$geometry, add=TRUE, pch ="-", col="red")
```

An Example

Generating predictions

```
preds <- predict(object=pred.stack, model=logistic.global, type="response")
plot(preds)
plot(pres.pts$geometry, add=TRUE, pch=3, col="blue")
plot(abs.pts$geometry, add=TRUE, pch ="-", col="red")
```

An Example

Generating predictions

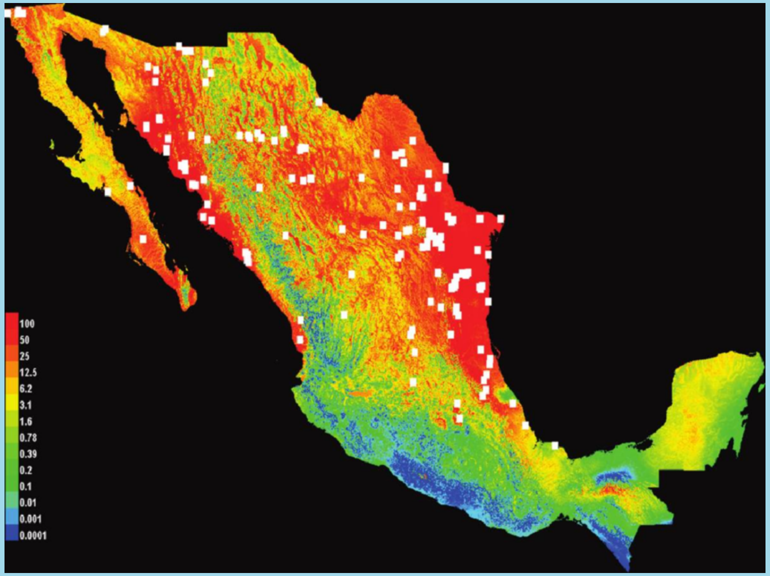
```
preds <- predict(object=pred.stack, model=logistic.rich, type="response")
plot(preds)
plot(pres.pts$geometry, add=TRUE, pch=3, col="blue")
plot(abs.pts$geometry, add=TRUE, pch ="-", col="red")
```

Key assumptions of logistic regression

- Dependent variable must be binary
- Observations must be independent (important for spatial analyses)
- Predictors should not be collinear
- Predictors should be linearly related to the log-odds
- **Sample Size**

Modelling Presence-Background Data

The sampling situation



From Lentz et al. 2008

- Opportunistic collection of presences only
- Hypothesized predictors of occurrence are measured (or extracted) at each presence
- Background points (or pseudoabsences) generated for comparison

The Challenge with Background Points

- What constitutes background?
- Not measuring *probability*, but relative likelihood of occurrence
- Sampling bias affects estimation
- The intercept

$$y_i \sim \mathrm{Bern}(p_i)$$

$$\mathrm{link}(p_i) = \mathbf{x_i}'\beta + \alpha$$

Point Process Models

- Poisson Point Process Models model location (not y)
- Number of points expected is given by a rate λ
- Model λ using Poisson regression