University of Kent NERC Innovation Workshop

New developments in the statistical modelling of presence/absence data

Rachel McCrea, Byron Morgan and Martin Ridout

NERC SCIENCE OF THE ENVIRONMENT

EPSRC
Engineering and Physical Sciences Research Council

ncse
National Centre for Statistical Ecology

University of Kent

Outline of the workshop

- Introductory Tutorial (Rachel)
- Covariates: Examples involving butterflies and newts (Byron)
- Break
- Abundance estimation using presence/absence data (Martin)
- Models for spatial replication (Martin)
- Occupancy as a hidden Markov model (Byron)

University of Kent NERC Innovation Workshop

New developments in the statistical modelling of presence/absence data

An Introductory Tutorial

Rachel McCrea

NERC SCIENCE OF THE ENVIRONMENT

EPSRC Engineering and Physical Sciences Research Council

ncse National Centre for Statistical Ecology

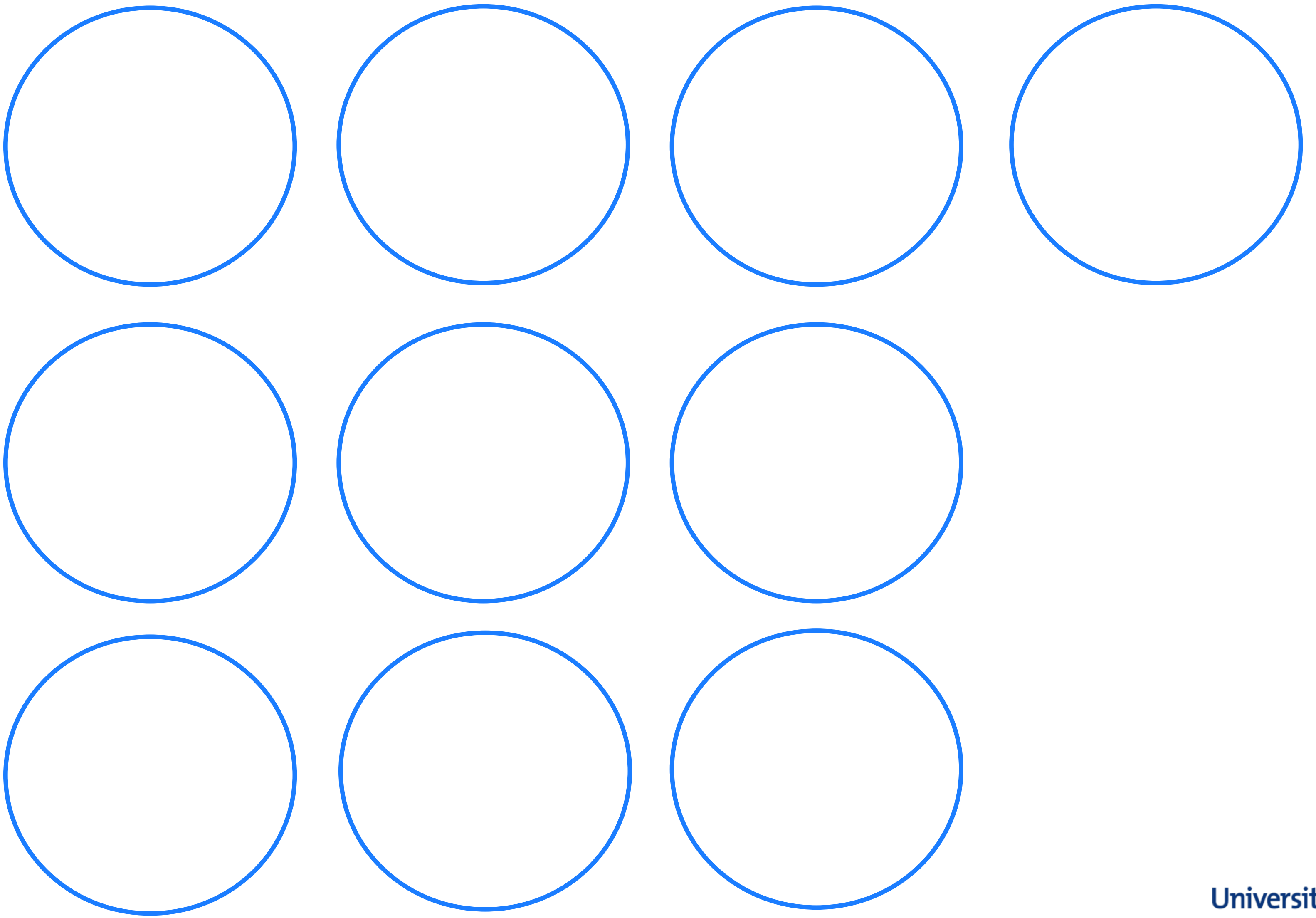
University of Kent

Overview

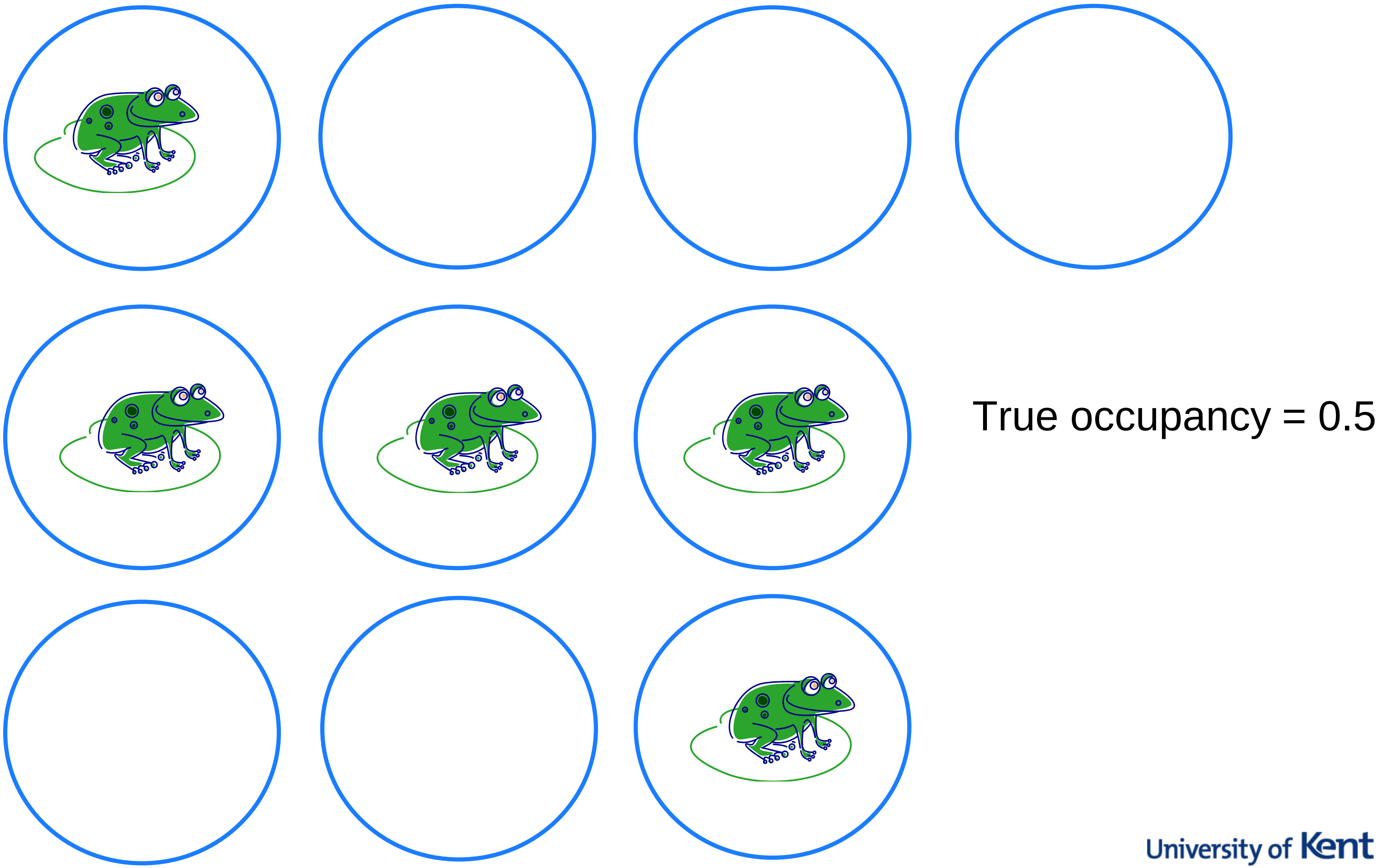
- Introduction
- Aside: likelihood theory
- Simple occupancy models
- Model selection
- Advanced occupancy models
- Computer software

- Further resources

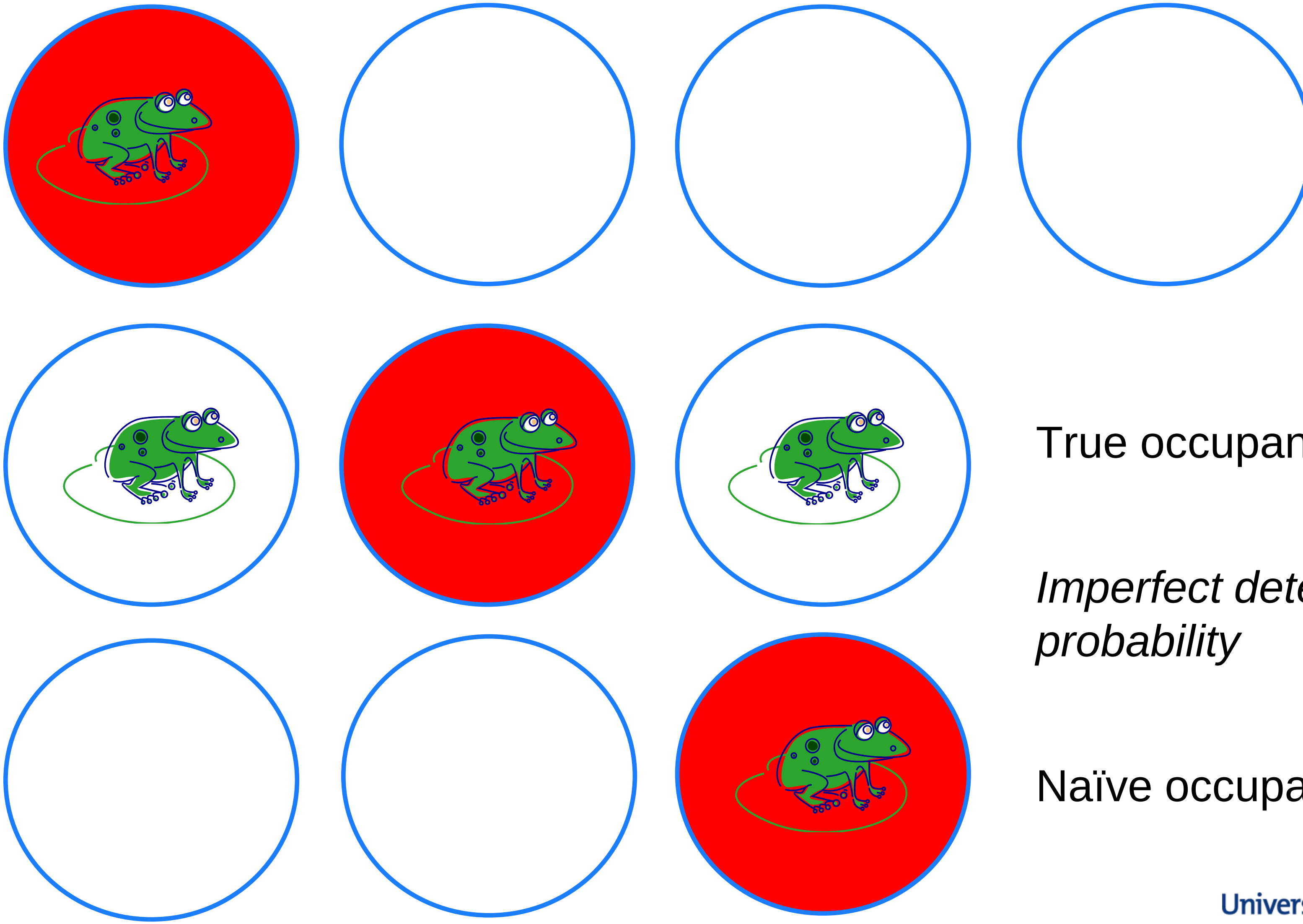
Do we need to account for detectability?



Do we need to account for detectability?



Do we need to account for detectability?



True occupancy = 0.5

Imperfect detection probability

Naïve occupancy = 0.3

Multinomial probabilities

- Suppose we have three buckets labelled A, B and C



- We try and throw N balls into the buckets.

Multinomial probabilities

- There are 4 possible outcomes for each of the throws:
  - Ball goes into bucket A
  - Ball goes into bucket B
  - Ball goes into bucket C
  - Ball doesn’t go into any of the buckets

Multinomial probabilities

- There are 4 possible outcomes for each of the throws:
  - Ball goes into bucket A
  - Ball goes into bucket B
  - Ball goes into bucket C
  - Ball doesn't go into any of the buckets

n_A n_B n_C

$$N - n_A - n_B - n_C$$

Multinomial likelihood

- Suppose the probability of getting the ball in bucket i is p_i
- We want to find the values of the parameters p_A, p_B, p_C which ***maximises the likelihood*** that we would observe the data that we did (n_A,n_B,n_C).

$$\mathcal{L}(p_A, p_B, p_C | N, n_A, n_B, n_C)$$
$$\propto {p_A}^{n_A} {p_B}^{n_B}\ {p_C}^{n_C} \times (1 - p_A - p_B - p_C)^{N - n_A - n_B - n_C}$$

- Note here we assume N is known.

Multinomial likelihood

- Suppose the probability of getting the ball in bucket i is p_i
- We want to find the values of the parameters p_A, p_B, p_C which ***maximises the likelihood*** that we would observe the data that we did (n_A,n_B,n_C).

$$\mathcal{L}(p_A, p_B, p_C | N, n_A, n_B, n_C)$$

$$\propto {p_A}^{n_A} {p_B}^{n_B} \, {p_C}^{n_C} \times (1 - p_A - p_B - p_C)^{N - n_A - n_B - n_C}$$

- Note here we assume N is known.

Multinomial likelihood

- Suppose the probability of getting the ball in bucket i is p_i
- We want to find the values of the parameters p_A, p_B, p_C which ***maximises the likelihood*** that we would observe the data that we did (n_A,n_B,n_C).

$$\mathcal{L}(p_A, p_B, p_C | N, n_A, n_B, n_C)$$

$$\propto {p_A}^{n_A} {p_B}^{n_B} \, {p_C}^{n_C} \times (1 - p_A - p_B - p_C)^{N - n_A - n_B - n_C}$$

- Note here we assume N is known.

Simple occupancy model

- Visit sites i = 1, …, S
- Multiple surveys j = 1, …, K
  - Can be temporal, or may be different teams conducting surveys
- Observed data: h_i
  - Detection history for each site
- Examples:
  - `0101`
  - `1110`
  - `0000`
  - …

Simple occupancy model

- Conceptual model:
  - A site may be occupied or not
  - If the site is occupied, there is some probability of detecting the species

Formalising the model

- Parameters:
  - ψ: probability the site is occupied
  - p_j: probability species is detected at survey j

- Construct probabilities

$$\Pr(h_i = 0101) = \psi(1-p_1)p_2(1-p_3)p_4$$

$$\Pr(h_i = 0000) = \psi(1-p_1)(1-p_2)(1-p_3)(1-p_4)+(\mathbf{1-}\psi)$$

University of Kent

Alternative models

- Constant detection probability
- Relate detection to covariate values
- Relate occupancy to covariate values
- Incorporate heterogeneity (finite and infinite mixtures)

- How do we select between these alternative models?

Model selection: AIC

- AIC can be calculated for each of the fitted models.
- AIC is a measure of the relative quality of a statistical model for a given set of data.
- It is a trade-off between how well the model fits the data and the complexity (number of parameters) of the model
- The smaller the AIC, the more support for the model
- Suppose you want to select between T candidate models:
  - $\Delta AIC_i = AIC_i - \min(AIC_1, \ldots, AIC_T)$

Example: Blue-ridge two-lined salamanders

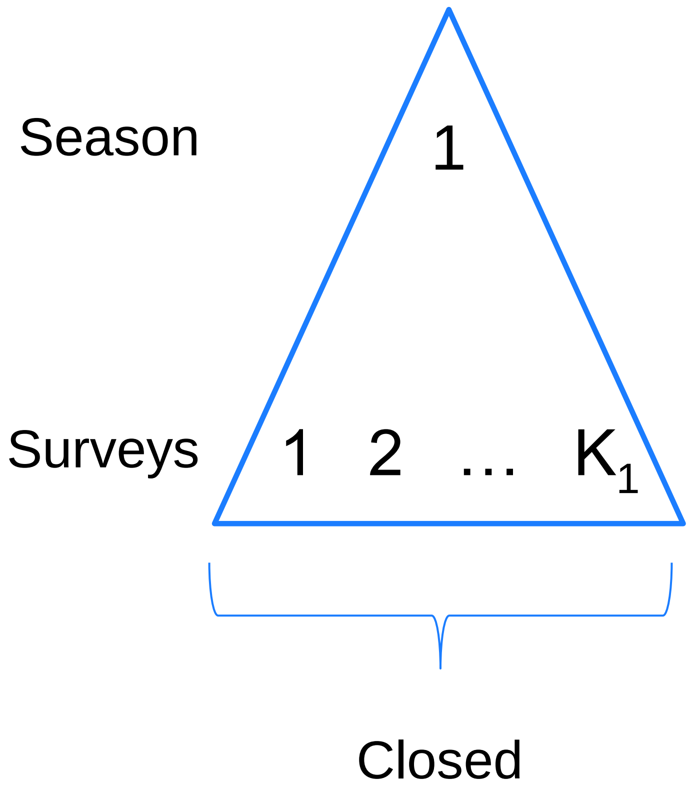
- MacKenzie et al (2006) p. 99
- s = 39 (number of sites)
- K = 5 (number of surveys)
- Two candidate models:
  - Occupancy and detection are constant
  - Constant occupancy, time-dependent detection
- Salamanders were detected at 18 of the 39 sites
  - Naïve occupancy estimate = 18/39 = 0.46

Results

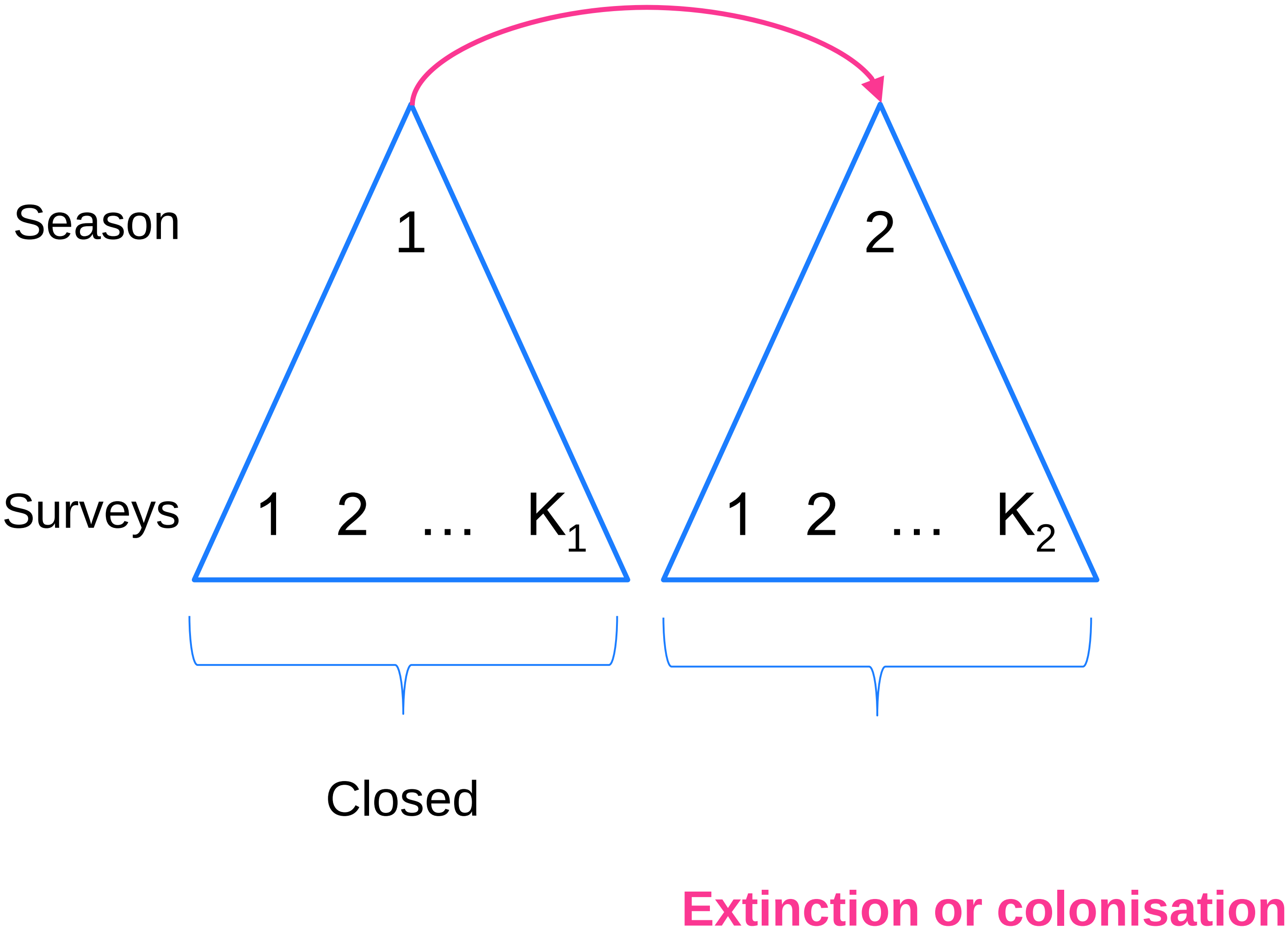
Model	ΔAIC	np	$\widehat{\psi}$	$\widehat{p_1}$	$\widehat{p_2}$	$\widehat{p_3}$	$\widehat{p_4}$	$\widehat{p_5}$
ψ(.),p(.)	0.00	2	0.60	0.26	0.26	0.26	0.26	0.26
ψ(.),p(t)	1.95	6	0.58	0.18	0.13	0.40	0.35	0.27

- Probability of false absence?
- Which model is best?

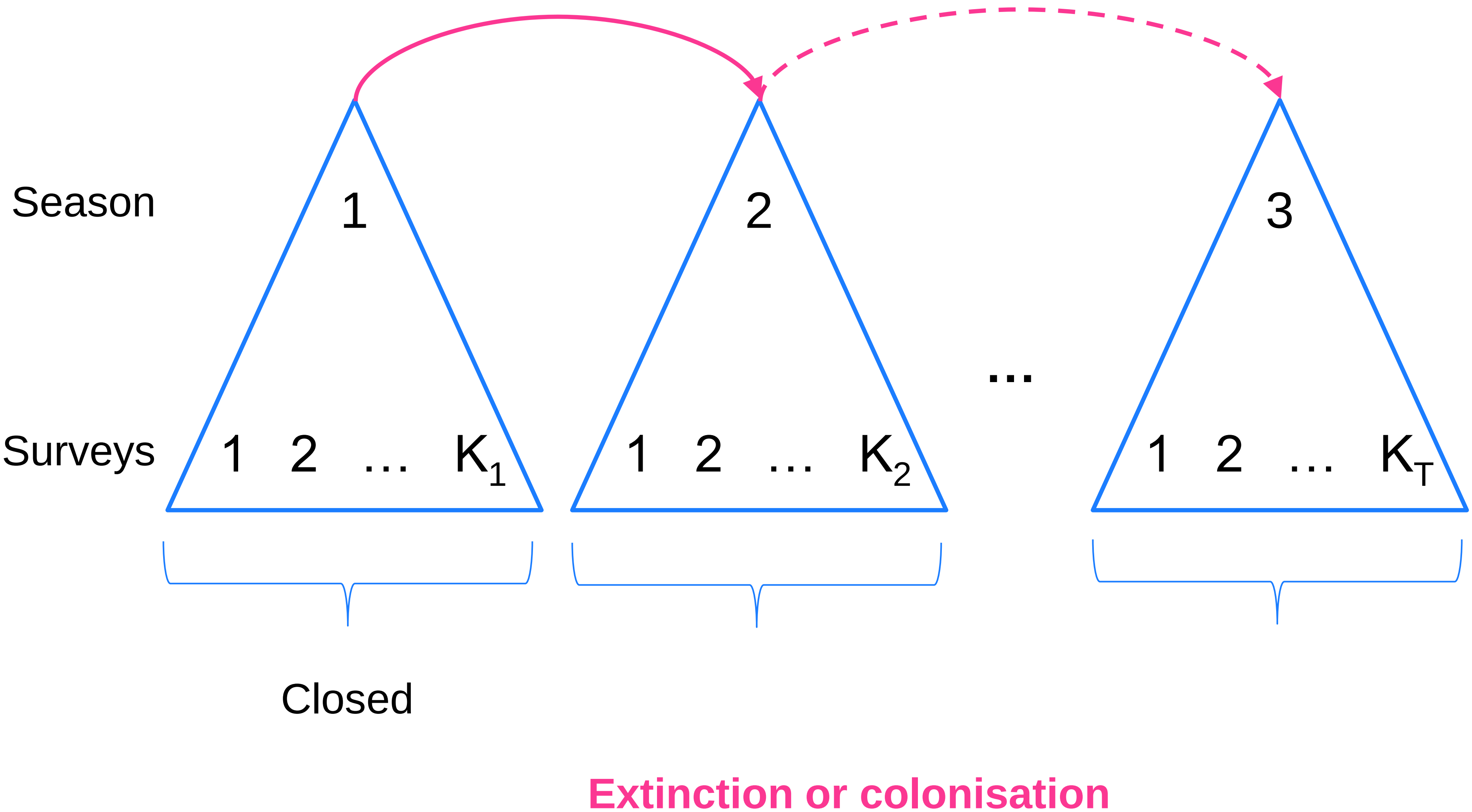
Advanced models: multiple-season models



Advanced models: multiple-season models



Advanced models: multiple-season models



Advanced models: multiple-season models

- Example detection history:

110 000 010

- New Parameters

COLONISATION

- γ_t: the probability that an unoccupied site in season t is occupied by the species in season t+1

EXTINCTION

- ε_t: the probability that a site occupied in season t is unoccupied by the species in season t+1

Advanced models: multiple-season models

- Example detection history:

110   000   010

- Extended Parameters
  - ψ_t: probability a site is occupied in season t
  - p_{tj}: probability of detecting the speices in the j^{th} survey of a site during season t

Example: Northern spotted owl

- MacKenzie et al (2006), p. 209
- s = 55 potential breeding territories
- Surveyed between 1997 and 2000 (T=5)
- $K_{average}$=5.3; K_{max}=8

Competing models

- Occupancy status does not change
  - $\psi(.)$,p(year)
- Random changes in occupancy (no dependence on whether previously occupied)
  - $\psi(1997)$, $\varepsilon=(1-\gamma)$, p(year)
- Markovian changes in occupancy
  - $\psi(1997)$, $\varepsilon(.)$, $\gamma(.)$, p(year)
- Constant occupancy and colonisation
  - $\psi(.)$, $\gamma(.)$, p(year)
  - ε: derived parameter determined from the dynamic process

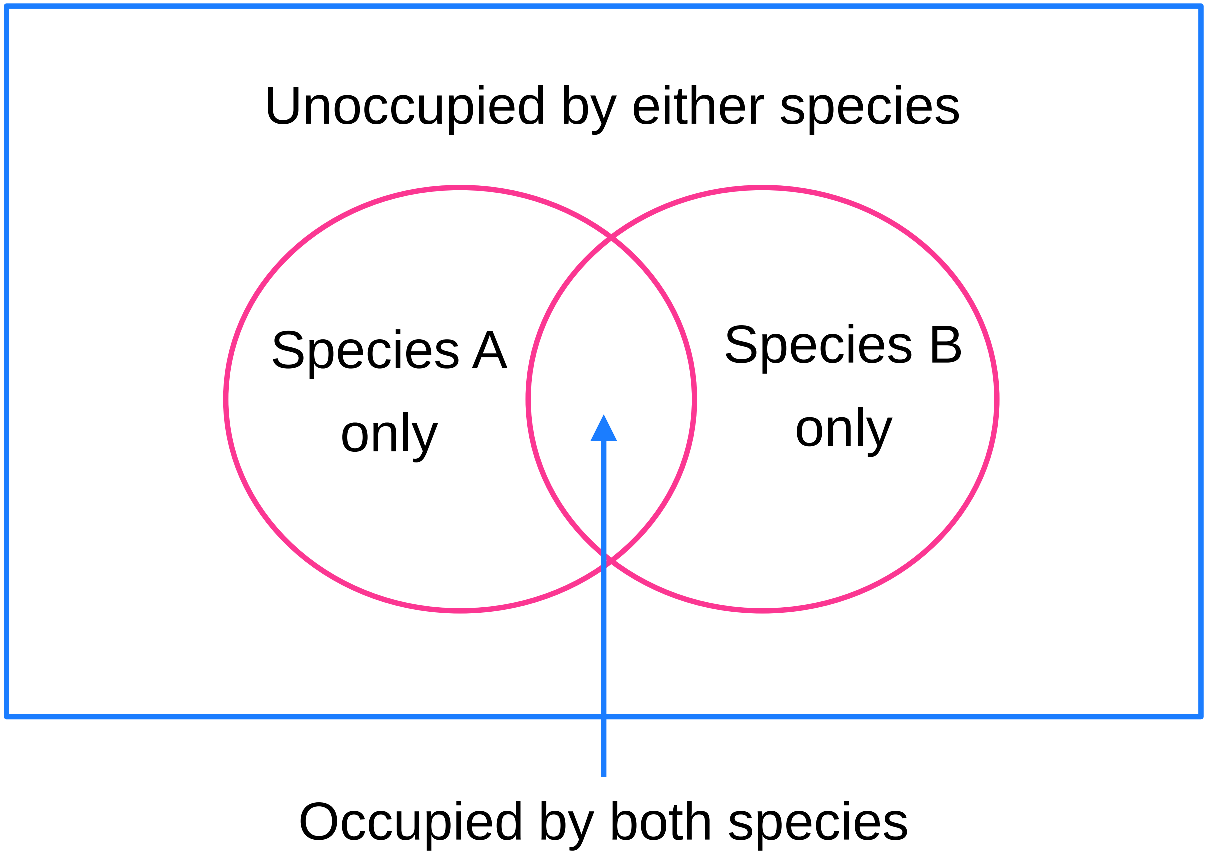
Results

Model	ΔAIC	np
ψ(.),γ(.),p(year)	0.00	7
ψ(1997),γ(.), ε(.), p(year)	1.57	8
ψ(1997),γ(year), ε(year), p(year)	3.69	14
ψ(1997),γ(.), {ε=1-γ}, p(year)	91.58	7
ψ(1997),γ(year), {ε=1-γ}, p(year)	97.37	10
ψ(.),p(year)	202.61	6

- Changes in occupancy best represented by Markov process
- Equilibrium state (no year-dependence in colonisation or extinction)

Advanced models: multiple species models

- Species interactions

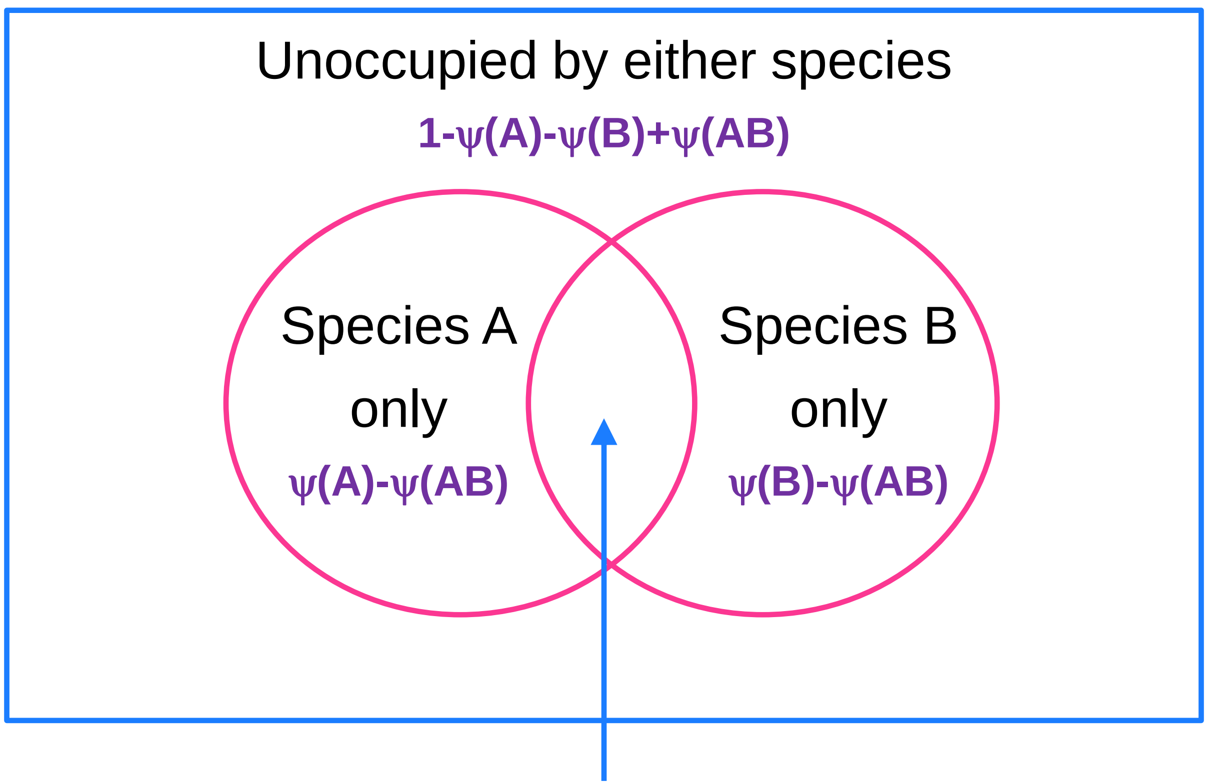


Advanced models: multiple species models

- Species interactions:
- ψ(A): probability species A occupies a site
- ψ(B): probability species B occupies a site
- ψ(AB): probability both species occupy a site site

Advanced models: multiple species models

- Species interactions



Occupied by both species ψ(AB)

Advanced models: multiple species models

- ψ(A): probability species A occupies a site
- ψ(B): probability species B occupies a site
- ψ(AB): probability both species occupy a site site

- p_j(A): probability of detecting species A during the j^{th} survey, given only species A is present

- p_j(B): probability of detecting species B during the j^{th} survey, given only species B is present

Advanced models: multiple species models

- $r_j(AB)$: probability of detecting both species during j^{th} survey, given both are present
- $r_j(Ab)$: probability of detecting species A but not B during j^{th} survey, given both are present
- $r_j(aB)$: probability of detecting species B but not A during j^{th} survey, given both are present
- $r_j(ab)$: probability of detecting neither species during j^{th} survey, given both are present

- $r_j(ab) = 1 - r_j(AB)-r_j(Ab)-r_j(aB)$

Advanced models: multiple species models

- Depending on parameters of inetrest, there are reparameterised forms:

- Species interaction factor

$$\varphi = \frac{\psi(AB)}{\psi(A)\psi(B)}$$

- “how much more or less likely the species are to co-occur at a site compared to what would be expected if they co-occurred independenly”

Computer software: Presence

- Presence can be downloaded here:
- http://www.mbr-pwrc.usgs.gov/software/presence.html
- The same webpage has a manual and tutorials to help you get started with the software.
- The citation for Presence is:
- Hines, J. E. (2006). *PRESENCE2 - Software to estimate patch occupancy and related parameters.* USGS-PWRC. http://www.mbr-pwrc.gov/software/presence.html

Computer software: unmarked in R

- R package `unmarked` can be used to fit occupancy models
- Details of unmarked can be found here:
  - http://cran.r-project.org/package=unmarked
- Fiske and Chandler (2011) unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software*. **43**, 1-23

Useful References

- MacKenzie, Nichols, Royle, Pollock, Bailey and Hines (2006) *Occupancy Estimation and Modeling: Inferring patterns and dynamics of species occurrence.* Academic Press.
- Guillera-Arroita, Ridout and Morgan (2010) Design of occupancy studies with imperfect detection. *Methods in Ecology and Evolution.* 1, 131-139

- Gurutzeta Guillera-Arroita's website and blog:
- https://gguilleraresearch.wordpress.com/