ROC curves for spatial point patterns and presence-absence data

Adrian Baddeley

joint work with

Ege Rubak, Suman Rakshit and Gopalan Nair

Dear Mr. Badly,

Dear Mr. Badly,

Could you please implement the Bloggs Technique in your ‘spatstat’ package?

Dear Mr. Badly,

Could you please implement the Bloggs Technique in your ‘spatstat’ package?

See attached paper by Bloggs (2015)

Yours sincerely,
A. User

Dear Mr. Badly,

Could you please implement the Bloggs Technique in your ‘spatstat’ package?

See attached paper by Bloggs (2015)

Yours sincerely,
A. User

PS. Please do it soon because my advisor wants the results on his desk on Monday morning

1. Introduction

Receiver Operating Characteristic (ROC) curve

1. Introduction

Receiver Operating Characteristic (ROC) curve

- measures the performance of a classifier/test

1. Introduction

Receiver Operating Characteristic (ROC) curve

- measures the performance of a classifier/test
- has recently been applied to **spatial data**

  to assess Species Distribution Models

Claims in the literature

The ROC for a Species Distribution Model is claimed to be

- "a measure of **goodness-of-fit** of the model"

Claims in the literature

The ROC for a Species Distribution Model is claimed to be

- "a measure of **goodness-of-fit** of the model"
- "a measure of **predictive power** of the model"

Claims in the literature

The ROC for a Species Distribution Model is claimed to be

- "a measure of **goodness-of-fit** of the model"
- "a measure of **predictive power** of the model"
- "useful for **model selection**"

Claims in the literature

The ROC for a Species Distribution Model is claimed to be

- "a measure of **goodness-of-fit** of the model"
- "a measure of **predictive power** of the model"
- "useful for **model selection**"
- "useful for **variable selection**"

Claims in the literature

The ROC for a Species Distribution Model is claimed to be

- "a measure of **goodness-of-fit** of the model"
- "a measure of **predictive power** of the model"
- "useful for **model selection**"
- "useful for **variable selection**"

Claims in the literature

The ROC for a Species Distribution Model is claimed to be

- "a measure of **goodness-of-fit** of the model"
- "a measure of **predictive power** of the model"
- "useful for **model selection**"
- "useful for **variable selection**"

Claims in the literature

The ROC for a Species Distribution Model is claimed to be

- "a measure of **goodness-of-fit** of the model"
- "a measure of **predictive power** of the model"
- "useful for **model selection**"
- "useful for **variable selection**"

Claims in the literature

The ROC for a Species Distribution Model is claimed to be

- "a measure of **goodness-of-fit** of the model"
- "a measure of **predictive power** of the model"
- "useful for **model selection**"
- "useful for **variable selection**"

Aims:

- clarify the meaning of ROC for spatial data
- identify strengths & weaknesses
- propose new extensions

(Skating over technicalities)

2. ROC curves

2. ROC curves

Assume there are two populations

- **Positive** (“infected”, “affected”)
- **Negative** (“not infected”, “not affected”)

2. ROC curves

Assume there are two populations

- **Positive** ("infected", "affected")
- **Negative** ("not infected", "not affected")

To determine the status of an individual, we can measure a quantity S ("**discriminant**", "clinical indicator")

Large values of S suggest that the individual is positive.

Assume there are two populations

- **Positive** ("infected", "affected")
- **Negative** ("not infected", "not affected")

To determine the status of an individual, we can measure a quantity S ("**discriminant**", "clinical indicator")

Large values of S suggest that the individual is positive.

$$\text{predicted status} = \begin{cases} \text{Positive} & \text{if } S > t \\ \text{Negative} & \text{if } S \leq t \end{cases}$$

where t is a threshold (that needs to be chosen).

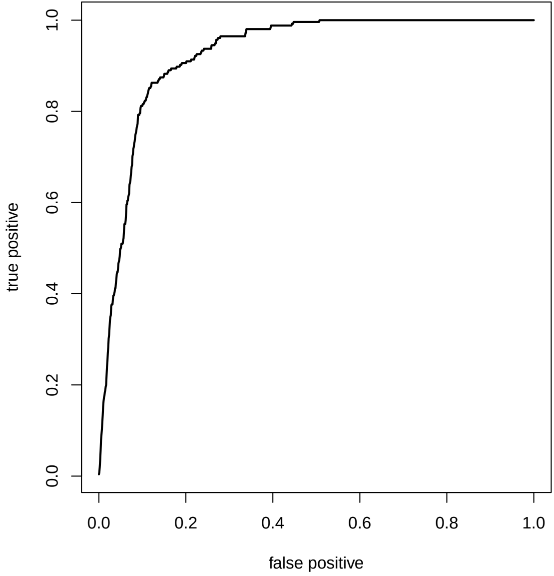
The **ROC curve** is a plot of the probability of a true positive

$$\mathrm{P}(S > t \mid \mathrm{Positive})$$

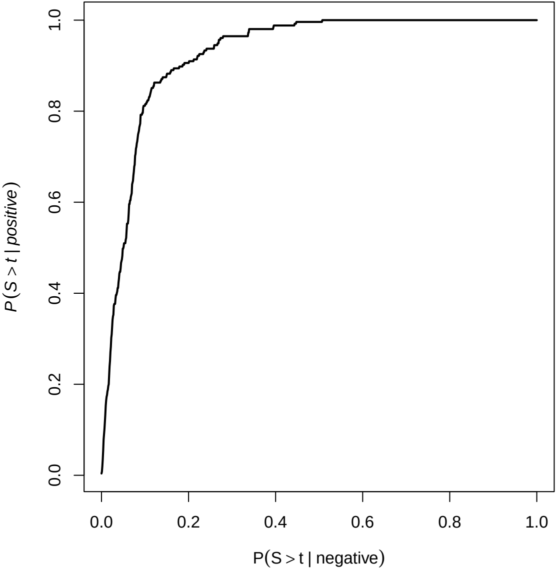
against the probability of a false positive

$$\mathrm{P}(S > t \mid \mathrm{Negative})$$

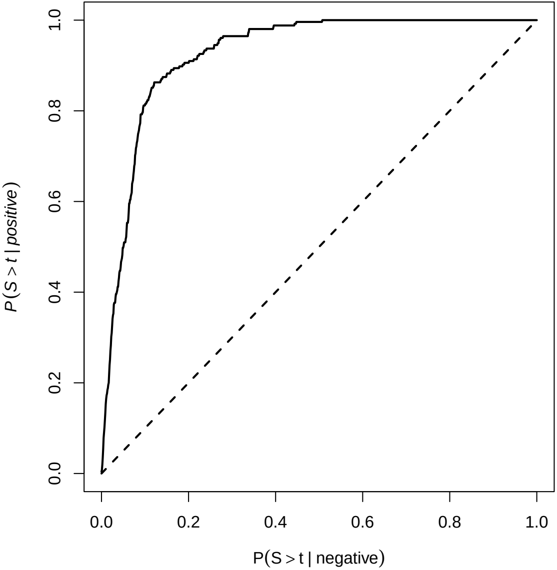
for all possible values of threshold t.



true positive
0.0
0.2
0.4
0.6
0.8
1.0
false positive
0.0
0.2
0.4
0.6
0.8
1.0



1.0
0.8
0.6
0.4
0.2
0.0
P(S > t | positive)
0.0
0.2
0.4
0.6
0.8
1.0
P(S > t | negative)


1.0
0.8
0.6
0.4
0.2
0.0
$P(S > t \mid positive)$
0.0
0.2
0.4
0.6
0.8
1.0
$P(S > t \mid negative)$

Other ways to say it:

Other ways to say it:

- The ROC curve is a plot of **power** against **size** (or **sensitivity** against "1− **specificity**") for the hypothesis test of

$$H_0 : \text{Negative}$$

vs

$$H_1 : \text{Positive}$$

which rejects H_0 when $S > t$.

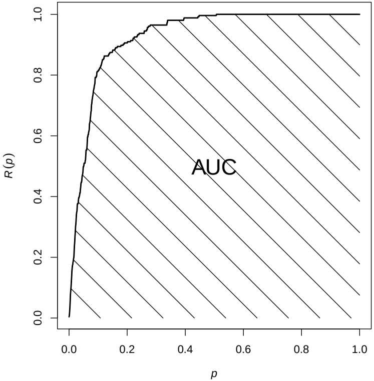
Other ways to say it:

- The ROC curve is a plot of **power** against **size** (or **sensitivity** against "1− **specificity**") for the hypothesis test of

$$\begin{array}{rl} & H_0 : \text{Negative} \\ \text{vs} & \\ & H_1 : \text{Positive} \end{array}$$

  which rejects H_0 when $S > t$.
- The ROC curve is a **P–P plot** comparing the distributions of the variable $-S$ in the Positive and Negative populations.

Area Under the Curve (AUC)



AUC = 1: perfect discrimination

AUC = $\frac{1}{2}$: no discrimination

Fun fact:

$$AUC = \mathbb{P}\{S(X) > S(Y)\}$$

where X, Y are independent, randomly selected members of the Positive and Negative populations respectively.

Fun fact:

$$AUC = \mathbb{P}\{S(X) > S(Y)\}$$

where X, Y are independent, randomly selected members of the Positive and Negative populations respectively.

If the two distributions are identical, then the ROC curve is the diagonal line, and AUC $= 1/2$.

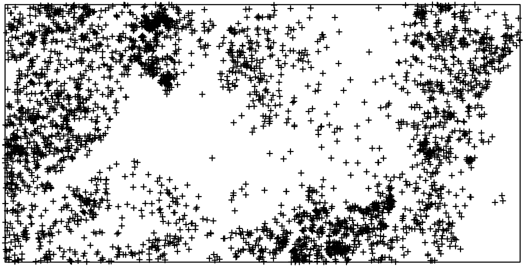
Krzanowski & Hand (2009)
ROC Curves for Continuous Data
Chapman and Hall/CRC

3. Spatial data

- spatial point patterns
- spatial presence-absence data

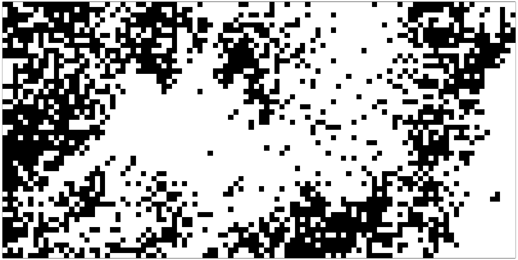
Spatial point pattern

Rainforest trees — mapped locations



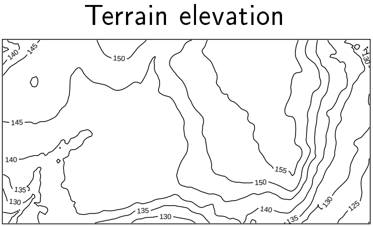
Spatial presence-absence data

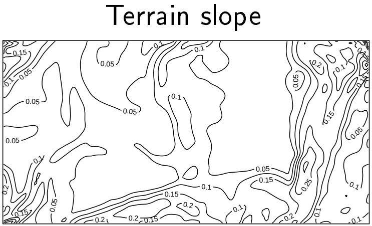
Rainforest trees — presence or absence in each 10×10 metre pixel



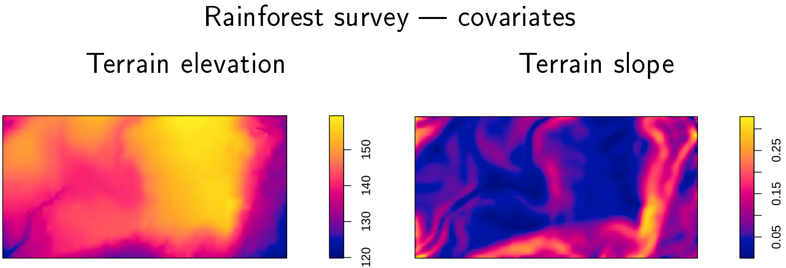
Spatial covariates

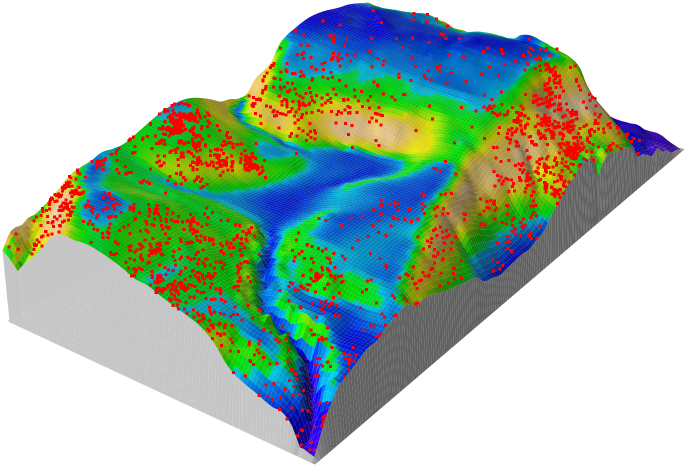
Rainforest survey — covariates



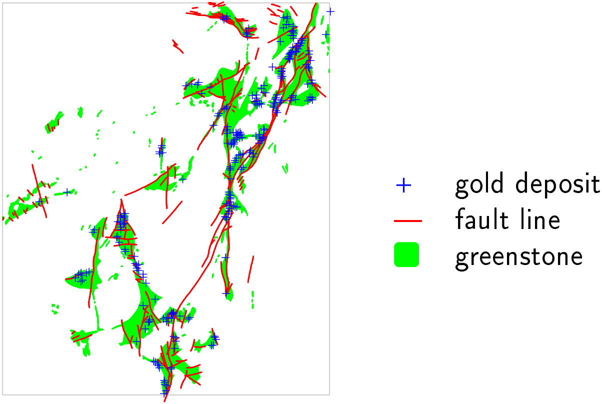


Spatial covariates





Geological survey



4. ROC for spatial model

Current practice:

4. ROC for spatial model

Current practice:

- Fit a statistical model to spatial presence-absence data

4. ROC for spatial model

Current practice:

- Fit a statistical model to spatial presence-absence data

$$\mathbb{P}\{\text{presence}\} = f(\text{covariates})$$

Current practice:

- Fit a statistical model to spatial presence-absence data

$$\mathbb{P}\{\text{presence in pixel } j\} = f(\text{covariates at pixel } j)$$

4. ROC for spatial model

Current practice:

- Fit a statistical model to spatial presence-absence data

$$\mathbb{P}\{\text{presence in pixel } j\} = f(\text{covariates at pixel } j)$$

- Calculate the predicted probability of presence in each pixel

4. ROC for spatial model

Current practice:

- Fit a statistical model to spatial presence-absence data

$$\mathbb{P}\{\text{presence in pixel } j\} = f(\text{covariates at pixel } j)$$

- Calculate the predicted probability of presence in each pixel
- Calculate the ROC curve using

  Positive 'population' = pixels with observed **presence**

  Negative 'population' = pixels with observed **absence**

  discriminant = predicted probability of presence

Franklin (2009)
Mapping Species Distributions: Spatial Inference and Prediction
Cambridge University Press

For each pixel j, let

$$\begin{aligned} x_j &= \text{value of covariate at } j \text{ (possibly vector)} \\ y_j &= \begin{cases} 1 & \text{if trees are present} \\ 0 & \text{if trees are absent} \end{cases} \\ p_j &= \mathbb{P}(Y_j = 1) \\ &= \mathbb{E}[Y_j] \end{aligned}$$

y_j = presence/absence indicator
x_j = covariate
$p_j = \mathbb{P}(Y_j = 1)$

y_j = presence/absence indicator
x_j = covariate
$p_j = \mathbb{P}(Y_j = 1)$

- Formulate a model for p_j as a function of x_j.

y_j = presence/absence indicator
x_j = covariate
$p_j = \mathbb{P}(Y_j = 1)$

- Formulate a model for p_j as a function of x_j.
- Fit the model and compute $\widehat{p}_j$.

y_j = presence/absence indicator
x_j = covariate
$p_j = \mathbb{P}(Y_j = 1)$

- Formulate a model for p_j as a function of x_j.
- Fit the model and compute $\widehat{p}_j$.
- For each possible threshold t, compute

y_j = presence/absence indicator
x_j = covariate
$p_j = \mathbb{P}(Y_j = 1)$

- Formulate a model for p_j as a function of x_j.
- Fit the model and compute $\widehat{p}_j$.
- For each possible threshold t, compute
  - estimated True Positive rate

$$\mathrm{TP}(t) = \frac{\sum_j y_j \, 1\{\widehat{p}_j > t\}}{\sum_j y_j}$$

y_j = presence/absence indicator
x_j = covariate
$p_j = \mathbb{P}(Y_j = 1)$

- Formulate a model for p_j as a function of x_j.
- Fit the model and compute $\widehat{p}_j$.
- For each possible threshold t, compute
  - estimated True Positive rate

$$\mathrm{TP}(t) = \frac{\sum_j y_j \, 1\{\widehat{p}_j > t\}}{\sum_j y_j}$$

  - estimated False Positive rate

$$\mathrm{FP}(t) = \frac{\sum_j (1 - y_j) 1\{\widehat{p}_j > t\}}{\sum_j (1 - y_j)}$$

y_j = presence/absence indicator
x_j = covariate
$p_j = \mathbb{P}(Y_j = 1)$

- Formulate a model for p_j as a function of x_j.
- Fit the model and compute $\widehat{p}_j$.
- For each possible threshold t, compute
  - estimated True Positive rate

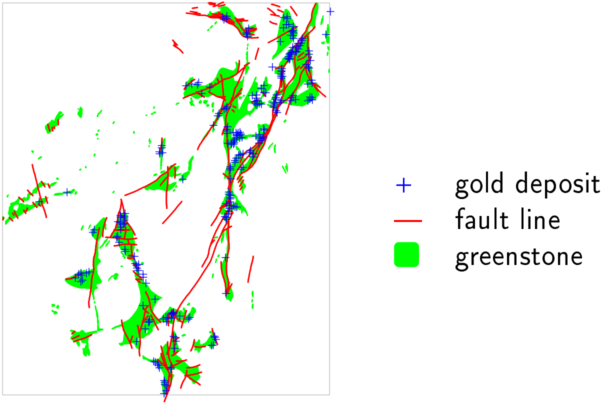
$$\mathrm{TP}(t) = \frac{\sum_j y_j \, 1\{\widehat{p}_j > t\}}{\sum_j y_j}$$

  - estimated False Positive rate

$$\mathrm{FP}(t) = \frac{\sum_j (1 - y_j) 1\{\widehat{p}_j > t\}}{\sum_j (1 - y_j)}$$

- Plot $\mathrm{TP}(t)$ against $\mathrm{FP}(t)$ for all t to produce the ROC curve.

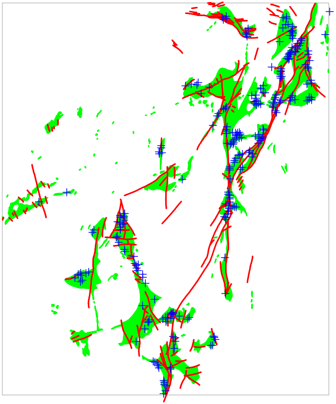
Geological survey



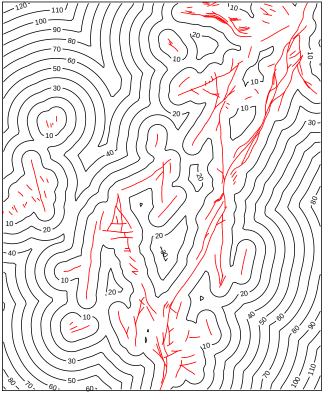
Model: logistic regression: at pixel j,

$$\log \frac{p_j}{1-p_j} = \beta_0 + \beta_1 \, d_j + \beta_2 \, g_j$$

where d_j = distance to nearest fault, g_j = greenstone indicator



Distance to nearest fault



Model: logistic regression:

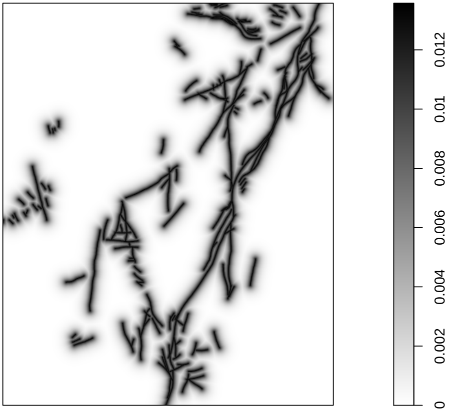
$$\log \frac{p_j}{1 - p_j} = \beta_0 + \beta_1\, d_j + \beta_2\, g_j$$

where d_j = distance to nearest fault, g_j = greenstone indicator

Model: logistic regression:

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \beta_1\, d_j + \beta_2\, g_j$$

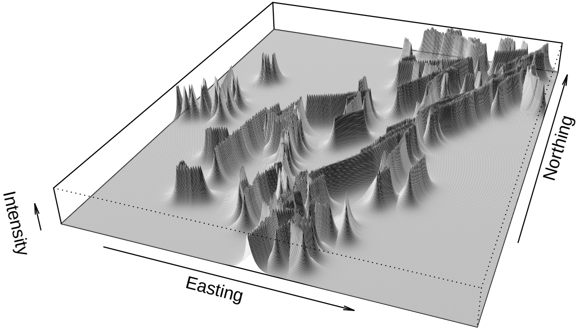
where d_j = distance to nearest fault, g_j = greenstone indicator



Model: logistic regression:

$$\log \frac{p_j}{1-p_j} = \beta_0 + \beta_1 d_j + \beta_2 g_j$$

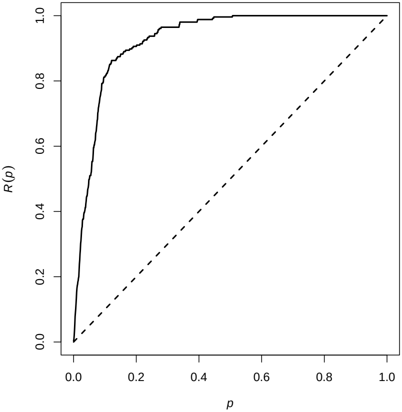
where d_j = distance to nearest fault, g_j = greenstone indicator



Model: logistic regression:

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \beta_1 \, d_j + \beta_2 \, g_j$$

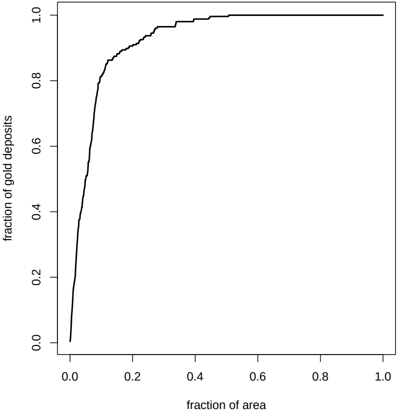
where d_j = distance to nearest fault, g_j = greenstone indicator



Model: logistic regression:

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \beta_1 \, d_j + \beta_2 \, g_j$$

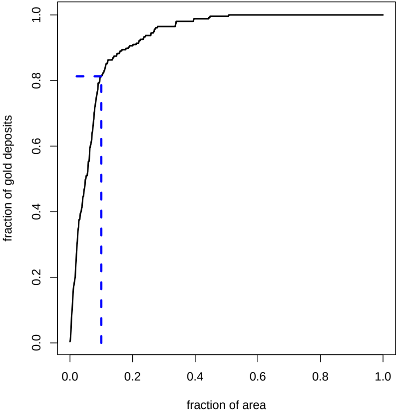
where d_j = distance to nearest fault, g_j = greenstone indicator



Model: logistic regression:

$$\log \frac{p_j}{1-p_j} = \beta_0 + \beta_1 \, d_j + \beta_2 \, g_j$$

where d_j = distance to nearest fault, g_j = greenstone indicator



AUC = 0.93

Interpretation:

Interpretation:

- Result is “good”

Interpretation:

- Result is “good”
- When the survey region is divided into regions of high and low probability of presence of gold (predicted by the fitted model),

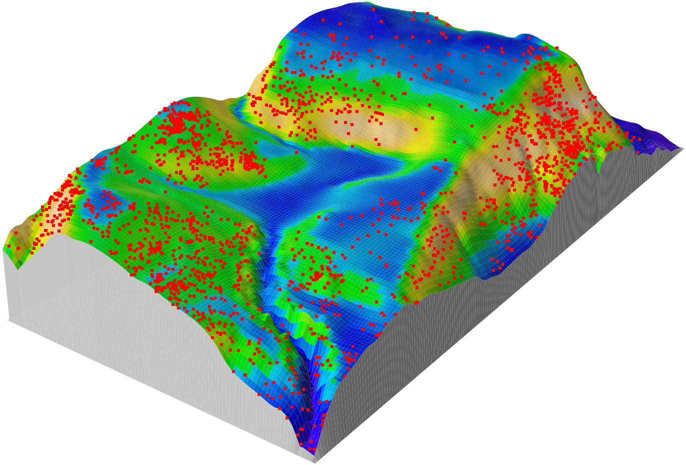
Interpretation:

- Result is "good"
- When the survey region is divided into regions of high and low probability of presence of gold (predicted by the fitted model),
  - ✓ the subdivision is *efficient*: 10% of the survey area contains 82% of the known gold deposits.

Interpretation:

- Result is "good"
- When the survey region is divided into regions of high and low probability of presence of gold (predicted by the fitted model),
  - ✓ the subdivision is *efficient*: 10% of the survey area contains 82% of the known gold deposits.
  - ✓ the model is *useful*: pixels with higher predicted probability of presence of gold are indeed much more likely to contain gold deposits

Rainforest



Model: logistic regression

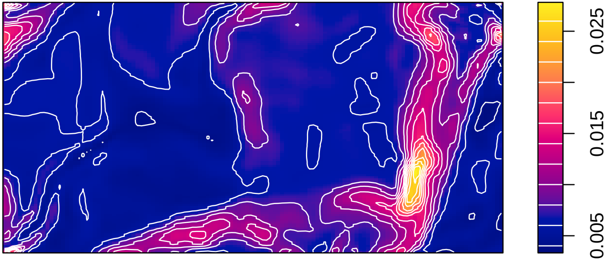
$$\log \frac{p_j}{1-p_j} = \beta_0 + \beta_1\, e_j + \beta_2\, s_j$$

where e_j = elevation, s_j = slope at pixel j

Model: logistic regression

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \beta_1 \, e_j + \beta_2 \, s_j$$

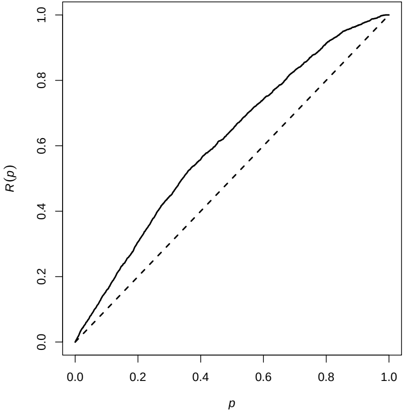
where e_j = elevation, s_j = slope at pixel j



Model: logistic regression

$$\log \frac{p_j}{1-p_j} = \beta_0 + \beta_1 e_j + \beta_2 s_j$$

where e_j = elevation, s_j = slope at pixel j



AUC = 0.61

Interpretation:

Interpretation:

- Result is “not so good”

Interpretation:

- Result is “not so good”
- Model does not efficiently segregate the rainforest into areas of high and low density of trees

✓ROC was a useful diagnostic in the two examples.

Weaknesses

Weaknesses

$\boxed{\text{(a) ROC depends on study region}}$

Weaknesses

(a) ROC depends on study region

The ROC curve depends crucially on the choice of the study region.

Weaknesses

(a) ROC depends on study region

The ROC curve depends crucially on the choice of the study region.

- The estimated false positive rate $\mathrm{FP}(t)$ is the fraction of *area in the study region* satisfying a constraint.

Weaknesses

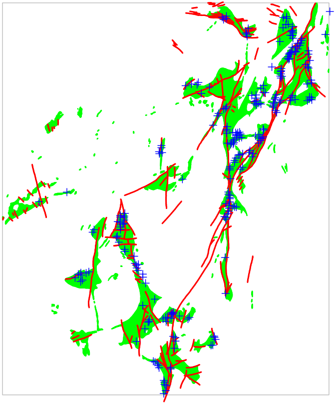
(a) ROC depends on study region

The ROC curve depends crucially on the choice of the study region.

- The estimated false positive rate $\mathrm{FP}(t)$ is the fraction of *area in the study region* satisfying a constraint.
- The estimated true positive rate $\mathrm{TP}(t)$ is the fraction of *individuals in the study region* (gold deposits, trees) satisfying a constraint.

Weaknesses

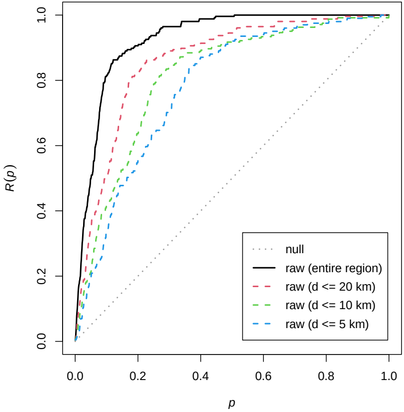
Example: Geological survey.



Weaknesses

Example: Geological survey.

Restrict the study region to those locations lying at most D kilometres from a fault.



Weaknesses

Weaknesses

- The ROC curve for a particular study region **cannot be extrapolated** to other study regions, even if the model is correct in both regions, and even if one region is a subset of the other

Weaknesses

- The ROC curve for a particular study region **cannot be extrapolated** to other study regions, even if the model is correct in both regions, and even if one region is a subset of the other
- Instances of Simpson's Paradox can occur

Weaknesses

Weaknesses

(b) ROC doesn't depend on details of model

Weaknesses

(b) ROC doesn't depend on details of model

Consider logistic regression on a single covariate z,

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \beta_1 \, z_j$$

Weaknesses

(b) ROC doesn't depend on details of model

Consider logistic regression on a single covariate z,

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \beta_1 \, z_j$$

Suppose $\widehat{\beta}_1 > 0$. Then $\widehat{p}_j$ **is an increasing function of** z_j

Weaknesses

(b) ROC doesn't depend on details of model

Consider logistic regression on a single covariate z,

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \beta_1 z_j$$

Suppose $\widehat{\beta}_1 > 0$. Then $\widehat{p}_j$ **is an increasing function of** z_j and, for any t,

$$\widehat{p}_j > t \quad \text{if and only if} \quad z_j > s$$

Weaknesses

(b) ROC doesn't depend on details of model

Consider logistic regression on a single covariate z,

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \beta_1 \, z_j$$

Suppose $\widehat{\beta}_1 > 0$. Then $\widehat{p}_j$ **is an increasing function of** z_j and, for any t,

$$\widehat{p}_j > t \quad \text{if and only if} \quad z_j > s$$

where

$$s = (\log \frac{t}{1-t} - \widehat{\beta}_0)/\widehat{\beta}_1 .$$

Weaknesses

The ROC curve for the **logistic regression on** z is the same as the ROC curve created by plotting

$$\mathrm{TP}(s) = \frac{\sum_j y_j \, 1\{z_j > s\}}{\sum_j y_j}$$

against

$$\mathrm{FP}(s) = \frac{\sum_j (1 - y_j) 1\{z_j > s\}}{\sum_j (1 - y_j)}$$

for all thresholds s.

This ROC curve is **based only on the covariate** z and **does not depend on the model!**

Weaknesses

Weaknesses

- All models in which the presence probability p is an increasing function of a single covariate z, have **the same ROC curve**

Weaknesses

- All models in which the presence probability p is an increasing function of a single covariate z, have **the same ROC curve**
- Models which are equivalent up to a monotone transformation of the mean response, have **the same ROC curve**

Weaknesses

- All models in which the presence probability p is an increasing function of a single covariate z, have **the same ROC curve**
- Models which are equivalent up to a monotone transformation of the mean response, have **the same ROC curve**
- The ROC curve of a model contains no information about the model's ability to predict absolute quantities
  (probability of presence, expected number of individuals)

Weaknesses

- All models in which the presence probability p is an increasing function of a single covariate z, have **the same ROC curve**
- Models which are equivalent up to a monotone transformation of the mean response, have **the same ROC curve**
- The ROC curve of a model contains no information about the model's ability to predict absolute quantities
  (probability of presence, expected number of individuals)
- AUC cannot be a measure of goodness-of-fit

What does ROC really measure?

The ROC for a spatial model measures the ability of the model to

What does ROC really measure?

The ROC for a spatial model measures the ability of the model to

- **segregate** the study region efficiently into subregions with high and low density of trees/deposits

What does ROC really measure?

The ROC for a spatial model measures the ability of the model to

- **segregate** the study region efficiently into subregions with high and low density of trees/deposits
- **rank** the pixels in increasing order of probability of presence of trees/deposits

5. ROC based on a spatial covariate

5. ROC based on a spatial covariate

Given a spatial covariate z, calculate an ROC curve based on z only,

5. ROC based on a spatial covariate

Given a spatial covariate z, calculate an ROC curve based on z only, by plotting

$$\mathrm{TP}(s) = \frac{\sum_j y_j \, 1\{z_j > s\}}{\sum_j y_j}$$

against

$$\mathrm{FP}(s) = \frac{\sum_j (1 - y_j) 1\{z_j > s\}}{\sum_j (1 - y_j)}$$

for all thresholds s.

5. ROC based on a spatial covariate

Given a spatial covariate z, calculate an ROC curve based on z only, by plotting

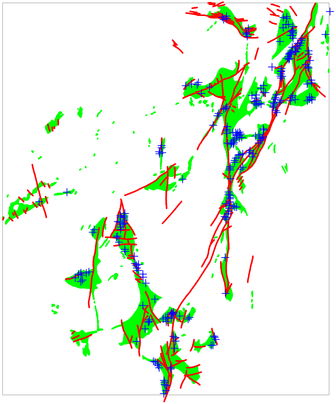
$$\mathrm{TP}(s) = \frac{\sum_j y_j \, 1\{z_j > s\}}{\sum_j y_j}$$

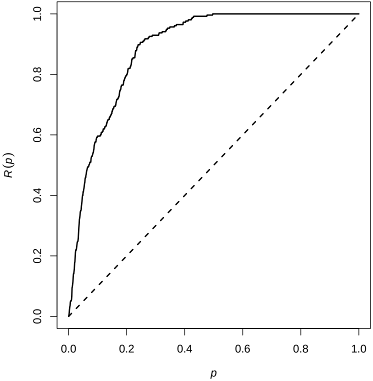
against

$$\mathrm{FP}(s) = \frac{\sum_j (1 - y_j) 1\{z_j > s\}}{\sum_j (1 - y_j)}$$

for all thresholds s. This ROC curve measures the ranking/segregating ability of the **covariate** z.

Geological survey: $z =$ distance to nearest fault





Interpretation:

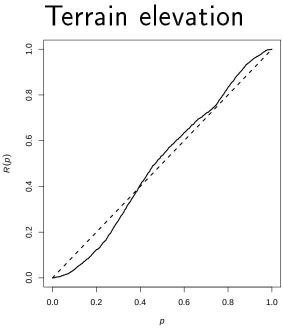
✓The geological survey region can be efficiently/usefully divided into subregions of high and low density of gold deposits, by specifying a threshold on the distance to the nearest major geological fault.

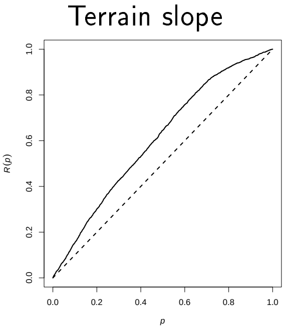
Fun fact: for the ROC curve based on a covariate Z,

$$AUC = \mathbb{P}\{Z(X) > Z(Y)\}$$

where X, Y are independent,
X is a randomly-selected **data point** (gold deposit),
Y is a randomly-selected **spatial location** in the study region.

Rainforest
ROC curves based on covariates





Interpretation:

In the rainforest study rectangle,

Interpretation:

In the rainforest study rectangle,

- × higher terrain elevations are **not** associated with higher densities of trees;

Interpretation:

In the rainforest study rectangle,

- × higher terrain elevations are **not** associated with higher densities of trees;
- ✓ steeper terrain slopes are *slightly* associated with higher densities of trees;

Interpretation:

In the rainforest study rectangle,

- × higher terrain elevations are **not** associated with higher densities of trees;
- ✓ steeper terrain slopes are *slightly* associated with higher densities of trees;
- ⚠ "reading" the ROC curve is complicated!

6. Dependence on a covariate

6. Dependence on a covariate

- How does forest density depend on terrain slope?

6. Dependence on a covariate

- How does forest density depend on terrain slope?
- How does presence of gold depend on proximity to faults?

6. Dependence on a covariate

- How does forest density depend on terrain slope?
- How does presence of gold depend on proximity to faults?

Suppose that the probability of presence p is a function of the covariate z,

$$p = \rho(z)$$

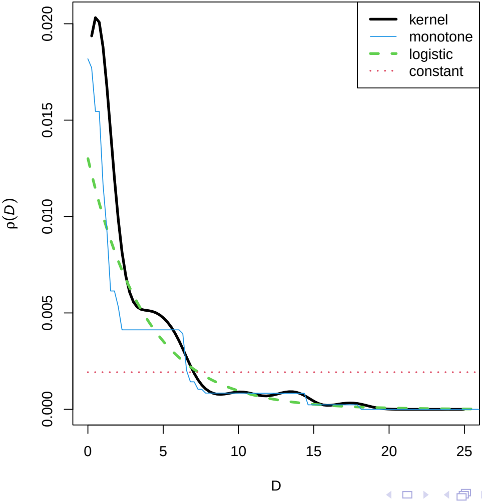
$\boxed{\rho(z) \text{ can be estimated from data}}$

$\rho(z)$ can be estimated parametrically ("species distribution model") or non-parametrically ("resource selection function").

$\rho(z)$ can be estimated from data

$\rho(z)$ can be estimated parametrically ("species distribution model") or non-parametrically ("resource selection function").

Geological survey, $z = D =$ distance to nearest fault:



$\boxed{\rho(z) \text{ is a "law"}}$

While ROC depends critically on the choice of study region, $\rho(z)$ does not: the equation

$$p = \rho(z)$$

is a "relation", "model" or "law" that could be extrapolated from one region to another.

The function $\rho(z)$ is directly interpretable.

What is the relationship between $\rho(z)$ *and the ROC for* z *?*

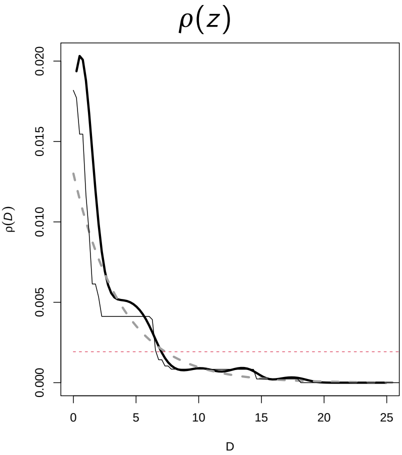
ρ is proportional to the slope of the ROC curve

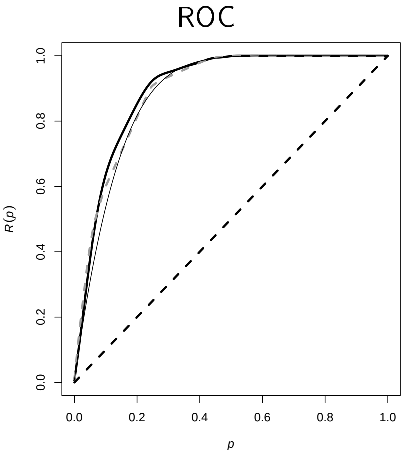
If the ROC curve is a function $p \mapsto R(p)$ for $0 \le p \le 1$, then

$$\rho(z) = \kappa \frac{\mathrm{d}}{\mathrm{d}p} R(p) \qquad \text{where} \quad p = \mathrm{FP}(z),$$

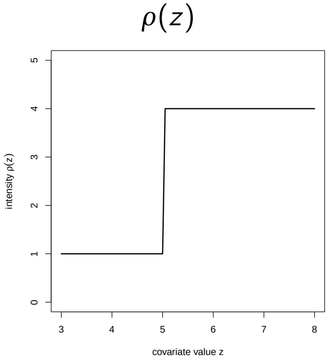
where κ is the average probability of presence.

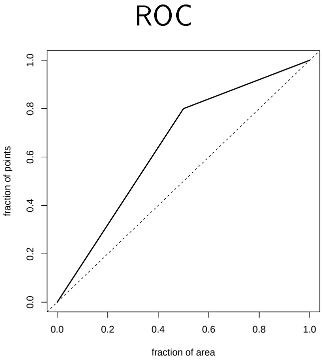
Geological survey, distance to nearest fault





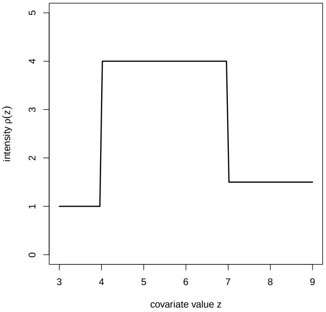
Corresponding shapes



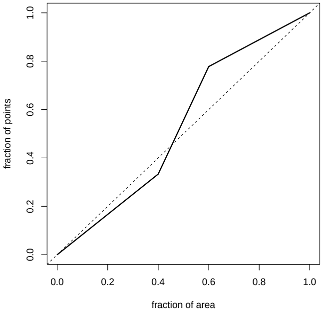


Corresponding shapes

$\rho(z)$



ROC



$\boxed{\text{The shape of the ROC curve is diagnostic}}$

$\boxed{\text{The shape of the ROC curve is diagnostic}}$

If the ROC curve is **concave**, then

$\boxed{\text{The shape of the ROC curve is diagnostic}}$

If the ROC curve is **concave**, then

✓ $\rho(z)$ is an increasing function of z

The shape of the ROC curve is diagnostic

If the ROC curve is **concave**, then

- ✓ $\rho(z)$ is an increasing function of z
- ✓ the most efficient way to segregate the region into high and low densities is to threshold the covariate z

The shape of the ROC curve is diagnostic

If the ROC curve is **concave**, then

- ✓ $\rho(z)$ is an increasing function of z
- ✓ the most efficient way to segregate the region into high and low densities is to threshold the covariate z
  (by the Neyman-Pearson Lemma)

The shape of the ROC curve is diagnostic

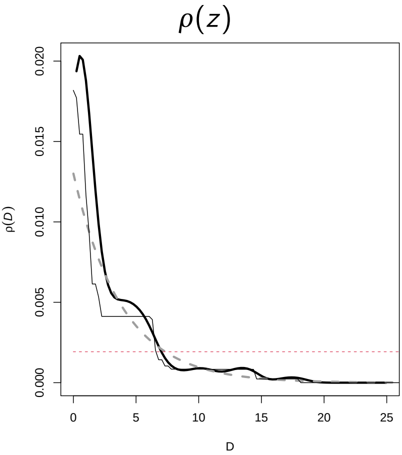
If the ROC curve is **concave**, then

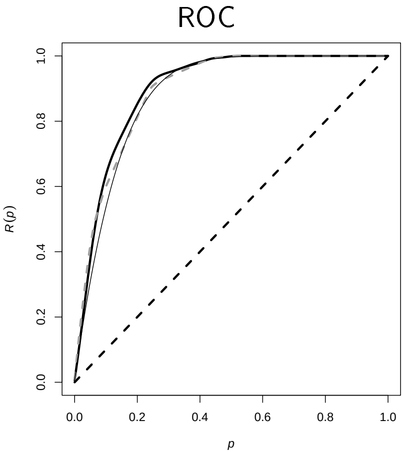
- ✓ $\rho(z)$ is an increasing function of z
- ✓ the most efficient way to segregate the region into high and low densities is to threshold the covariate z

  (by the Neyman-Pearson Lemma)
- ✓ the ROC and AUC are appropriate summaries 👍

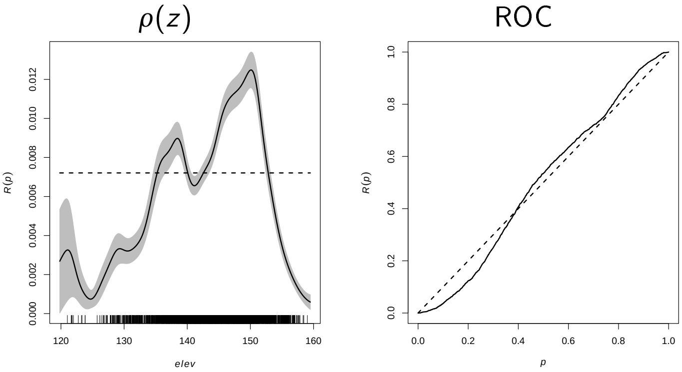
If the ROC curve is **not concave**, thresholding the covariate z is not optimal for predicting presence/absence.

Geological survey, distance to nearest fault

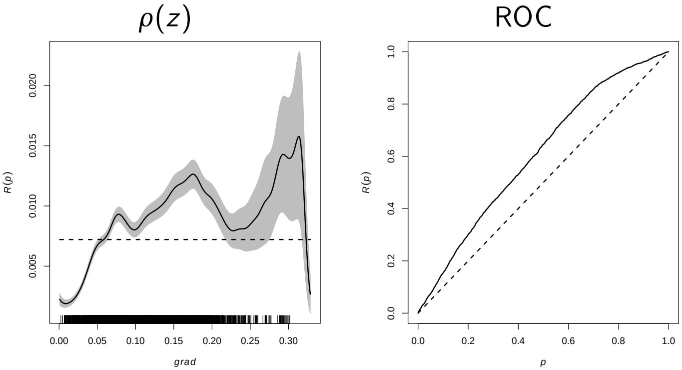




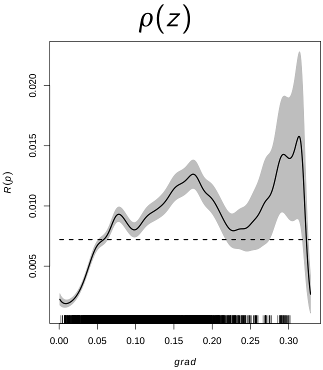
Rainforest, terrain elevation

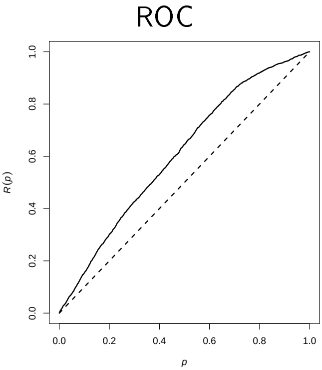


Rainforest, terrain slope



Rainforest, terrain slope





To decide whether $\rho(z)$ is an increasing function of z, it may be safer to use the ROC curve, which is not affected by smoothing artefacts.

7. Other ways to use ROC

7. Other ways to use ROC

As originally defined, the ROC curve is a comparison between two probability distributions

As originally defined, the ROC curve is a comparison between two probability distributions (the distribution of the discriminant S in the Positive and Negative populations).

7. Other ways to use ROC

As originally defined, the ROC curve is a comparison between two probability distributions (the distribution of the discriminant S in the Positive and Negative populations).

In applications to spatial data, **"the"** ROC curve has been interpreted narrowly:

7. Other ways to use ROC

As originally defined, the ROC curve is a comparison between two probability distributions (the distribution of the discriminant S in the Positive and Negative populations).

In applications to spatial data, **"the"** ROC curve has been interpreted narrowly:

- S = fitted probability of presence

7. Other ways to use ROC

As originally defined, the ROC curve is a comparison between two probability distributions (the distribution of the discriminant S in the Positive and Negative populations).

In applications to spatial data, **"the"** ROC curve has been interpreted narrowly:

- S = fitted probability of presence
- Positive "population" = observed presence pixels

7. Other ways to use ROC

As originally defined, the ROC curve is a comparison between two probability distributions (the distribution of the discriminant S in the Positive and Negative populations).

In applications to spatial data, **"the"** ROC curve has been interpreted narrowly:

- S = fitted probability of presence
- Positive "population" = observed presence pixels
- Negative "population" = observed absence pixels

There are many other potential uses of ROC curves based on different choices of S and the two "populations".

"Traditional" ROC for spatial model

- S = fitted probability of presence $\widehat{p}_j$
- Positive "population" = observed presence pixels
- Negative "population" = observed absence pixels

$$
\begin{aligned}
\mathrm{TP}(t) &= \frac{\sum_j y_j \, 1\{\widehat{p}_j > t\}}{\sum_j y_j} \\
\mathrm{FP}(t) &= \frac{\sum_j (1 - y_j) 1\{\widehat{p}_j > t\}}{\sum_j (1 - y_j)}
\end{aligned}
$$

"Traditional" ROC for spatial model

- S = fitted probability of presence $\widehat{p}_j$
- Positive "population" = observed presence pixels
- Negative "population" = observed absence pixels

$$\begin{aligned} \mathrm{TP}(t) &= \frac{\sum_j y_j \, 1\{\widehat{p}_j > t\}}{\sum_j y_j} \\ \mathrm{FP}(t) &= \frac{\sum_j (1 - y_j) 1\{\widehat{p}_j > t\}}{\sum_j (1 - y_j)} \end{aligned}$$

Recommendation: calculate $\widehat{p}_j$ using leave-one-out estimate

Empirical ROC based on a spatial covariate Z

- S = value of covariate Z
- Positive "population" = observed presence pixels
- Negative "population" = observed absence pixels

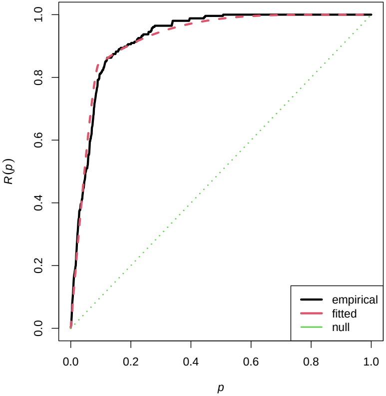
$$
\begin{aligned}
\mathrm{TP}(t) &= \frac{\sum_j y_j \, 1\{z_j > t\}}{\sum_j y_j} \\
\mathrm{FP}(t) &= \frac{\sum_j (1 - y_j) 1\{z_j > t\}}{\sum_j (1 - y_j)}
\end{aligned}
$$

Predicted ROC of spatial model

- S = fitted probability of presence $\widehat{p}_j$
- Positive population = all pixels, weight $\propto \widehat{p}_j$
- Negative population = all pixels, weight $\propto (1 - \widehat{p}_j)$

$$
\begin{aligned}
\mathrm{TP}(t) &= \frac{\sum_j \widehat{p}_j \, 1\{\widehat{p}_j > t\}}{\sum_j \widehat{p}_j} \\
\mathrm{FP}(t) &= \frac{\sum_j (1 - \widehat{p}_j) 1\{\widehat{p}_j > t\}}{\sum_j (1 - \widehat{p}_j)}
\end{aligned}
$$

Geological survey
Logistic regression on distance and greenstone

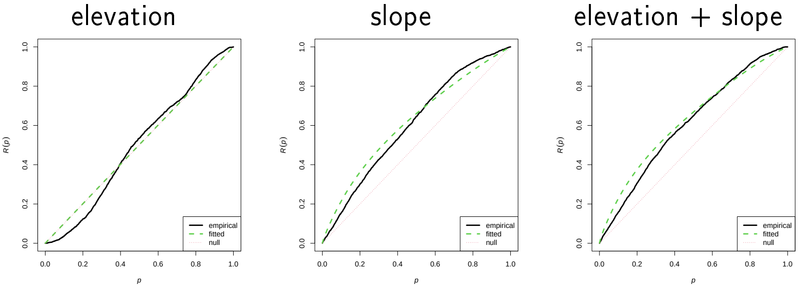


$\boxed{\text{The \textbf{predicted ROC} of a fitted spatial model is always concave.}}$

The **predicted ROC** of a fitted spatial model is always concave.

Discrepancies between the shapes of the empirical and predicted ROC curve suggest the model is inadequate.

Rainforest
Logistic regressions
Empirical and predicted ROC curves



Partial ROC

After fitting a model and computing predicted presence probabilities $\tilde{p}_j$, consider adding a new variable Z to the model.

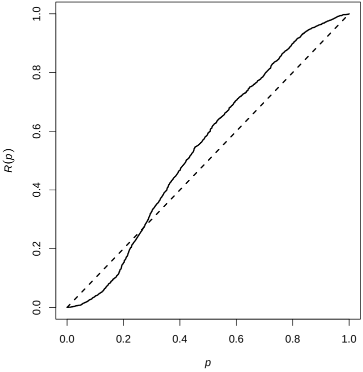
- S = value of new covariate z_j
- Positive population = observed presence pixels
- Negative population = all pixels, weight $\propto \tilde{p}_j$

$$
\begin{aligned}
\mathrm{TP}(t) &= \frac{\sum_j y_j \, 1\{z_j > t\}}{\sum_j y_j} \\
\mathrm{FP}(t) &= \frac{\sum_j \tilde{p}_j 1\{z_j > t\}}{\sum_j \tilde{p}_j}
\end{aligned}
$$

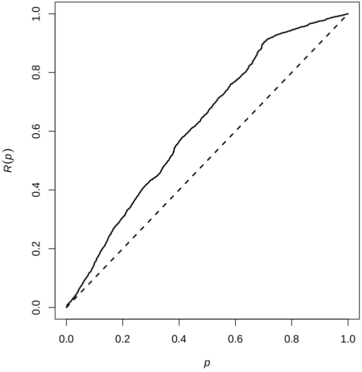
The partial ROC indicates the “benefit” of adding the variable Z to the existing model.

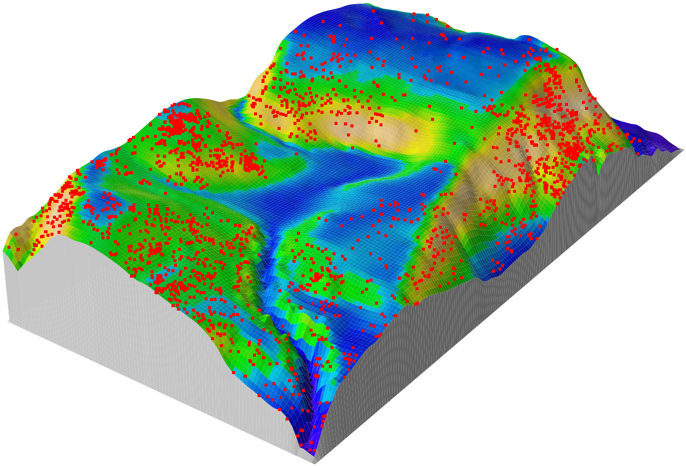
Rainforest
Logistic regressions
Partial ROC curves for adding a covariate

regression on **slope**
add variable: **elevation**



regression on **elevation**
add variable: **slope**





Colour = probability predicted by logistic regression on **slope**

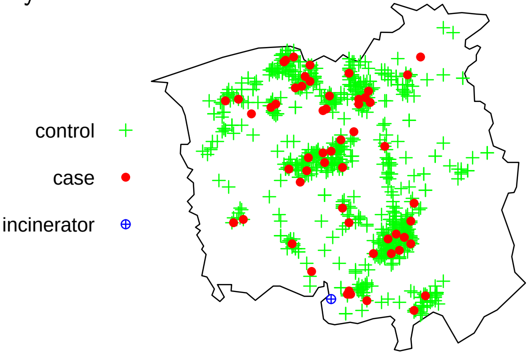
ROC for spatial case-control data

A spatial case-control dataset consists of a point pattern of **"cases"** and a point pattern of **"controls"** in the same study region.

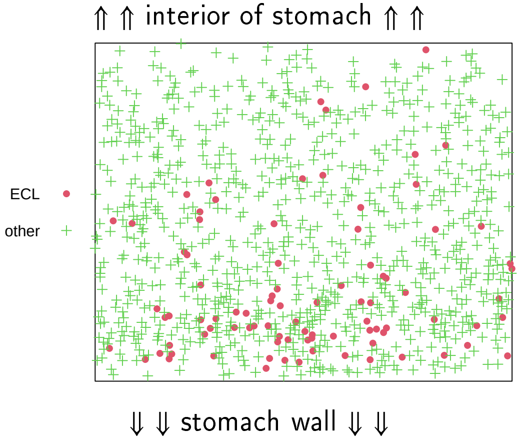
ROC for spatial case-control data

A spatial case-control dataset consists of a point pattern of **"cases"** and a point pattern of **"controls"** in the same study region.

Cancer registry data



Stomach cells



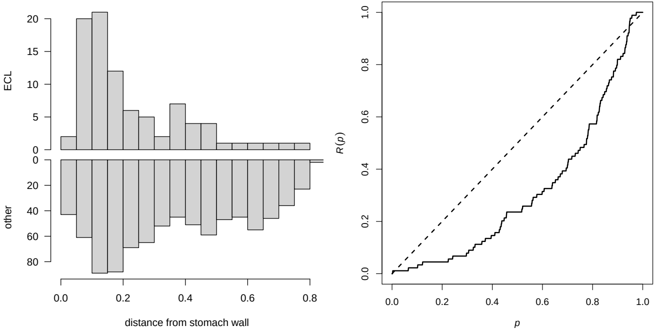
ROC for spatial case-control data

For a spatial covariate z, create the ROC with

- S = value of covariate z_j
- Positive population = cases
- Negative population = controls

ROC for spatial case-control data

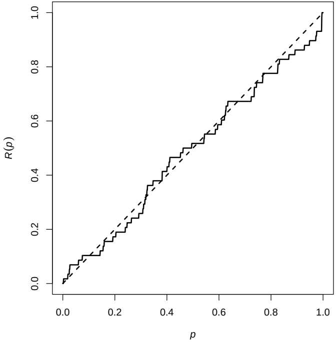
Stomach cells, distance from stomach wall (≡ vertical coordinate)



AUC = 0.32

ROC for spatial case-control data

Cancer registry, distance to incinerator



AUC = 0.49 😐

8. What does AUC measure?

AUC = Area Under the ROC Curve

Some writers claim that "AUC is a measure of **goodness-of-fit** of the fitted model", in the sense that a **large** value of AUC indicates that the model is a **good fit** to the data.

A goodness-of-fit test is a hypothesis test of

$$\begin{array}{rl} & H_0 : \text{model is true} \\ \text{vs} & \\ & H_1 : \text{model is false} \end{array}$$

A **large** value of the test statistic would cause us to **reject** H_0 and conclude that the model **does not fit** the data.

Berman, Lawson and Waller developed hypothesis tests to decide whether probability of presence depends on a spatial covariate Z. They are goodness-of-fit tests of

$$H_0 : \mathbb{P}\{\text{presence}\} \text{ is constant}$$

against the one-sided alternative

$$H_1 : \mathbb{P}\{\text{presence}\} \text{ is an increasing function of } Z$$

Berman, Lawson and Waller developed hypothesis tests to decide whether probability of presence depends on a spatial covariate Z. They are goodness-of-fit tests of

$$H_0 : \mathbb{P}\{\text{presence}\} \text{ is constant}$$

against the one-sided alternative

$$H_1 : \mathbb{P}\{\text{presence}\} \text{ is an increasing function of } Z$$

Berman's "Z_2 test" rejects H_0 if $T > t$, where the test statistic T turns out to be

$$T = \sqrt{12n}\left(\text{AUC} - \frac{1}{2}\right)$$

where n is the number of presence pixels or data points, and AUC is calculated for the ROC curve based on Z.

Berman, Lawson and Waller developed hypothesis tests to decide whether probability of presence depends on a spatial covariate Z. They are goodness-of-fit tests of

$$H_0 : \mathbb{P}\{\text{presence}\} \text{ is constant}$$

against the one-sided alternative

$$H_1 : \mathbb{P}\{\text{presence}\} \text{ is an increasing function of } Z$$

Berman's "Z_2 test" rejects H_0 if $T > t$, where the test statistic T turns out to be

$$T = \sqrt{12n}\left(\text{AUC} - \frac{1}{2}\right)$$

where n is the number of presence pixels or data points, and AUC is calculated for the ROC curve based on Z.

That is, AUC is a measure of **badness-of-fit** of the **null** model of uniform probability of presence.

AUC is

- a measure of **badness-of-fit** of the null model of uniform probability of presence

AUC is

- a measure of **badness-of-fit** of the null model of uniform probability of presence
- not adjusted for sample size

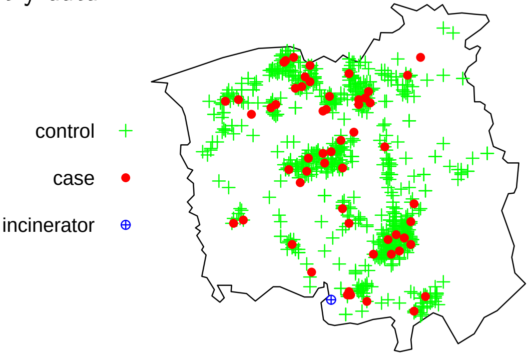
AUC is

- a measure of **badness-of-fit** of the null model of uniform probability of presence
- not adjusted for sample size
- analogous to a measure of **effect size** summarising the ranking/segregating ability of the covariate or fitted model.

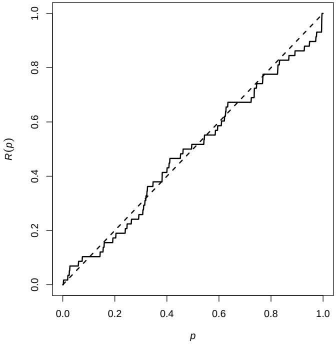
AUC is

- a measure of **badness-of-fit** of the null model of uniform probability of presence
- not adjusted for sample size
- analogous to a measure of **effect size** summarising the ranking/segregating ability of the covariate or fitted model.
- an **aggregate** over the whole population; insensitive to effects occurring in small sub-populations

Cancer registry data

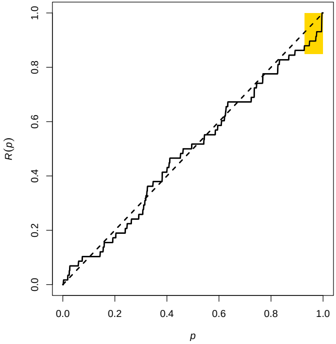


Cancer registry, distance to incinerator



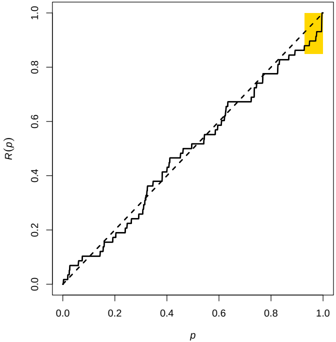
AUC = 0.49

Cancer registry, distance to incinerator



AUC = 0.49

Cancer registry, distance to incinerator



AUC = 0.49

Proximity to the incinerator causes a statistically significant increase in cancer risk even though it only affects a small fraction of the population.

Diggle & Rowlingson (1994)

Conclusions

ROC and AUC

- × do not measure goodness-of-fit
- × do not measure predictive performance
- ✓ do measure “ranking”/ “segregating” ability
- ✓ do contain diagnostic information
- ⚠ are bound to the study region
- ⚠ are insensitive to details of the fitted model
- ✓ are useful for variable selection
- 💡 can be modified/extended to serve many useful purposes

References

A. Baddeley, E. Rubak, S. Rakshit, G. Nair (2023)
ROC curves for spatial point patterns and presence-absence data.
In preparation.

A. Baddeley et al (2021)
Optimal thresholding of predictors in mineral prospectivity analysis.
Natural Resources Research **30**, 923–969

P. Diggle, B. Rowlingson (1994)
A conditional approach to point process modelling of elevated risk.
J Roy Statist Soc A **157**, 443–440.

J. Franklin (2009)
Mapping Species Distributions: Spatial Inference and Prediction
Cambridge University Press

W. Krzanowski, D. Hand (2009)
ROC Curves for Continuous Data
Chapman and Hall/CRC

References

A. Baddeley, E. Rubak, S. Rakshit, G. Nair (2023)
ROC curves for spatial point patterns and presence-absence data.
In preparation.

A. Baddeley et al (2021)
Optimal thresholding of predictors in mineral prospectivity analysis.
Natural Resources Research **30**, 923–969

P. Diggle, B. Rowlingson (1994)
A conditional approach to point process modelling of elevated risk.
J Roy Statist Soc A **157**, 443–440.

J. Franklin (2009)
Mapping Species Distributions: Spatial Inference and Prediction
Cambridge University Press

W. Krzanowski, D. Hand (2009)
ROC Curves for Continuous Data
Chapman and Hall/CRC

adrian.baddeley@curtin.edu.au