



Workshop presencial:

Mapeamento probabilístico da distribuição de espécies baseado na integração de dados

André L. Luza – Universidade Federal de Santa Maria (luza.andre@gmail.com)

Viviane Zulian – North Carolina State University (zulian.vi@gmail.com)

28 a 31 de julho de 2022.



Sumário - Sexta-feira, 29 de julho

- O que são modelos de distribuição de espécies
- Tipos de modelos
- Métodos para estimativa de parâmetros
- Diferenças entre abordagem frequentista e bayesiana
- Prática em R e Jags
- Exercício



Por que modelar distribuição de espécies é importante?

Por que modelar distribuição de espécies é importante?

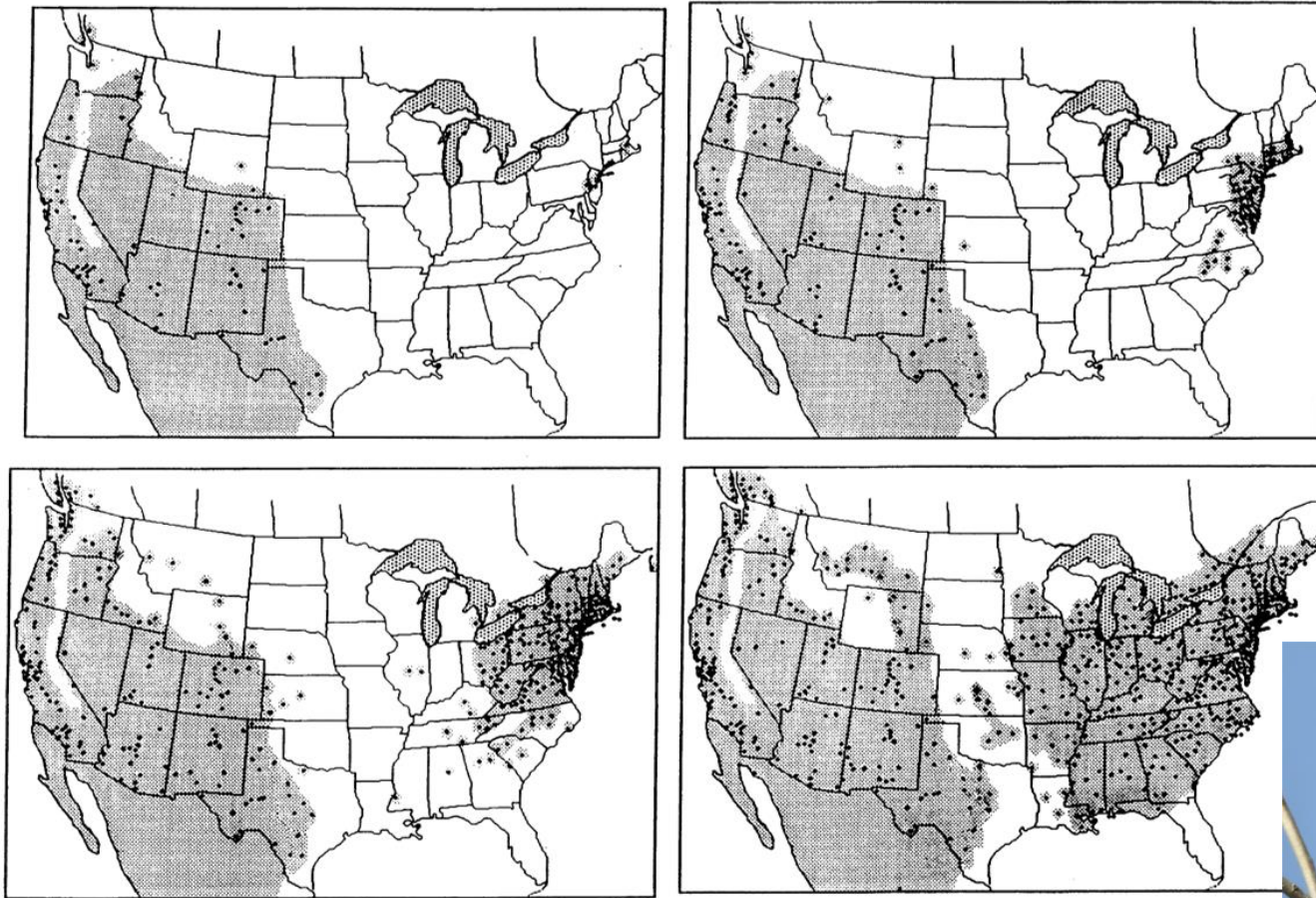
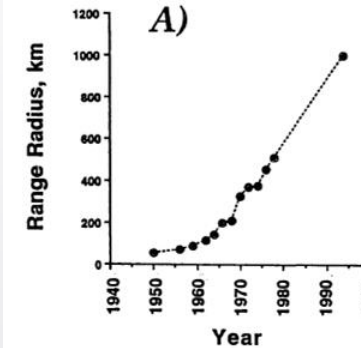


FIG. 1.—Spread of the house finch in North America. Upper left, 1958–1961; lower left, 1968–1971; upper right, 1978–1981; lower right, 1988–1990.



House Finch Density, Western Long Island and Connecticut

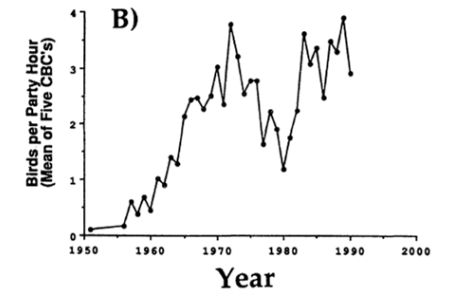


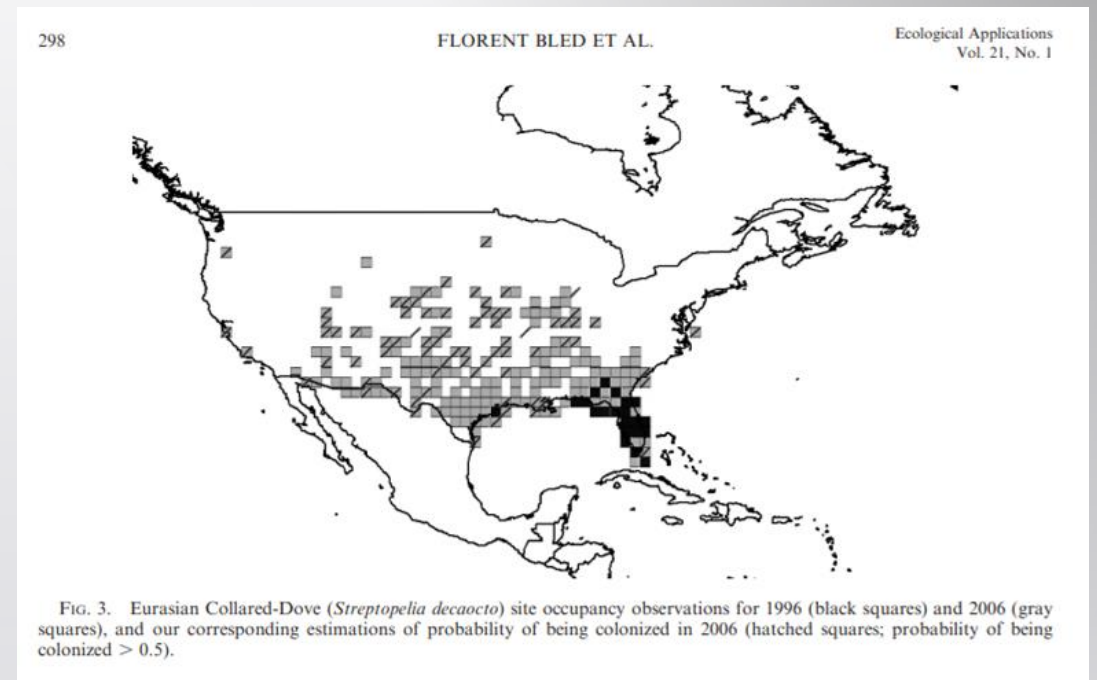
FIG. 2.—A, Spread of the house finch in North America. The range radius is calculated as the radius of a semicircle covering the area invaded from the Eastern seaboard. B, Mean number of house finches on five Christmas bird counts within the core area of their range. Number per square kilometer was estimated by dividing the total number counted by the area of a 15-mi-diameter circle.



House Finch
Haemorhous mexicanus

O que são modelos de distribuição de espécies?

- Conceito
 - “SDMs typically correlate the **presence (or presence/absence) of species** at multiple locations with relevant **environmental covariates** to estimate **habitat preferences or predict distributions**; these outputs are commonly used to inform ecological and biogeographical theory as well as conservation decisions”





O que são modelos de distribuição de espécies?

- Tipos de modelos
 - Modelos lineares/ não-lineares (geralmente logísticos)
 - Modelos lineares mistos
 - Modelos hierárquicos
- Aplicações
 - Distribuição e dinâmica de populações e comunidades
 - Mapeamentos diversos
 -



Métodos para estimativa de parâmetros

- Abordagem frequentista
- Abordagem Bayesiana



Métodos para estimativa de parâmetros

- Abordagem frequentista
 - Os parâmetros geralmente são assumidos como fixos e conhecidos;
 - Os dados são aleatórios e usados para estimar os parâmetros usando Máxima Verossimilhança.

Definimos probabilidade em termos da frequência relativa de algum evento em uma sequência infinita de experimentos aleatórios e replicáveis.

$P(\text{dados} \mid \text{parâmetros})$



Métodos para estimativa de parâmetros

- Abordagem Bayesiana
 - Variáveis aleatórias são usadas para modelar todas as fontes de incerteza. Os parâmetros são incertos, então eles terão uma distribuição!
 - Os dados são tratados e corrigidos e a inferência é realizada de forma condicional aos dados.
 - As análises levam a uma distribuição posterior que descreve ‘crenças’ sobre parâmetros e hipóteses, condicionadas aos dados observados.

Todos os parâmetros são desconhecidos e sua incerteza é representada por distribuições de probabilidade (distribuições ***a priori*** - antes de quaisquer dados serem coletados e distribuições ***a posteriori*** após os dados terem sido coletados e analisados).



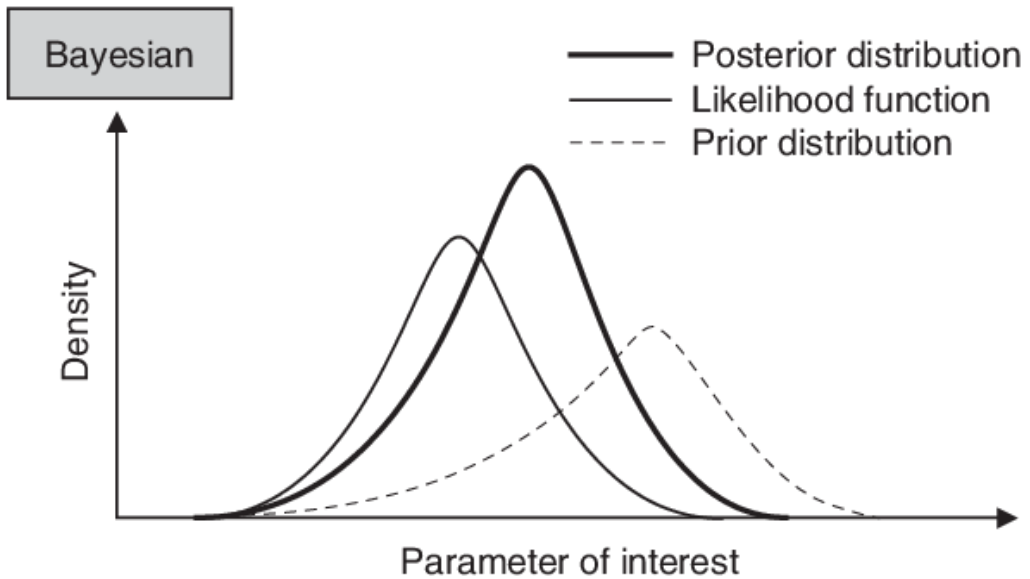
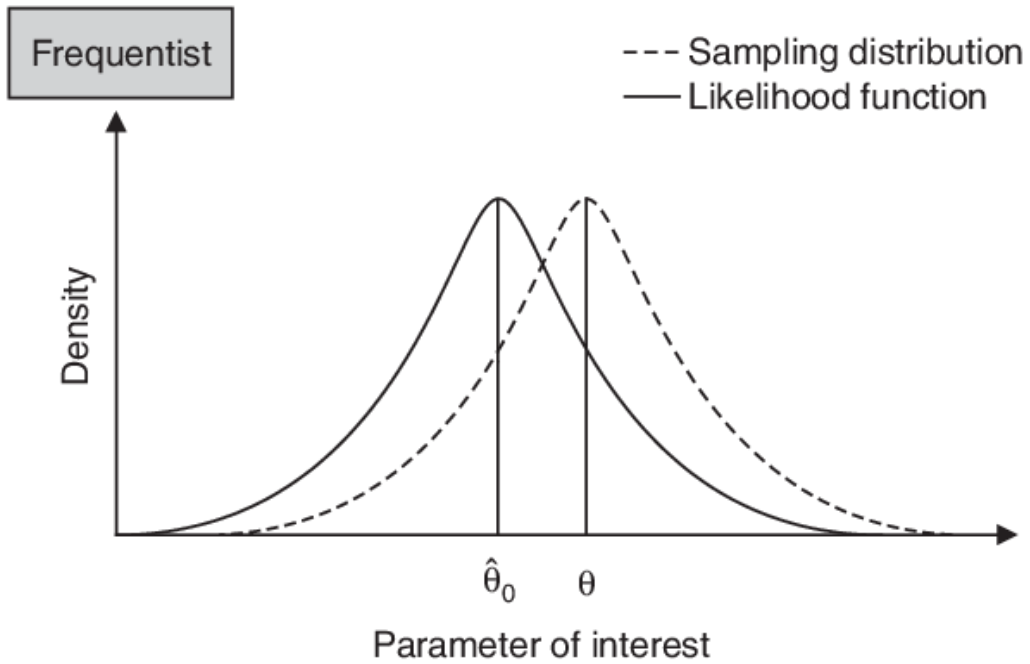
Teorema de Bayes

$$p(A | B) = \frac{p(B | A)p(A)}{p(B)}$$

$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)}$$

θ - parâmetros
 x - dados

Resumindo...









Sumário - Sábado, 30 de julho

- Coleta, organização e mapeamento de dados de ciência cidadã
- Introdução aos modelos de integração de dados
- Introdução ao teste de ajuste dos modelos
- Socialização de projetos



O que precisamos para modelar a distribuição de
uma espécie?



O que precisamos para modelar a distribuição de uma espécie?

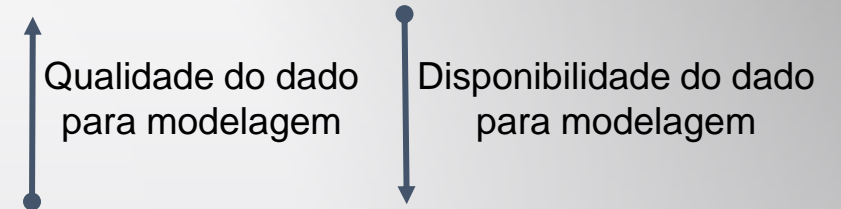
Depende da pergunta que queremos responder!

O primeiro requisito é obter detecções da espécie foco.

Como são os dados?

- Tipo:

- Detecção / não detecção
- Presença / ausência
- Somente detecção (presence-only, background-presence)



- Protocolo

- Padronizados
 - Protocolos amostrais comuns
- Não padronizados
 - Diversos protocolos amostrais



Tipo de dado para modelo hierárquico

Table 2.1 Typical data structure for the classes of HMs implemented in the unmarked package.							
	Detection Data			Site Covariate	Observation Covariate		
	Visit1	Visit2	Visit3	Habitat	Date1	Date2	Date3
Site 1	1	1	1	Good	3	6	10
Site 2	0	0	0	Good	1	7	11
Site 3	1	0	0	Bad	2	9	12
Site 4	0	0	1	Bad	5	6	10

Implicações de assumir uma detecção perfeita

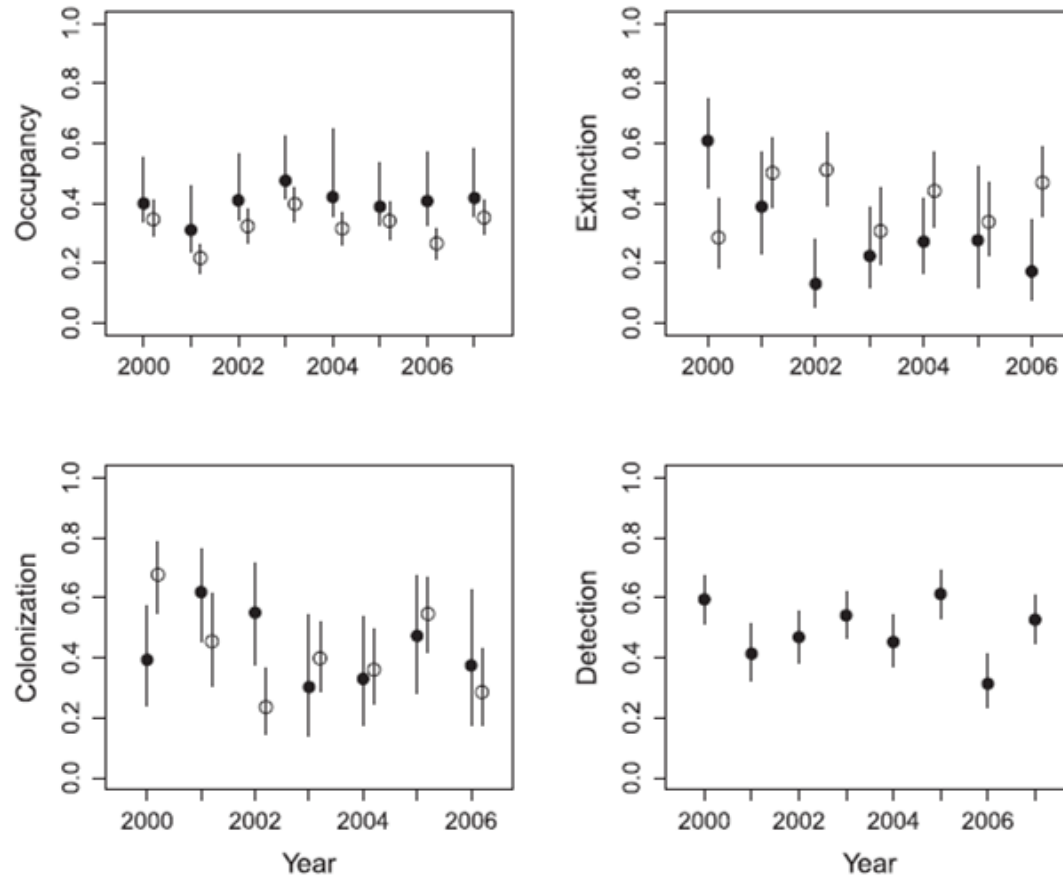


Figure 2 Annual variation in range dynamics and detection parameters for the European crossbill (*Loxia curvirostra*) in Switzerland from 2000 to 2007 under the dynamic occupancy model where detection probability (p) is estimated (filled circles) and under the traditional metapopulation model with the assumption that $p = 1$ (open circles). Lines represent 95% confidence intervals based on asymptotic standard errors except for occupancy probability, where they are based on nonparametric bootstrapping (1000 replicates).

Implicações de assumir uma detecção perfeita

Table 1 Parameter estimates (maximum-likelihood estimates, MLEs, with asymptotic standard errors, ASEs) of the best dynamic occupancy (Dynocc) model, in terms of Akaike's information criterion (AIC), for the range dynamics of the European crossbill (*Loxia curvirostra*) in Switzerland from 2000 to 2007. For comparison, MLEs and ASEs are also given for the best traditional, 'naïve' metapopulation model that assumes perfect detection (i.e. $p = 1$). Terms not in the model selected by AIC are denoted by –. Models contain parameters for elevation (elev), forest cover (forest), year and survey date (date). Dots denote interactions.

Parameter	Dynocc model (p estimated)		'Naïve' model ($p = 1$ fixed)	
	MLE	ASE	MLE	ASE
Initial occupancy (ψ_1)				
Intercept	–0.199	0.243	–0.244	0.281
elev	1.869	0.354	1.994	0.338
elev ²	–	–	–0.140	0.341
forest	0.813	0.255	0.215	0.360
forest ²	–0.142	0.197	0.005	0.210
elev.forest	0.548	0.254	0.501	0.277
elev.forest ²	–0.621	0.243	–0.444	0.276
elev ² .forest	–	–	1.018	0.373
elev ² .forest ²	–	–	–	–



Modelos de ocupação de sítio

Processo biológico

Processo amostral

Modelos de ocupação de sítio

Processo biológico

Ocorre no sítio i ?

$$z_i \sim \text{Bernoulli}(\psi_i)$$

Processo amostral

Qual a probabilidade de ocorrência dado as covariáveis de sítio?

$$\text{logit}(\psi_i) = \beta_0 + \beta_1 * X1_i + \dots + \beta_n * Xn_i$$

Modelos de ocupação de sítio

Processo biológico

Ocorre no sítio i ?

$$z_i \sim \text{Bernoulli}(\psi_i)$$

Qual a probabilidade de ocorrência dado as covariáveis de sítio?

$$\text{logit}(\psi_i) = \beta_0 + \beta_1 * X1_i + \dots + \beta_n * Xn_i$$

Processo amostral

Detectei a espécie no sítio i e visita j ?



$$Y_{ij} \sim \text{Bernoulli}(z_i, p_{ij})$$

Qual a probabilidade de detecção dado as covariáveis de amostragem?

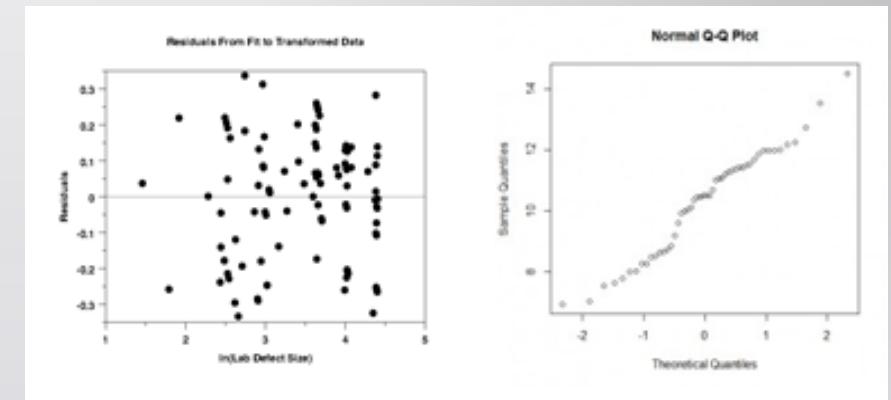
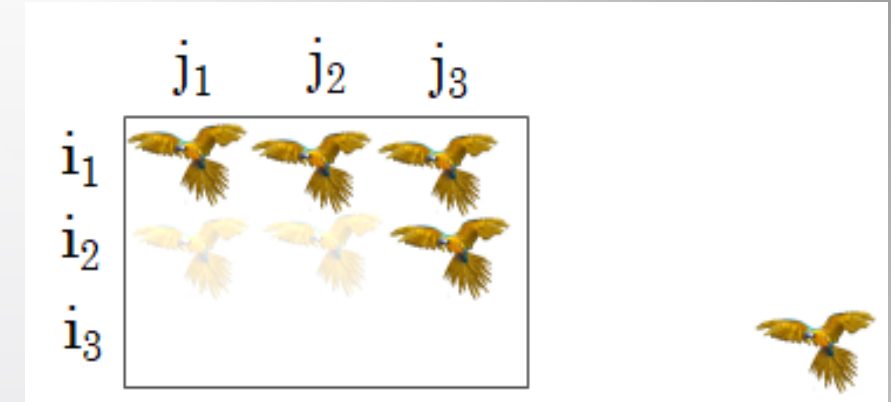
$$\text{logit}(p_{ij}) = \alpha_0 + \alpha_1 * X1_{ij} + \dots + \alpha_n * Xn_{ij}$$

Pressupostos do modelo

- População fechada de $j = 1 \dots J$ visitas (ocasiões amostrais)
- Ausência de falsos positivos;
- Detecção independente;
- Homogeneidade de detecção entre N_i indivíduos;



- Pressupostos de modelos paramétricos;





Implicações na amostragem

- Conhecimento básico sobre a biologia da espécie foco;
- Visitas replicadas;
- Independência entre visitas;
- Intervalo entre visitas curto, de modo que a realidade biológica seja a mesma;
- Coleta de dados de covariáveis de sítio em cada estação amostral (ex: altitude, cobertura florestal);
- Coleta de dados de covariáveis de amostragem em cada visita (ex: vento, umidade do ar);
- Coleta de dados de esforço amostral (ex: tempo amostrando, distância amostrada);



Como modelar no R?

- Modelo 1: sem consideração de esforço amostral e detecção imperfeita – GLM, relacionando a 'presença' da ema com covariáveis de ambiente (área de campo e agricultura por município).
- Modelo 2: considerando o esforço amostral e detecção imperfeita para o mesmo conjunto de dados – usando o pacote *unmarked* e *WinBugs*.



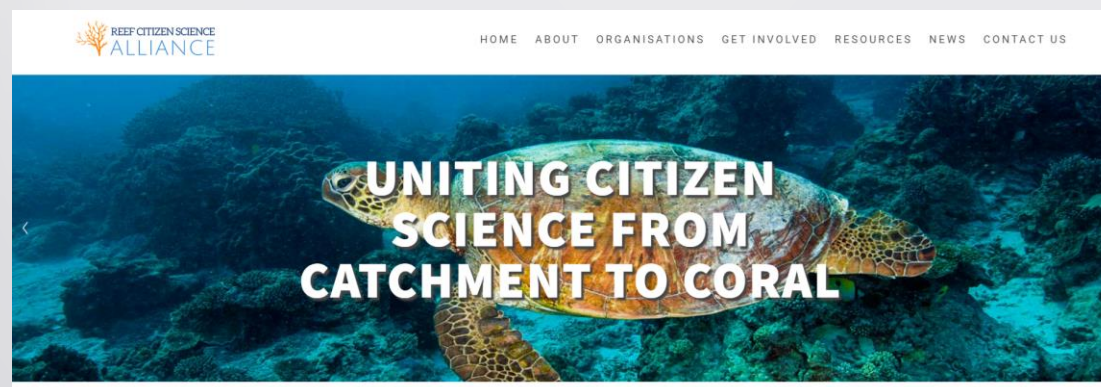
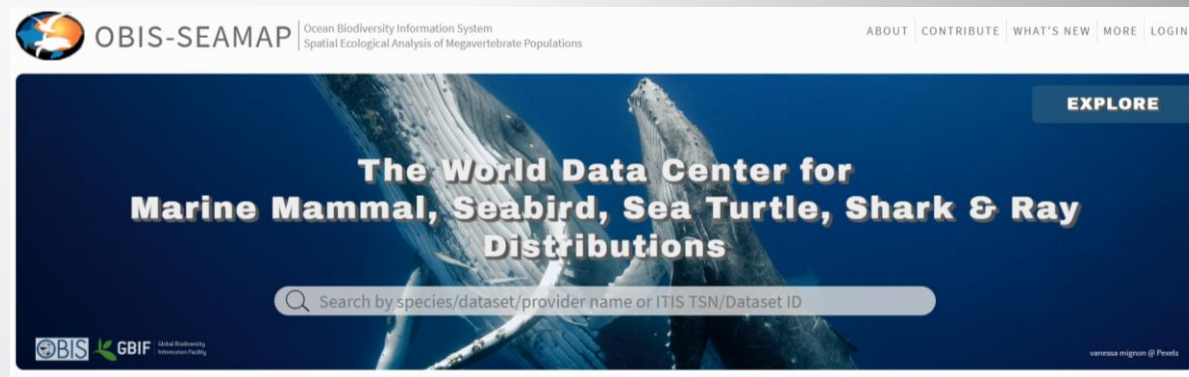
Até agora...


Necessidade de amostragem replicada

Áreas extensas para amostrar

Recursos limitados

Diversas plataformas de ciência cidadã disponíveis





O que são modelos de integração de dados?

- “Integrated methods combine multiple datasets to improve predictions made about species distributions.”
 - Combinação de dados descrevendo observações de indivíduos e espécies em diferentes escalas, geralmente com diferentes estruturas de dados, e níveis de padronização do esforço
 - Integrated population models
 - Spatial capture-recapture models
 - **Data integration models**
- **Todos usam *joint likelihood***

Como a integração é feita?

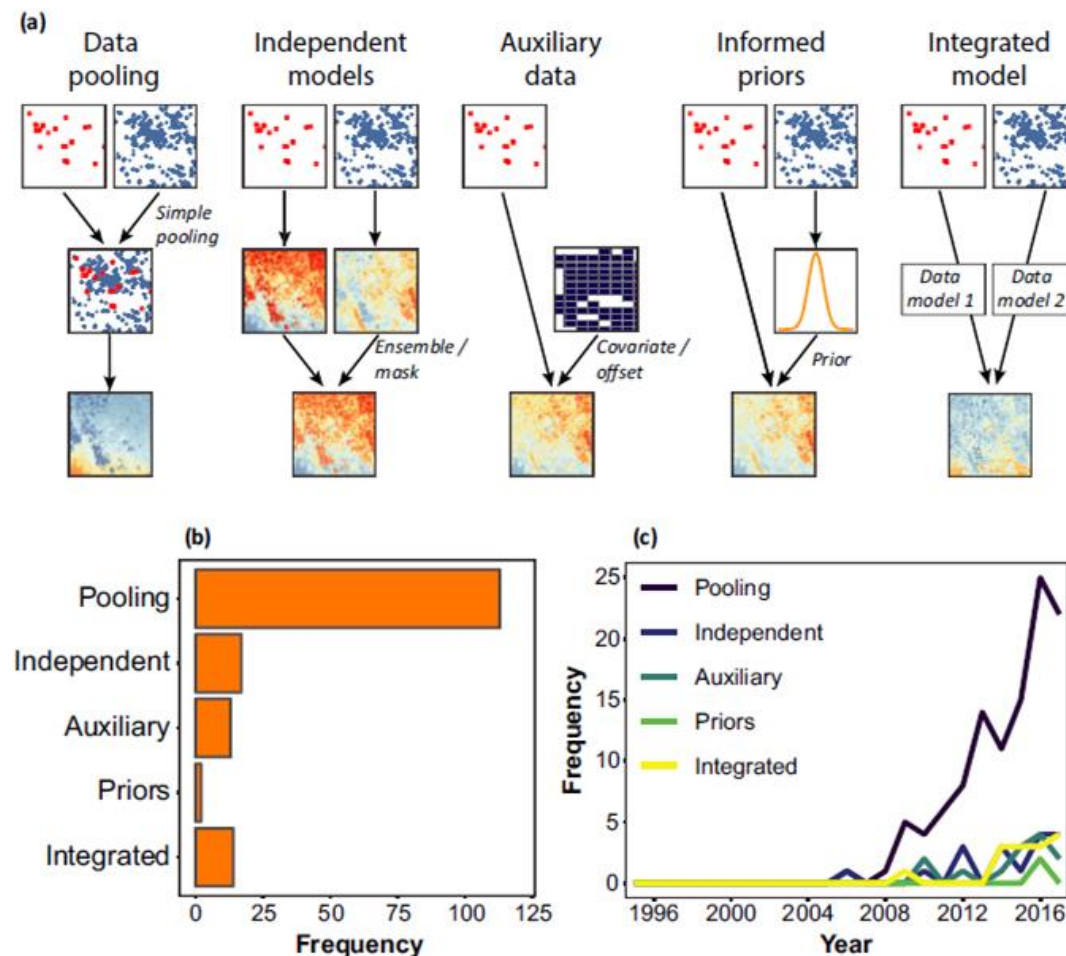
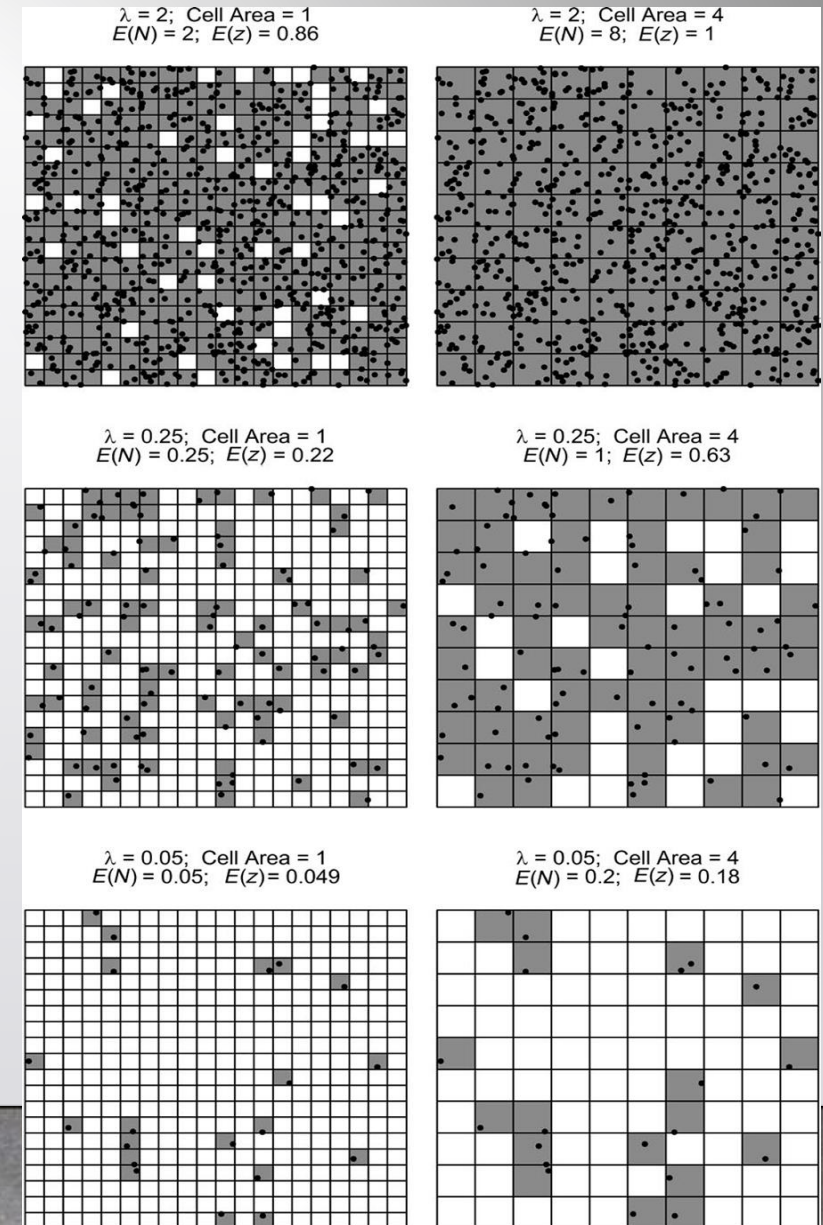


FIG. 2. Combining data for predicting species distributions. Across 353 articles reviewed, (a) there were five general approaches to combining data. First, simple pooling occurred, where different data sources were combined and a single model was fit. Second, independent distribution models were fit to different data sources and results were combined, either through ensemble techniques or through masking/clipping. Third, auxiliary data (not directly species occurrence or abundance) were used in modeling building, typically through the use of covariates or offsets. Fourth, one data source was used to create an informative prior for modeling the primary data source in a Bayesian modeling framework. Finally, multiple data sources were formally integrated by developing separate data models for each source that could then be combined, typically through the use of joint likelihoods. (b) The overall frequency of each approach and (c) the frequency of each approach over time.

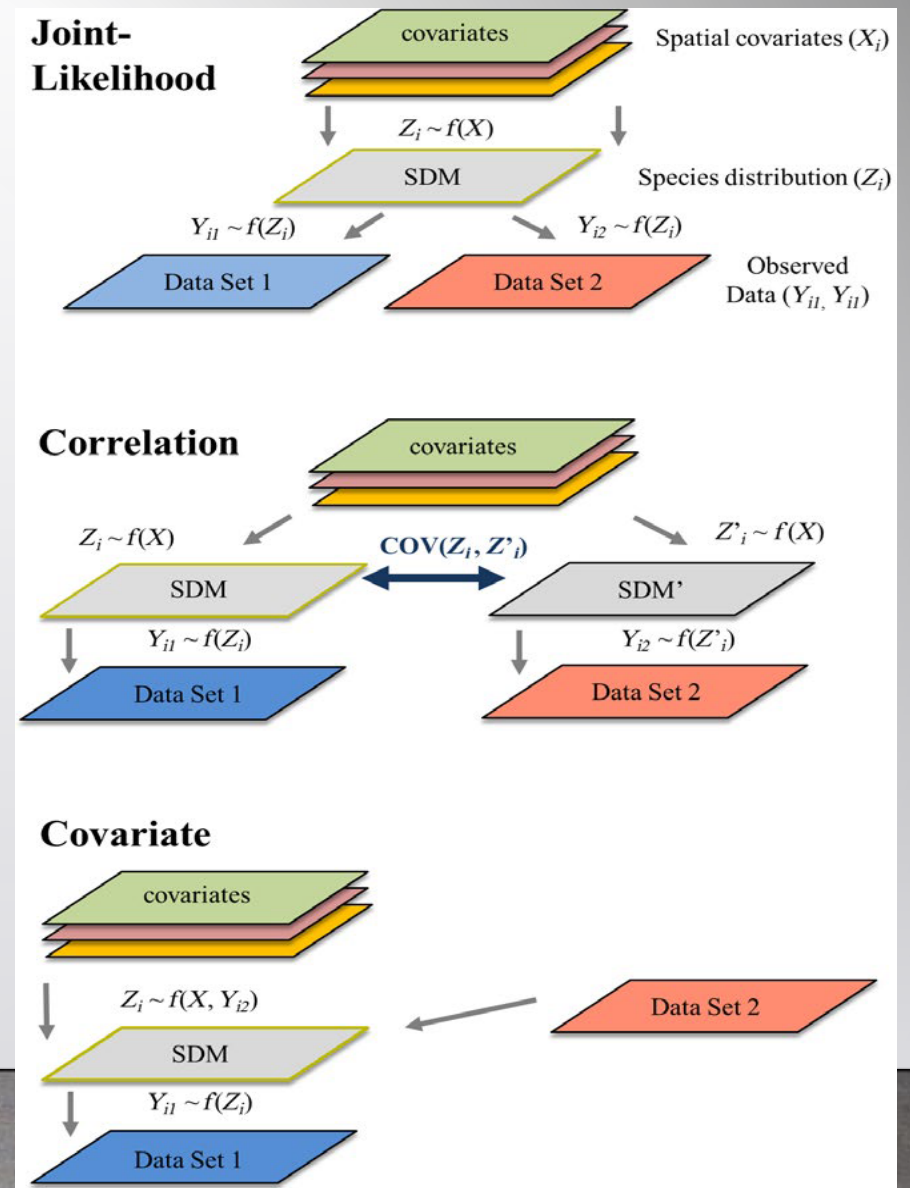
Conceitos chave em modelos de integração de dados

- *Spatial point processes*
 - individual data can be summarized using a density function (λ)



Conceitos chave em modelos de integração de dados

- *Joint-likelihood methods*
- os parâmetros compartilhados são estimados pela maximização da verossimilhança entre todos os conjuntos de dados





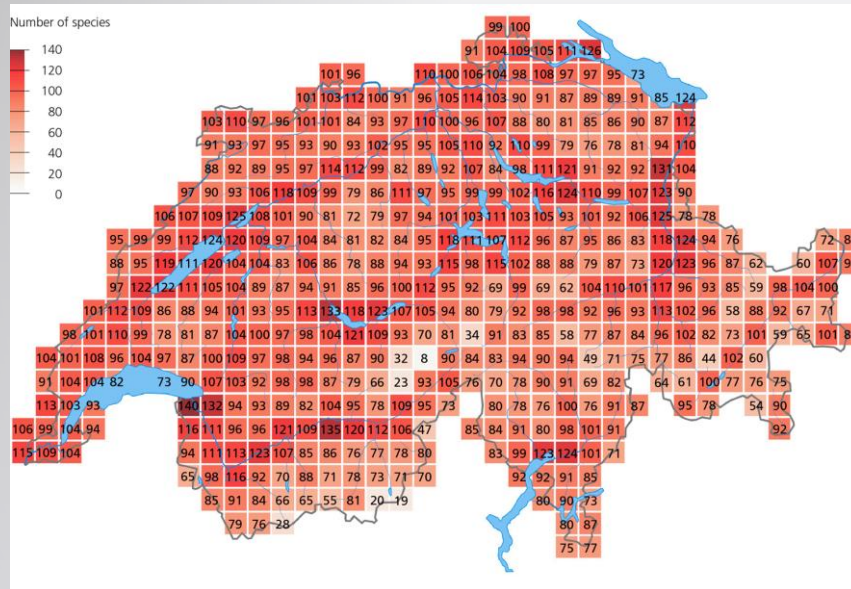
Como modelar no R?

- Modelo 3: sem consideração de esforço amostral e detecção imperfeita – dados agregados de diversas plataformas.

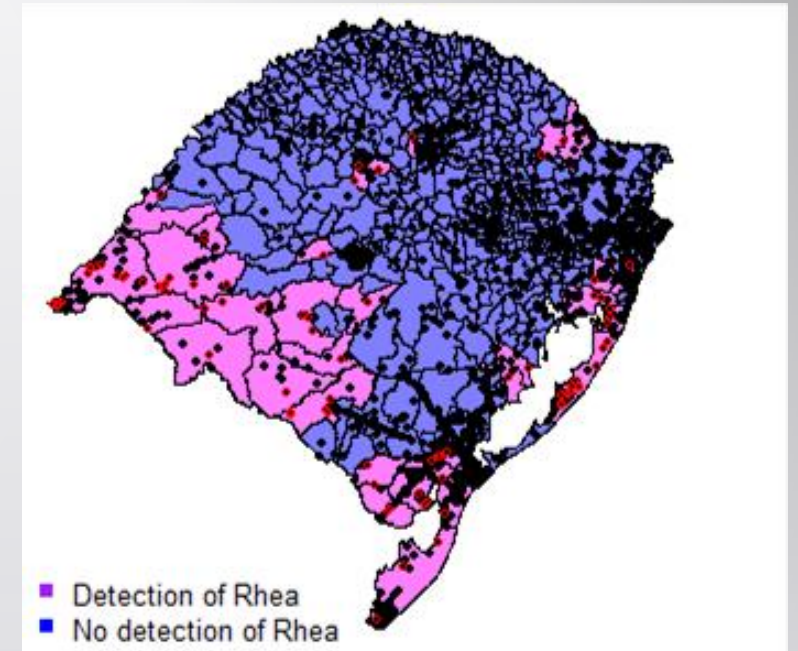


Lidando com detecção imperfeita em DIM

- Dados com protocolos padronizados
 - e.g., eBird, North American BB Survey, Swiss BB Atlas
- Dados não padronizados (adaptações)
 - GBIF, WikiAves, VertNet, iNaturalist



<https://www.vogelwarte.ch/en/atlas/evolution/number-of-breeding-bird-species>



$$P[i]^* = 1 - (1-p)^{E[i]}$$



Lidando com detecção imperfeita em DIM

- Dados com protocolos padronizados



Listas são **réplicas**.

Número de espécies em cada lista; Tempo de observação; Distância percorrida são **medidas de esforço**.



Lidando com detecção imperfeita em DIM

- Dados com protocolos padronizados



Listas são **réplicas**.

Número de espécies em cada lista; Tempo de observação; Distância percorrida são **medidas de esforço**.

- Dados não padronizados (adaptações)



Número de fotos por unidade espacial; (e/ou) Número de sons por unidade espacial; (e/ou) Número de espécies são **medidas de esforço**.



Unidade espacial deve ser compatível com a resolução dos dados



Dados por município, sem coordenada específica.



Cada lista tem uma coordenada geográfica – permite o uso de grids.



Exemplos de organização dos dados

ESTADO	MUNICÍPIO	Nº FOTOS	Nº SONS	Nº ESPECIES	A_VINACEA	IDENT
Rio Grande do Sul	Porto Alegre	5.931	186	225	0	3222
Santa Catarina	Urubici	1.566	85	204	1	2265

ESTADO	MUNICÍPIO	Nº ESPECIES	Nº SONS	A_VINACEA	IDENT
Rio Grande do Sul	Campinas do Sul	3	0	0	510
Bahia	Ilhéus	274	342	0	3287

Modelos de ocupação de sítio

Processo biológico

Ocorre no sítio i ?

$$z_i \sim \text{Bernoulli}(\psi_i)$$

Qual a probabilidade de ocorrência dado as covariáveis de sítio?

$$\text{logit}(\psi_i) = \beta_0 + \beta_1 * X1_i + \dots + \beta_n * Xn_i$$

Processo amostral

Detectei a espécie no sítio i e visita j ?

$$Y_{ij} \sim \text{Bernoulli}(z_i, p_{ij})$$

Qual a probabilidade de detecção dado as covariáveis de amostragem?

$$\text{logit}(p_{ij}) = \alpha_0 + \alpha_1 * X1_{ij} + \dots + \alpha_n * Xn_{ij}$$

Modelos de ocupação de sítio com integração de dados

Processo biológico

Ocorre no sítio i ?

$$z_i \sim \text{Bernoulli}(\psi_i)$$

Qual a probabilidade de ocorrência dado as covariáveis de sítio?

$$\text{logit}(\psi_i) = \beta_0 + \beta_1 * X1_i + \dots + \beta_n * Xn_i$$

Processo amostral

$$Y_{ij} \sim \text{Bernoulli}(z_i, p_j)$$

$$p_j^* = 1 - (1 - p)^{E_j}$$

$$E_j^{EB} = \alpha_2 * SSee_j + \alpha_3 * TObs_j + \alpha_4 * RLen_j$$

$$E_j^{WA} = \alpha_5 * NPho_j + \alpha_6 * NAud_j$$

$$E_j^{GB} = \alpha_7 * NSP_j$$



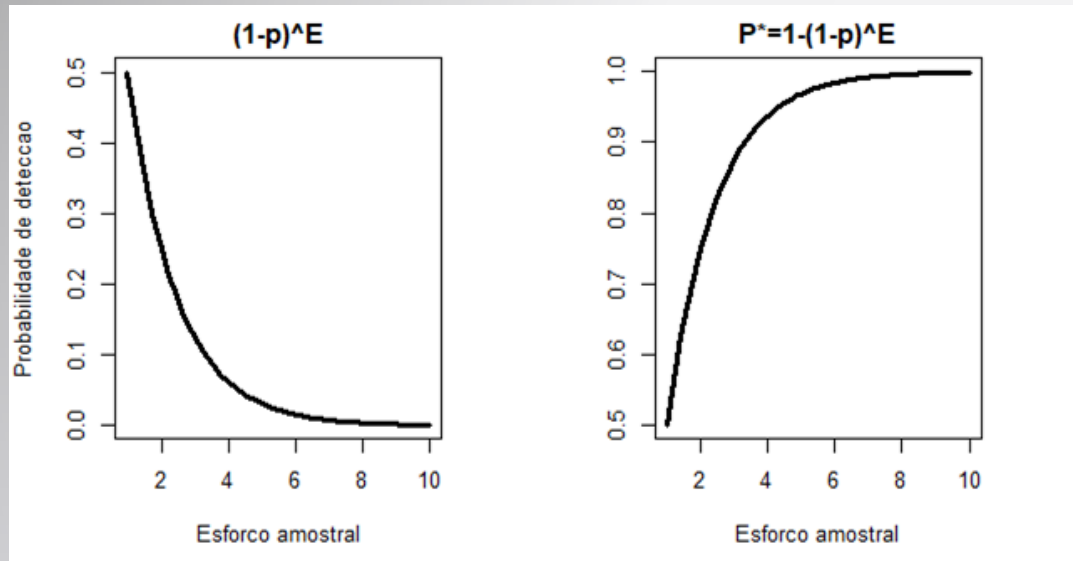
Exemplos de modelo: GLM sem intercepto

$$p_i^* = 1 - (1 - p)^{Ei}$$

$$E[i] = \text{ALPHA.PICT} * \text{nPIC.wikiaves}[i] + \text{ALPHA.SONG} * \text{nSONG.wikiaves}[i]$$

$$E[i,j] <- \text{ALPHA.DIST} * \text{dist}[i,j] + \text{ALPHA.DURATION} * \text{duration}[i,j] + \text{ALPHA.OBSERVER} * \text{observer}[i,j]$$

$$E[i] <- \text{ALPHA.GBIF} * \text{nSP.gbif}[i]$$



$$zPn_i < -Pn_i * z_i$$

$$yn_i \sim \text{dbern}(zPn_i)$$

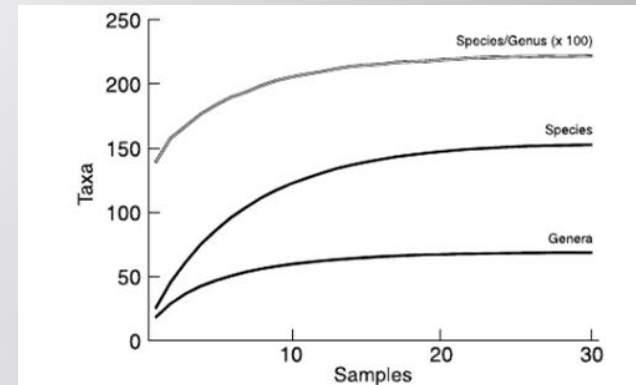


Figure 5 Taxon sampling curves for species and for the genera to which they belong, with the species–genus ratio. Note that the curve for genera reaches its asymptote at a smaller number of samples than the species curve. For this reason, the ratio of species to genera is nonlinear. This pattern is inevitable for any case of category–subcategory sampling curves. The curves are based on a sample of hummingbird specimens (Colwell 2000b; appendix B, pooled, distributed into “samples” at random, and then repeatedly re-sampled using *EstimateS* (Colwell 2000a).



Cuidados com integração de dados

- **Dados desalinhados espacialmente ou temporalmente** (ex: variável 1 medida em nível de município e variável 2 medida em nível de coordenada geográfica).
- **Problema de definição de unidades espaciais** (ex: agregação de dados em área maior; diferenças em tamanho e formato das unidades espaciais).
- **Falácia ecológica** (ex: resposta individual a uma covariável difere dos resultados para o grupo de indivíduos).

SPECIAL FEATURE: DATA INTEGRATION FOR POPULATION MODELS

Ecology, 100(6), 2019, e02709

© 2019 The Authors. *Ecology* published by Wiley Periodicals, Inc. on behalf of Ecological Society of America

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

Resolving misaligned spatial data with integrated species distribution models

KRISHNA PACIFICI,^{1,4} BRIAN J. REICH,² DAVID A. W. MILLER,³ AND BRENT S. PEASE¹



Como resolver:

1. Defina o modelo estocástico para o processo biológico na escala mais fina possível.
2. Determine a escala desejada para as previsões (ex: a escala que as decisões de conservação e manejo serão tomadas).
3. Identifique a melhor maneira de integrar as fontes de dados com base no processo biológico. Uma segunda fonte de dados pode fornecer uma diversidade de informações, incluindo fontes de erro ou esforço.
4. Desenvolva o modelo de integração para fontes de dados e para o processo biológico e realize as inferências.
5. Conduza a avaliação do modelo e verifique a sensibilidade do modelo, especificamente ao alterar a escala dos dados.



Como modelar no R?

- Modelo 4: considerando o esforço amostral e detecção imperfeita – com integração e dados de diferentes plataformas separadamente.

Vantagem: Permite considerar as especificidades de cada plataforma.

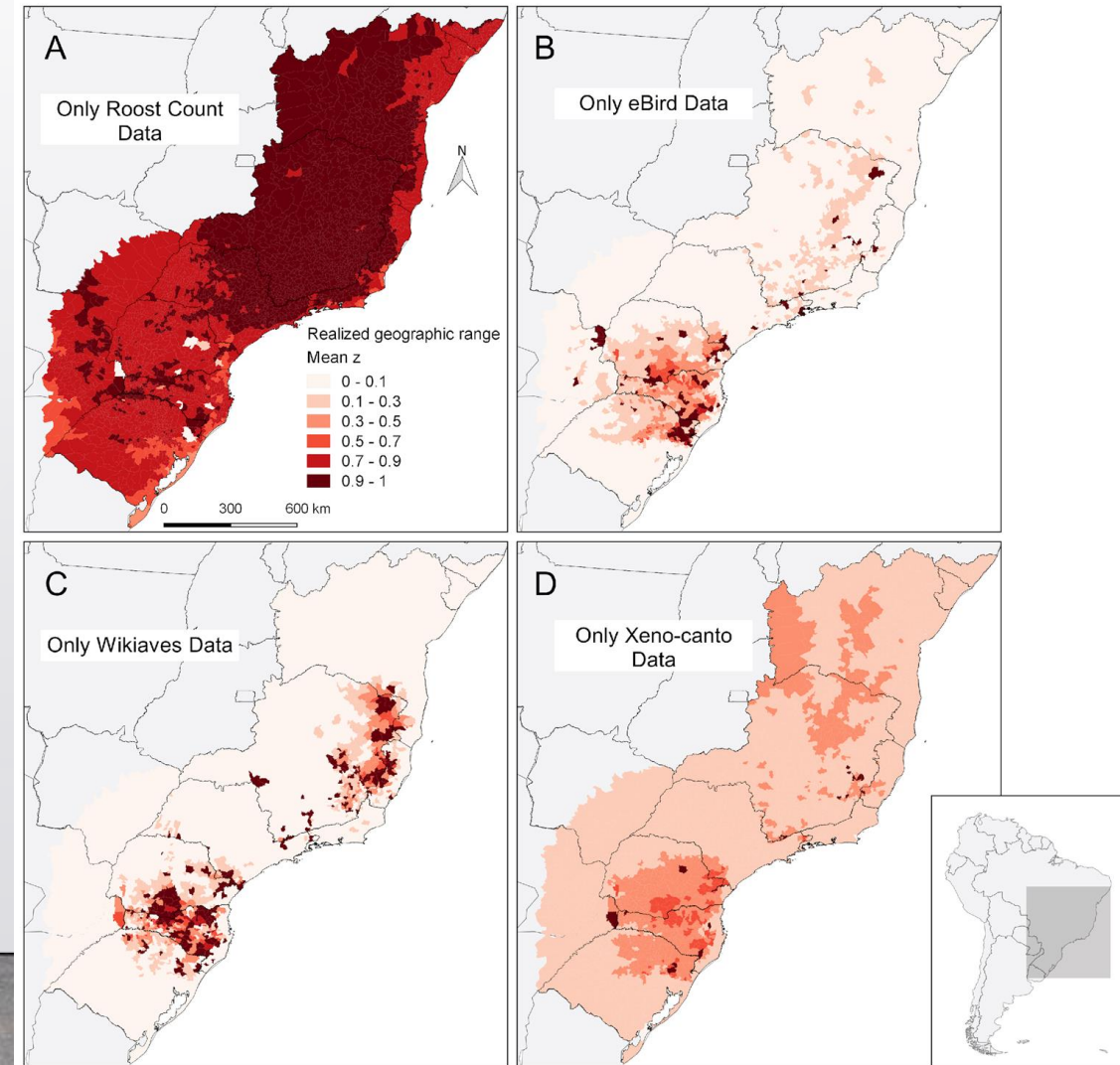
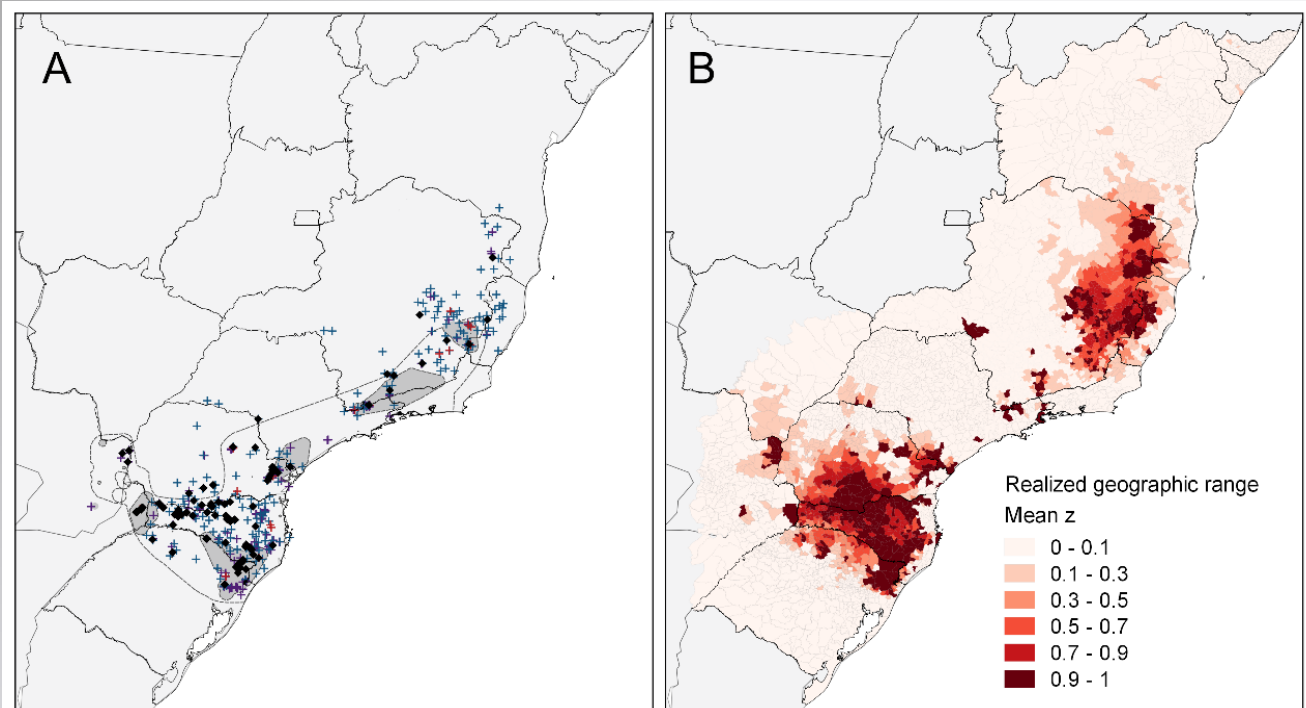




Sumário - Domingo, 31 de julho

- Interpretação das estimativas
- Produção dos mapas, análise do efeito de covariáveis
- Teste de ajuste dos modelos
- Encerramento

Integrar é melhor do que modelar cada base de dados sozinha.



Comparando modelos

Medida de deviance:

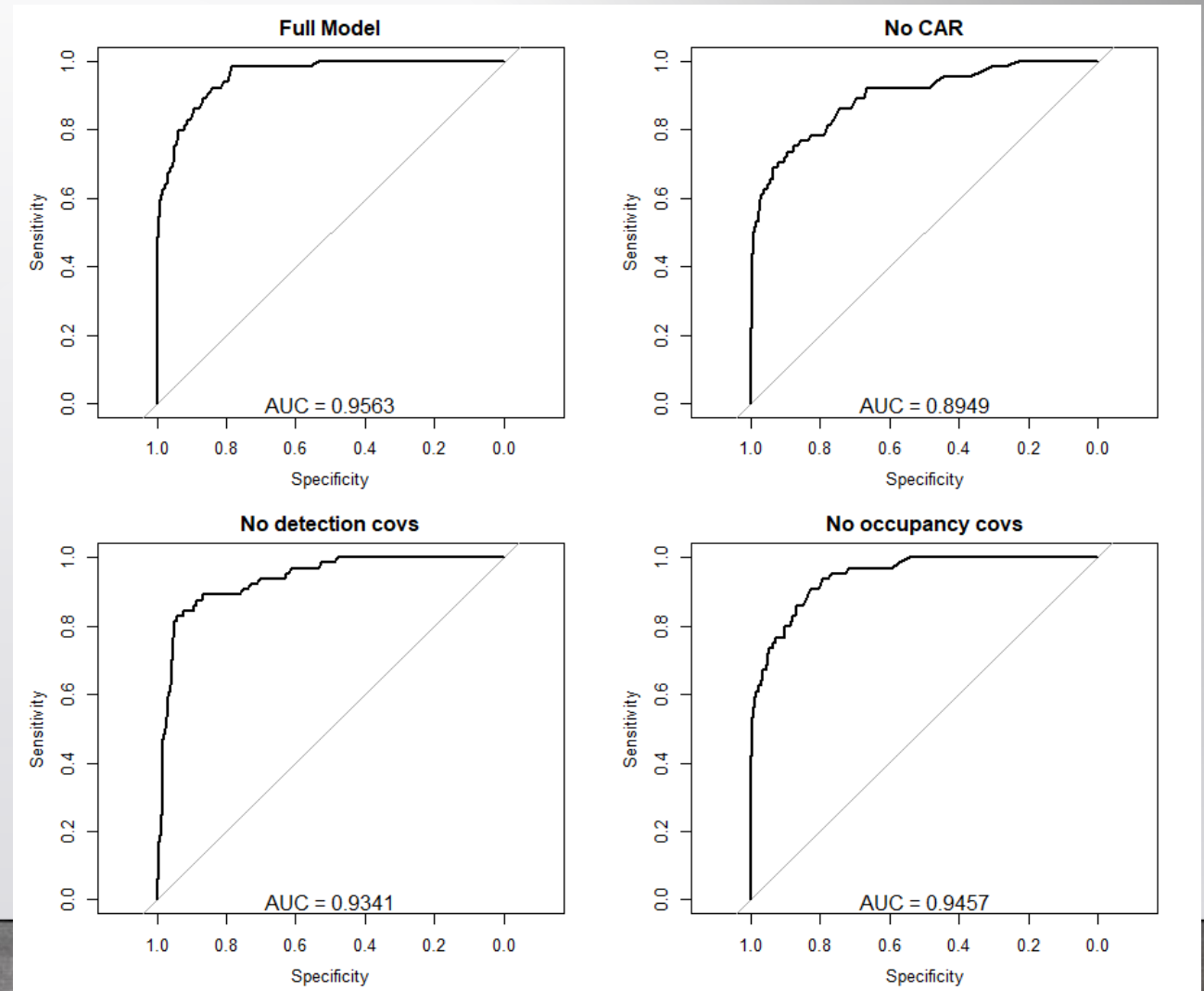
$$D = -2 \sum \log(\mathcal{L})$$

$$\mathcal{L} = \hat{y}^y * (1 - \hat{y})^{1-y}$$

Models	Deviance in each data set				
	Total Deviance	RC	EB	WA	XC
1. Full Model	440.85	28.84	281.19	103.35	27.46
2. No CAR	581.32	50.97	362.58	139.56	28.20
3. No detection covs.	952.84	57.21	735.61	133.60	26.41
4. No occupancy covs.	477.06	26.04	315.34	107.79	27.87

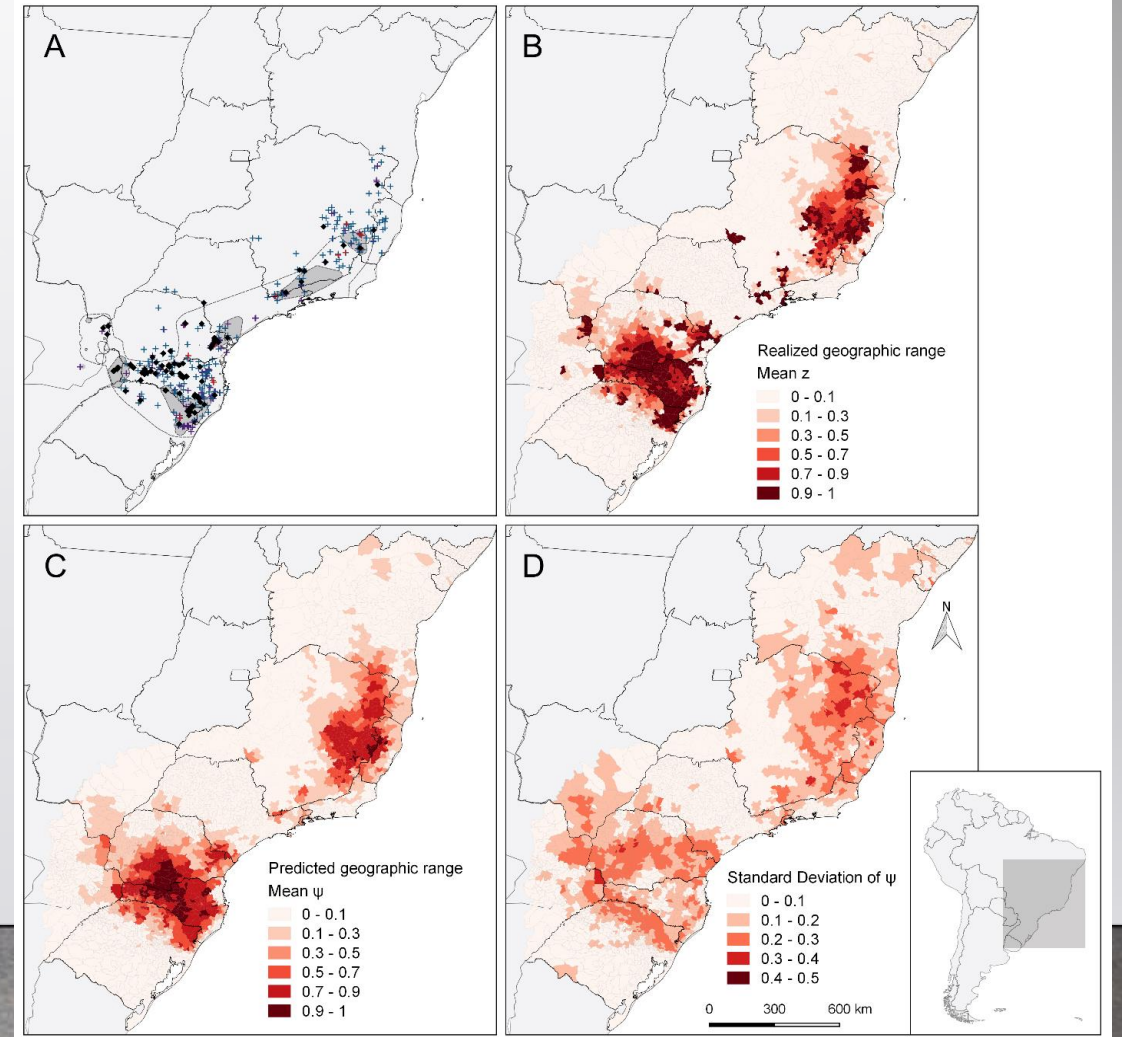
Comparando modelos

Curvas AUC:



Medida de área de distribuição (com incerteza associada):

$$A = \sum \psi_i * area_i$$






Considerações finais:

Ter perguntas definidas é sempre o primeiro e um dos mais importantes passos em qualquer projeto.

Existem diferentes formas de responder as perguntas. É importante conhecer as possibilidades de análise antes de coletar os dados.



To a man with a hammer,
Everything looks like a nail.