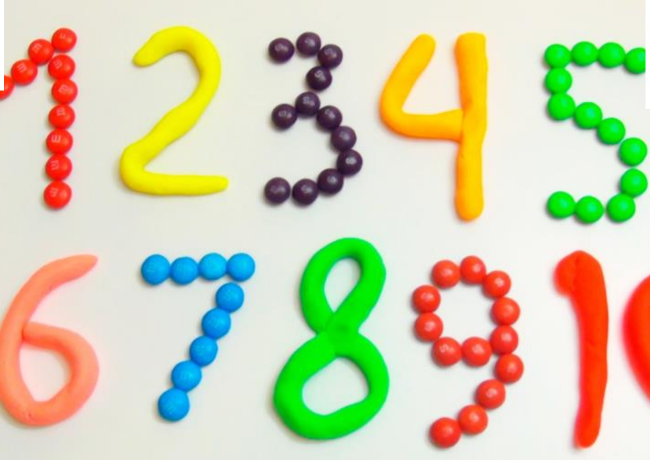
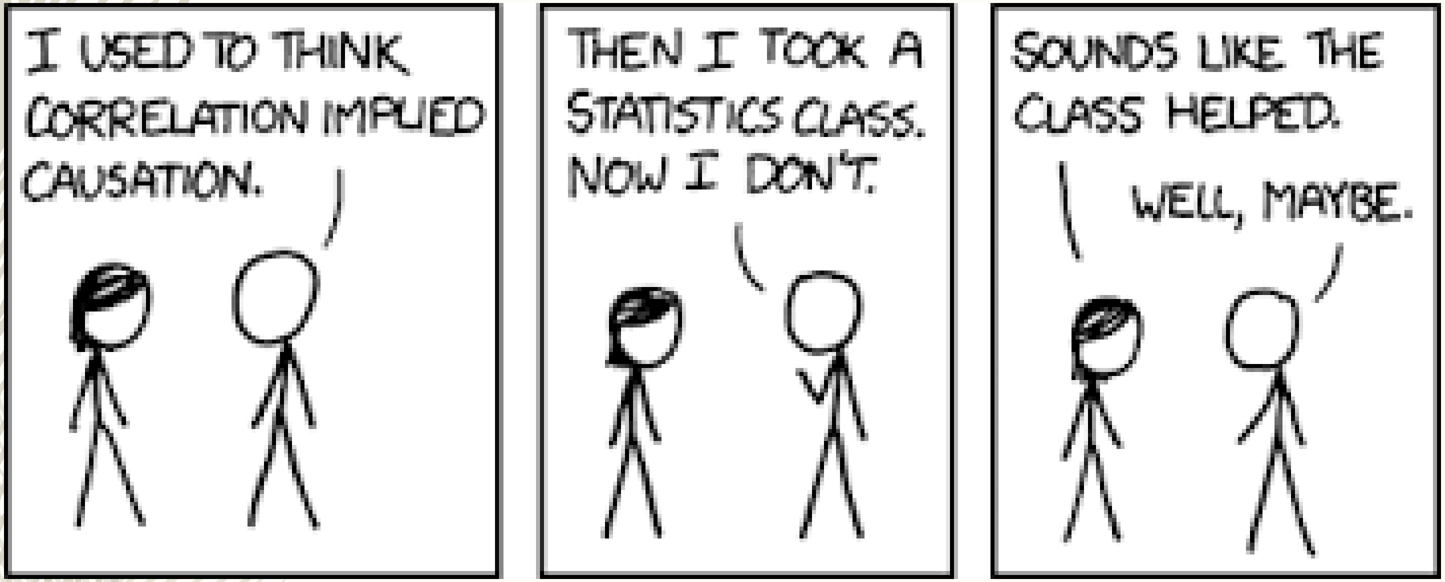
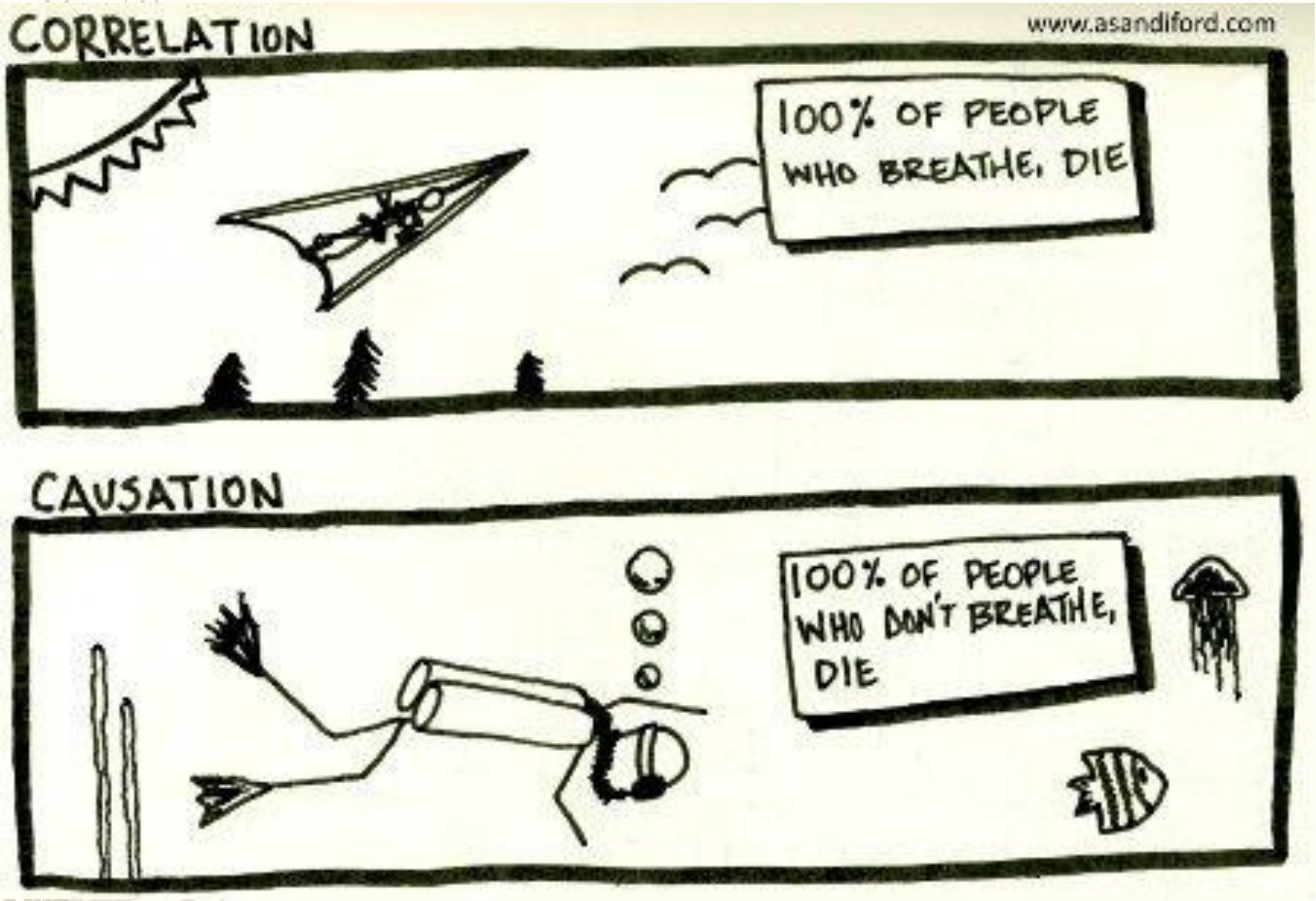
Aula 19 Goodies*



* Goodies related to animals, plants and numbers…

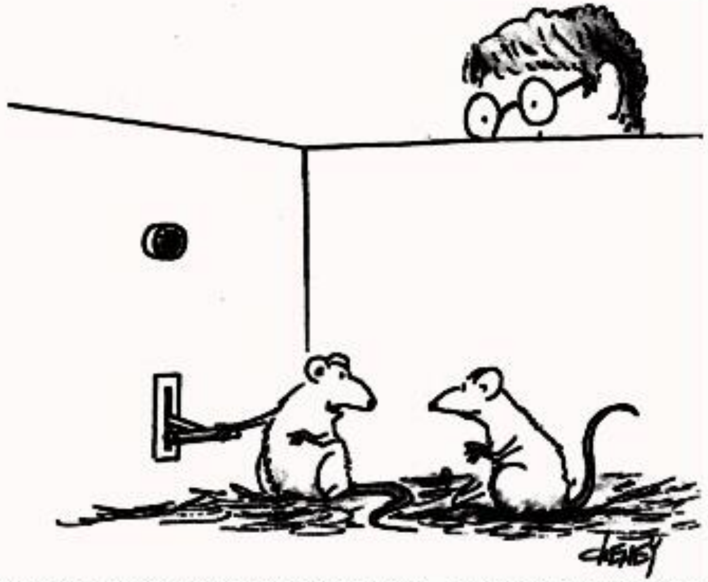

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.



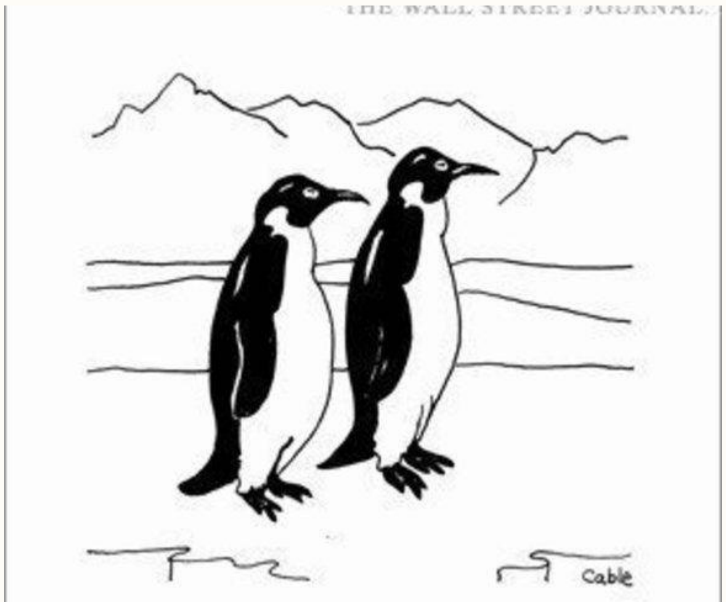
CORRELATION

CAUSATION

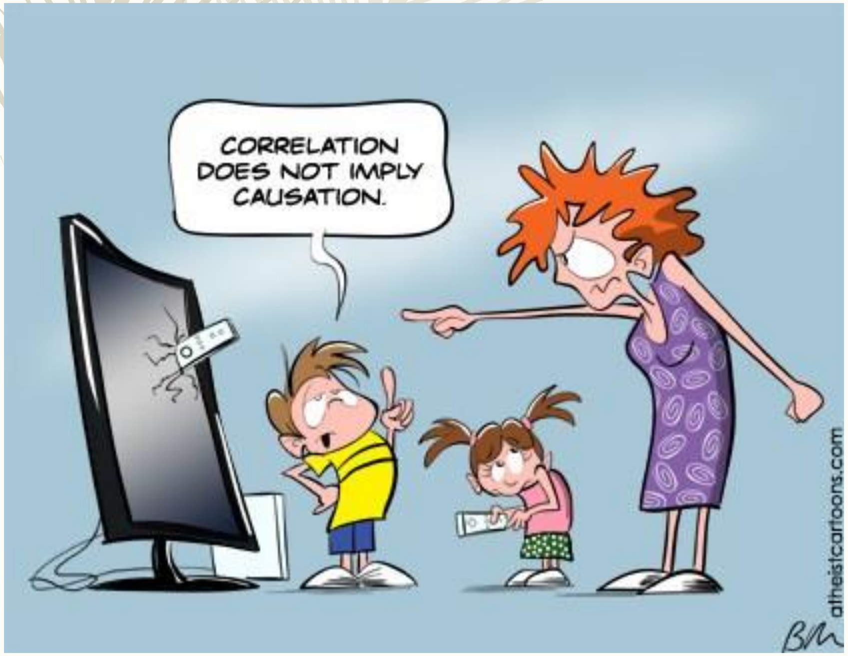
www.asandiford.com



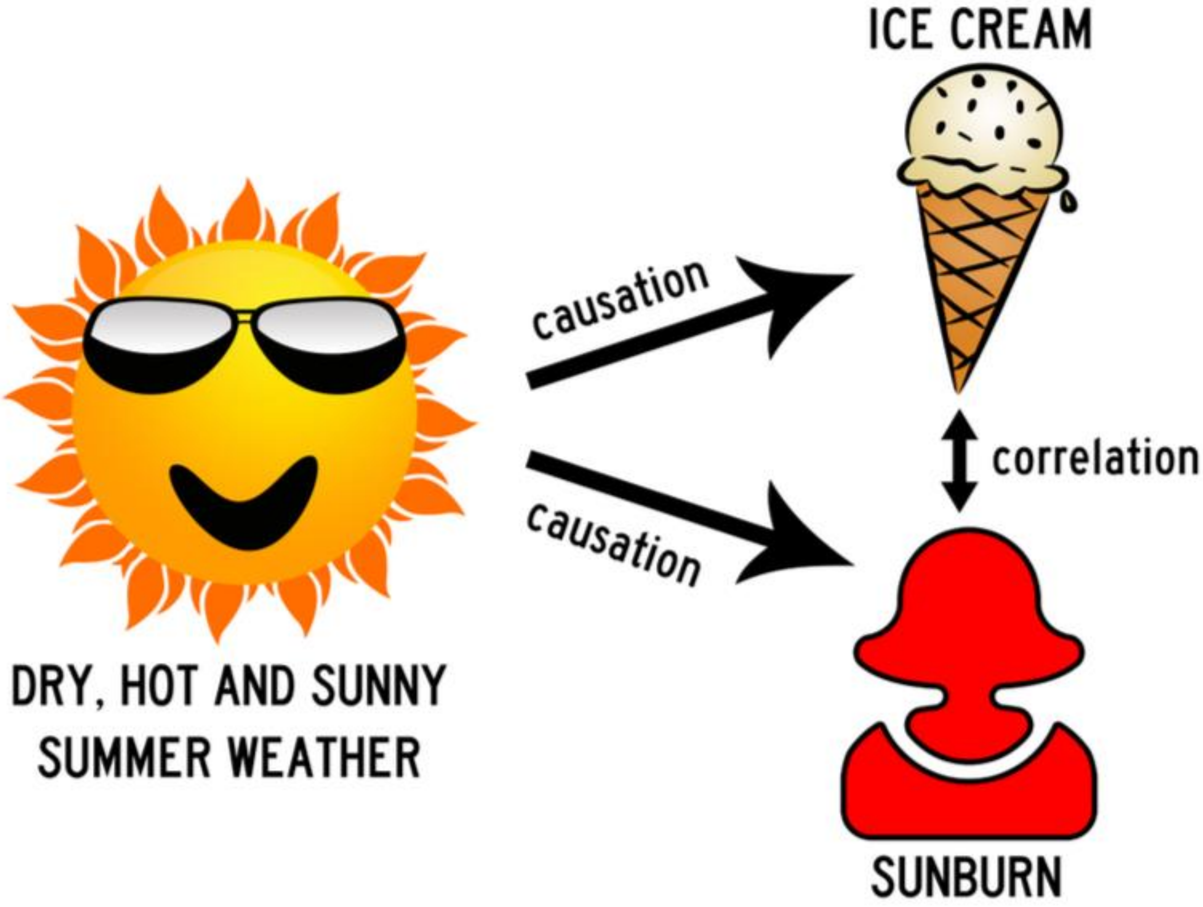
It's a rather interesting phenomenon. Every time I press this lever, that post-graduate student breathes a sigh of relief.



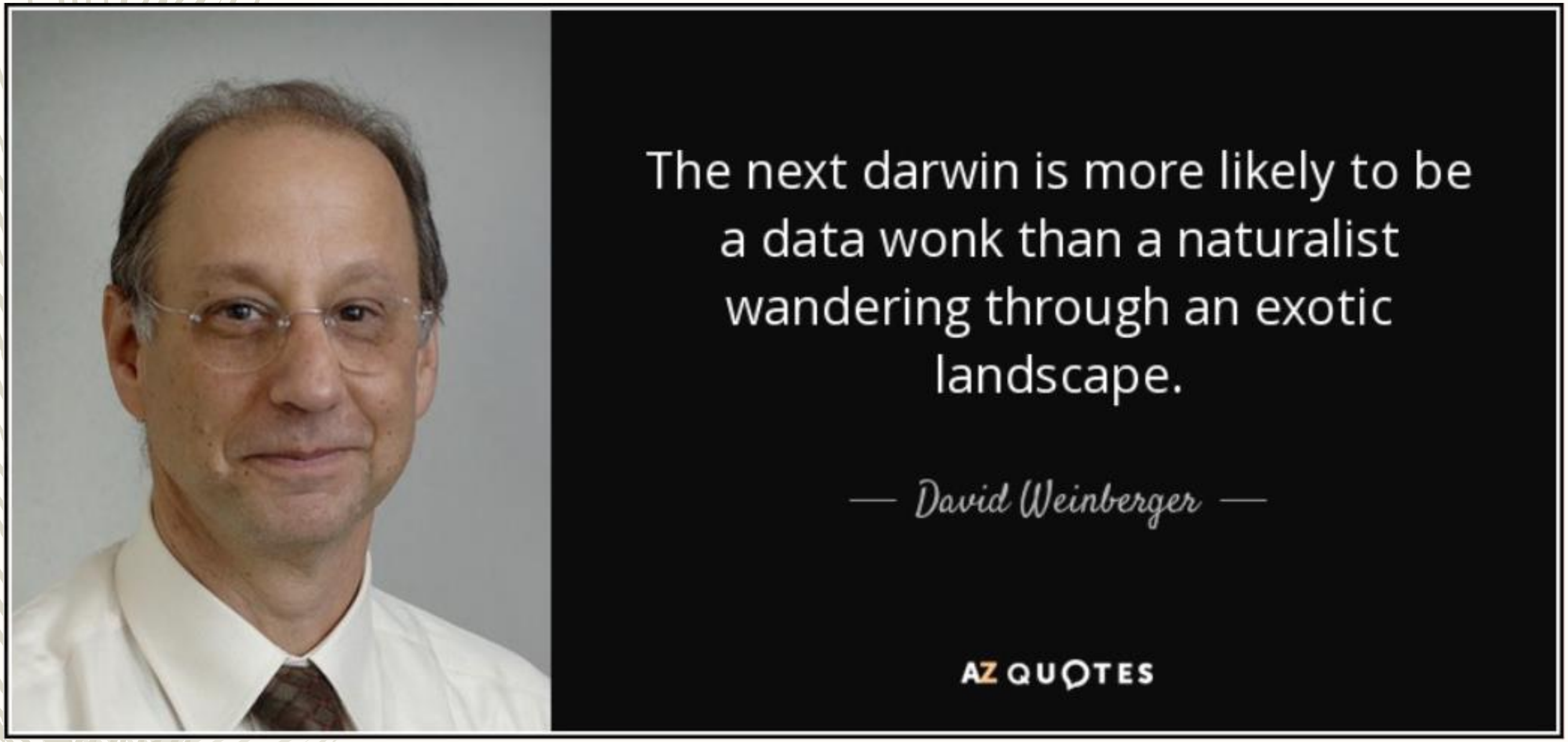
"Do you think all these film crews brought on global warming or did global warming bring on all these film crews?"



"WE'VE ESTABLISHED A CLEAR LINK."


ICE CREAM
causation
causation
correlation
DRY, HOT AND SUNNY SUMMER WEATHER
SUNBURN

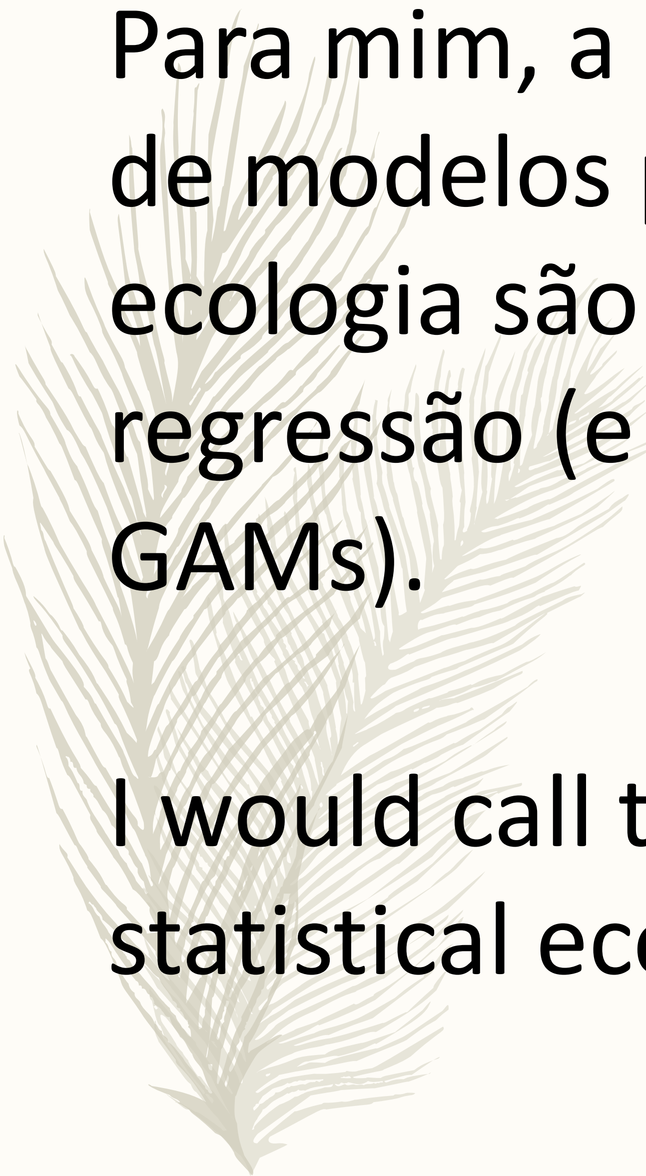
Ecologia Numérica - Aula Teórica 19 – 19-11-2018



The next darwin is more likely to be a data wonk than a naturalist wandering through an exotic landscape.

— David Weinberger —

https://www.azquotes.com/quote/671155



Para mim, a mais importante classe de modelos para quem trabalha em ecologia são os modelos de regressão (e dentro destes, GLMs e GAMs).

I would call these the cornerstone of statistical ecology.

Exemplos práticos da utilização de **GLMs** e **GAMs** em ecologia

Sebastián-González, E.; Camp, R. J.; Tanimoto, A. M.; de Oliveira, P. M.; Lima, B. B.; Marques, T. A. & Hart, P. J. 2018 Density estimation of sound-producing terrestrial animals using single automatic acoustic recorders and distance sampling Avian Conservation and Ecology, 13:7

Fig. 1. Picture and spectrogram of an ʻamakihi at the study area.

Fig. 2. Methodological steps suggested to estimate density from sound-emitting species.

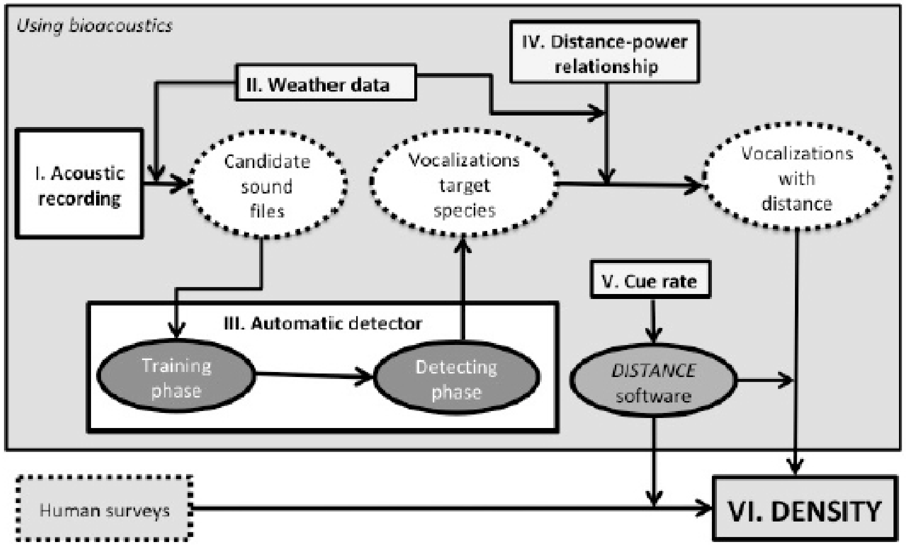
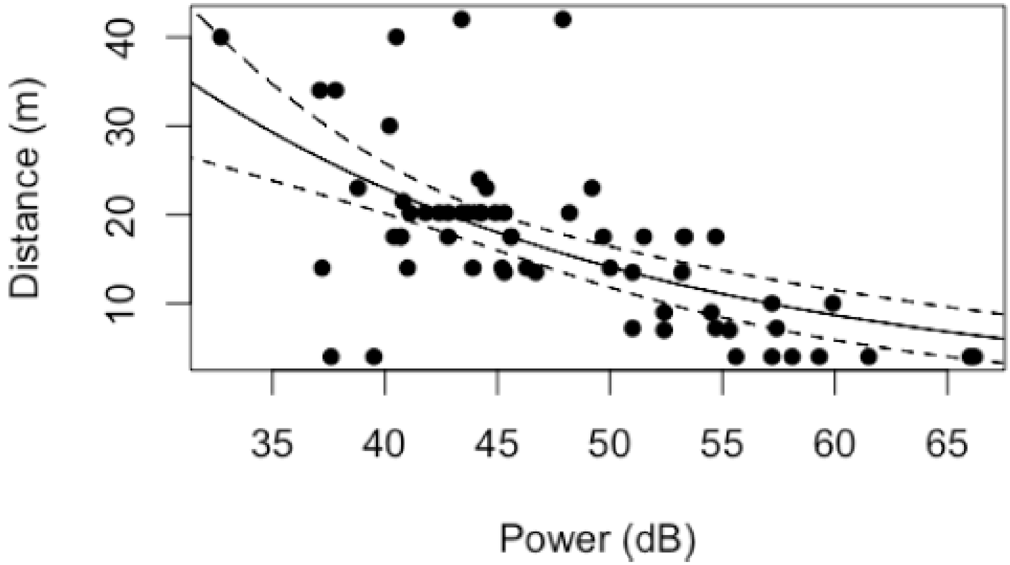


Fig. 3. Relationship between power and distance to the vocalizing bird. We also represent the regression line fron model used to estimate the distances.



The power of the cue was significantly related to the distance of the vocalizing individual (GLM, coefficient ± SE = -0.048 ± 0.009, Intercept ± SE = 5.077 ± 0.406, t-value = -5.221, DF = 60, $P < 0.001$, explained deviance = 37.45%) (Fig. 3, Fig. A1.1). This relationship was not affected by the presence of rain or wind (variables not included in the model with the lowest AIC). Also, the measured distance was significantly related to the predicted one (LM, coefficient ± SE = 1.043 ± 1.04, Intercept ± SE = -0.807 ± 3.085, $P < 0.001$, $R^2 = 0.36$) (Fig. A1.2).

Garcia-Barón, I.; Cortés-Avizanda, A.; Verburg, P. H.; Marques, T. A.; Moreno-Opo, R.; Pereira, H. M. & Donázar, J. A. 2018 How to fit the distribution of apex scavengers into land-abandonment scenarios? The Cinereous vulture in the Mediterranean biome Diversity and Distributions

We used generalized linear m
ence/absence of Cinereous vult
logit function). The data set wa
742 presences, 4,403 and 3,805
respectively). Thus, to balance
sences, 1,000 independent sam
selected for nesting and foragi
for each sample. Models were f
stepwise procedures, using Aka
select the best model of each tr
environment (version 3.1.1, R De
the function *glm* in the "stats" pa

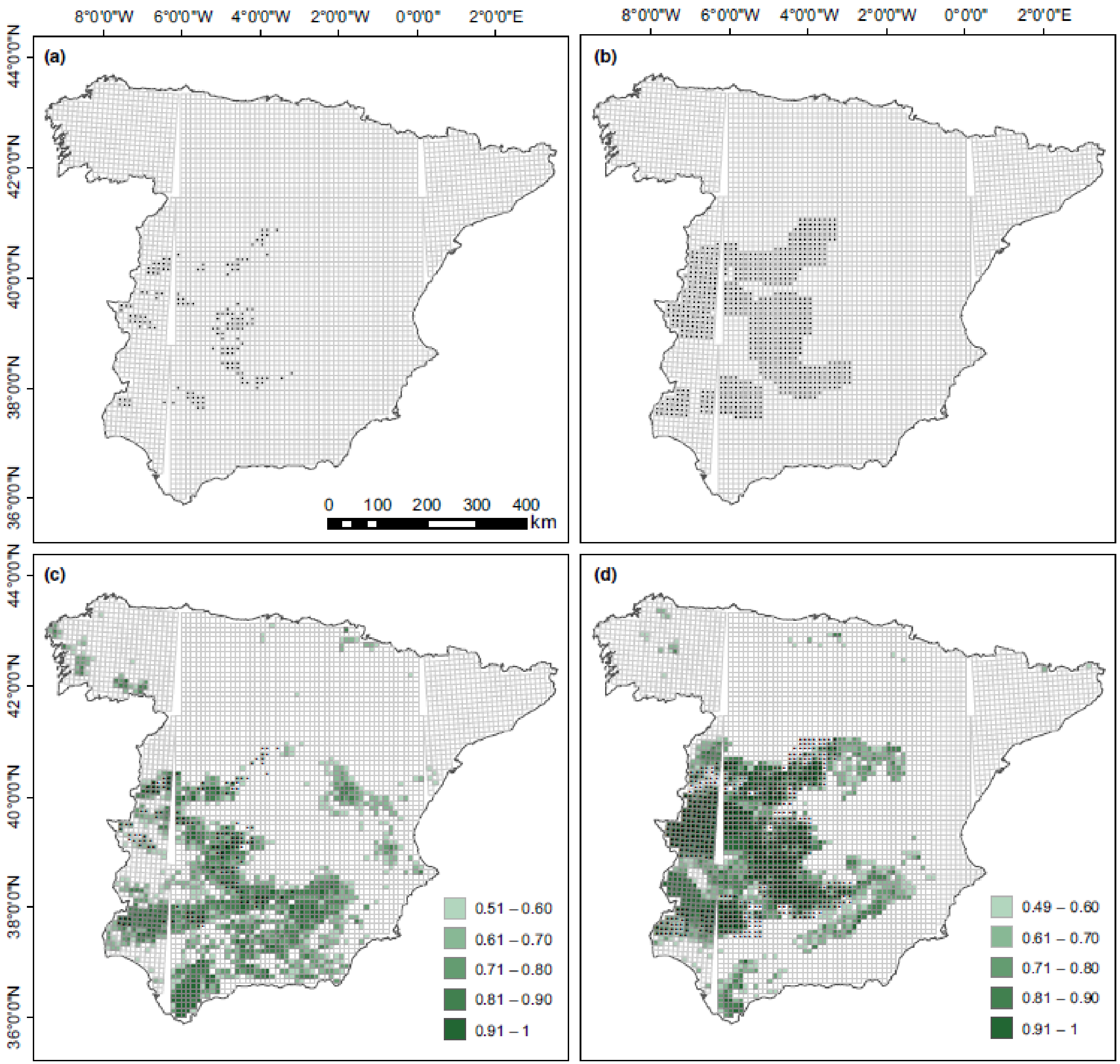


FIGURE 1 Distribution of Cinereous vultures in Peninsular Spain (a) nesting (based on Martí & del Moral, 2003); and (b) foraging (created according to inference from radio-tracking studies (Carrete & Donázar, 2005; Moreno-Opo et al., 2010). The last two maps represent the average prediction of the 1,000 final models showing each UTM 10 × 10 km cells predicted as suitable according to the cut-off value, (>0.51 for nest-site habitat and >0.49 for foraging habitat): (c) nest-site habitat; and (d) foraging habitat

Um exemplo
de
aplicação
de um
GAM:
a primeira estimativa
de
densidade de cetáceos
usando
acústica passiva

Estimating cetacean population density using fixed passive acoustic sensors: An example with Blainville's beaked whales

Tiago A. Marques[a] and Len Thomas
Centre for Research into Ecological and Environmental Modelling, The Observatory, University of St Andrews, St Andrews KY16 9LZ, Scotland

Jessica Ward and Nancy DiMarzio
Naval Undersea Warfare Center Division, 1176 Howell Street, Newport, Rhode Island 02841

Peter L. Tyack
Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543

(Received 2 December 2008; revised 3 February 2009; accepted 5 February 2009)

Methods are developed for estimating the size/density of cetacean populations using data from a set of fixed passive acoustic sensors. The methods convert the number of detected acoustic cues into animal density by accounting for (i) the probability of detecting cues, (ii) the rate at which animals produce cues, and (iii) the proportion of false positive detections. Additional information is often required for estimation of these quantities, for example, from an acoustic tag applied to a sample of animals. Methods are illustrated with a case study: estimation of Blainville's beaked whale density over a 6 day period in spring 2005, using an 82 hydrophone wide-baseline array located in the Tongue of the Ocean, Bahamas. To estimate the required quantities, additional data are used from digital acoustic tags, attached to five whales over 21 deep dives, where cues recorded on some of the dives are associated with those received on the fixed hydrophones. Estimated density was 25.3 or 22.5 animals/1000 km^2, depending on assumptions about false positive detections, with 95% confidence intervals 17.3–36.9 and 15.4–32.9. These methods are potentially applicable to a wide variety of marine and terrestrial species that are hard to survey using conventional visual methods. © *2009 Acoustical Society of America.* [DOI: 10.1121/1.3089590]

PACS number(s): 43.30.Sf, 43.80.Ka [WWA] Pages: 1982–1994

I. INTRODUCTION

Cetaceans (whales and dolphins) form a key part of marine ecosystems, and yet many species are potentially threatened with extinction by human activities. One essential element of an effective conservation or management strategy is a reliable estimate of population size ("abundance") or, equivalently, number per unit area ("density"). However, most cetacean species are hard to survey, since they live at low density over large areas of ocean and spend almost all of their time underwater. The object of this paper is to increase the repertoire of tools available for making species assessments, by developing and demonstrating methods for estimating cetacean density from surveys of their vocalizations collected from fixed passive acoustic sensors.

Currently, the main method for obtaining estimates of density is through visual line transect surveys. A set of randomly placed lines is traversed by an observation platform (e.g., ship, airplane, or helicopter) and all sighted animals of the target species are recorded, together with their perpendicular distance from the line. In the standard method, it is assumed that all animals on the transect line (i.e., at zero distance) are seen with certainty, but that probability of detection declines with increasing distance from the line. The distribution of observed detection distances is then used to estimate the average probability of detection, and this in turn allows estimation of population abundance or density. Line transects are a special case of distance sampling methods, which are described in detail in the two standard texts by Buckland *et al.* (2001, 2004).

Visual line transect methods have a number of disadvantages for surveying cetaceans: they can only be performed during daylight hours and are strongly dependent on good weather conditions; they do not work well for species that spend long periods of time underwater; they are expensive to do well and have restricted temporal coverage. On the other hand, some cetacean species make frequent and characteristic vocalizations, and this has led to increasing recent interest in the use of passive acoustic methods for monitoring cetacean populations (see review by Mellinger *et al.*, 2007b). One solution is to replace or supplement the visual observers on a shipboard line transect survey with a towed passive acoustic platform, since even a simple two-element hydrophone array can be used to obtain locations of repeatedly vocalizing animals, and hence the required perpendicular distances. This has proved particularly effective for sperm whales (*Physeter macrocephalus*), which are long, deep divers and hence hard to detect visually, but produce loud

---

[a] Electronic mail: tiago@mcs.st-and.ac.uk; URL: http://www.creem.st-and.ac.uk/decaf/

0001-4966/2009/125(4)/1982/13/$25.00

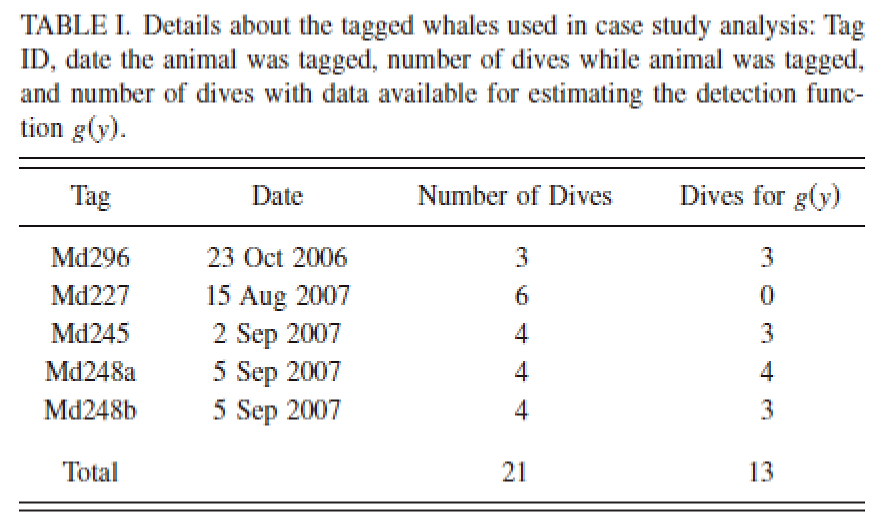


TABLE I. Details about the tagged whales used in case study analysis: Tag ID, date the animal was tagged, number of dives while animal was tagged, and number of dives with data available for estimating the detection function $g(y)$.

Tag	Date	Number of Dives	Dives for $g(y)$
Md296	23 Oct 2006	3	3
Md227	15 Aug 2007	6	0
Md245	2 Sep 2007	4	3
Md248a	5 Sep 2007	4	4
Md248b	5 Sep 2007	4	3
Total		21	13

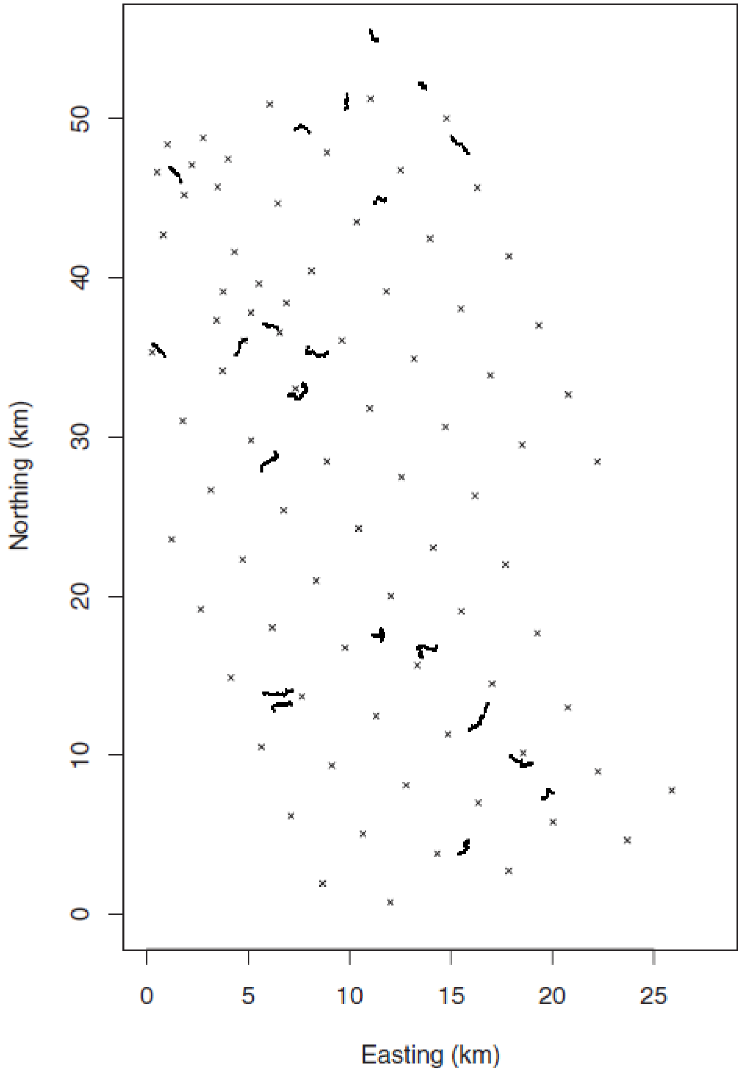


FIG. 1. The spatial layout of the AUTEC hydrophones that were recording during the collection of the primary survey data, represented by small crosses. Also shown as dots (perceived as solid lines) are the locations of the tagged Blainville's beaked whales when each click recorded on the DTag was produced.

Marques, T. A.; Thomas, L.; Ward, J.; DiMarzio, N. & Tyack, P. L. 2009 Estimating cetacean population density using fixed passive acoustic sensors: an example with Blainville's beaked whales *The Journal of the Acoustical Society of America* **125**:1982-1994

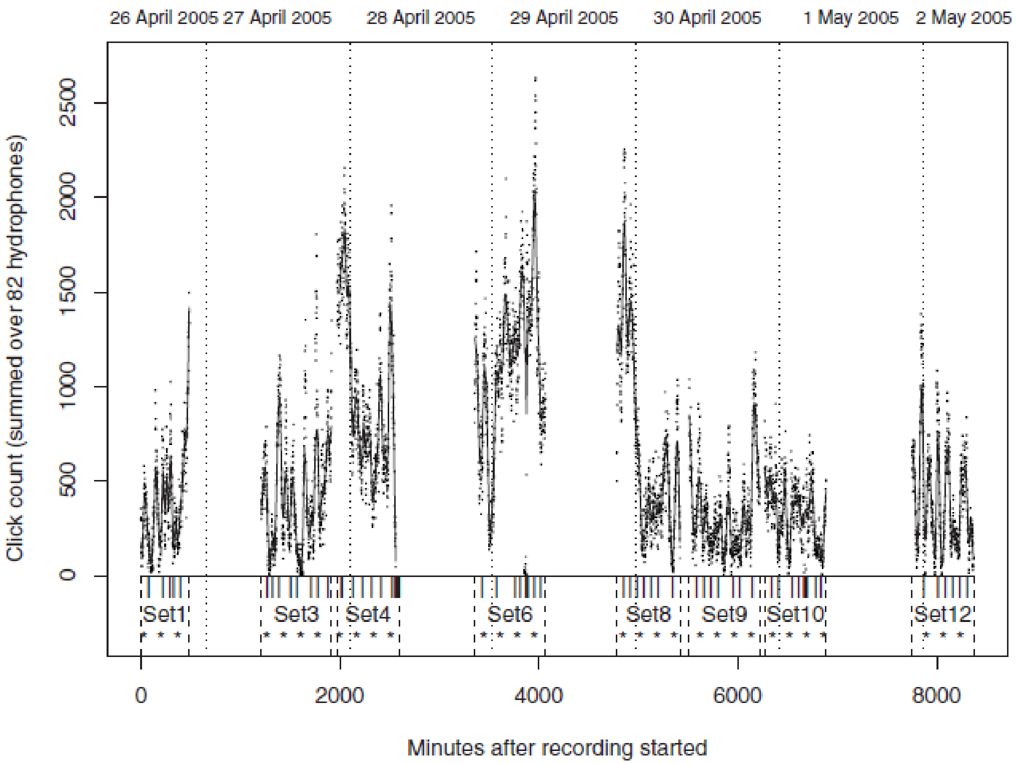


FIG. 2. Click counts per minute, summed over the 82 recording hydrophones, for the 6 day period of the primary dataset. Time is indexed as minutes since recording started. For operational reasons the data were divided into sets, and some sets (2, 5, 7 and 11) were not used. A standard lowess smooth of click counts over time is shown for the sets used. The small black vertical dashes ("|") are scattered minutes within the 8 sets used which were faulty and hence removed from the data. The sample periods used for the estimation of the false positive proportion are represented by " *". The limits of each day and set are represented by dotted and dashed lines, respectively.

Marques, T. A.; Thomas, L.; Ward, J.; DiMarzio, N. & Tyack, P. L. 2009 Estimating cetacean population density using fixed passive acoustic sensors: an example with Blainville's beaked whales *The Journal of the Acoustical Society of America* **125**:1982-1994

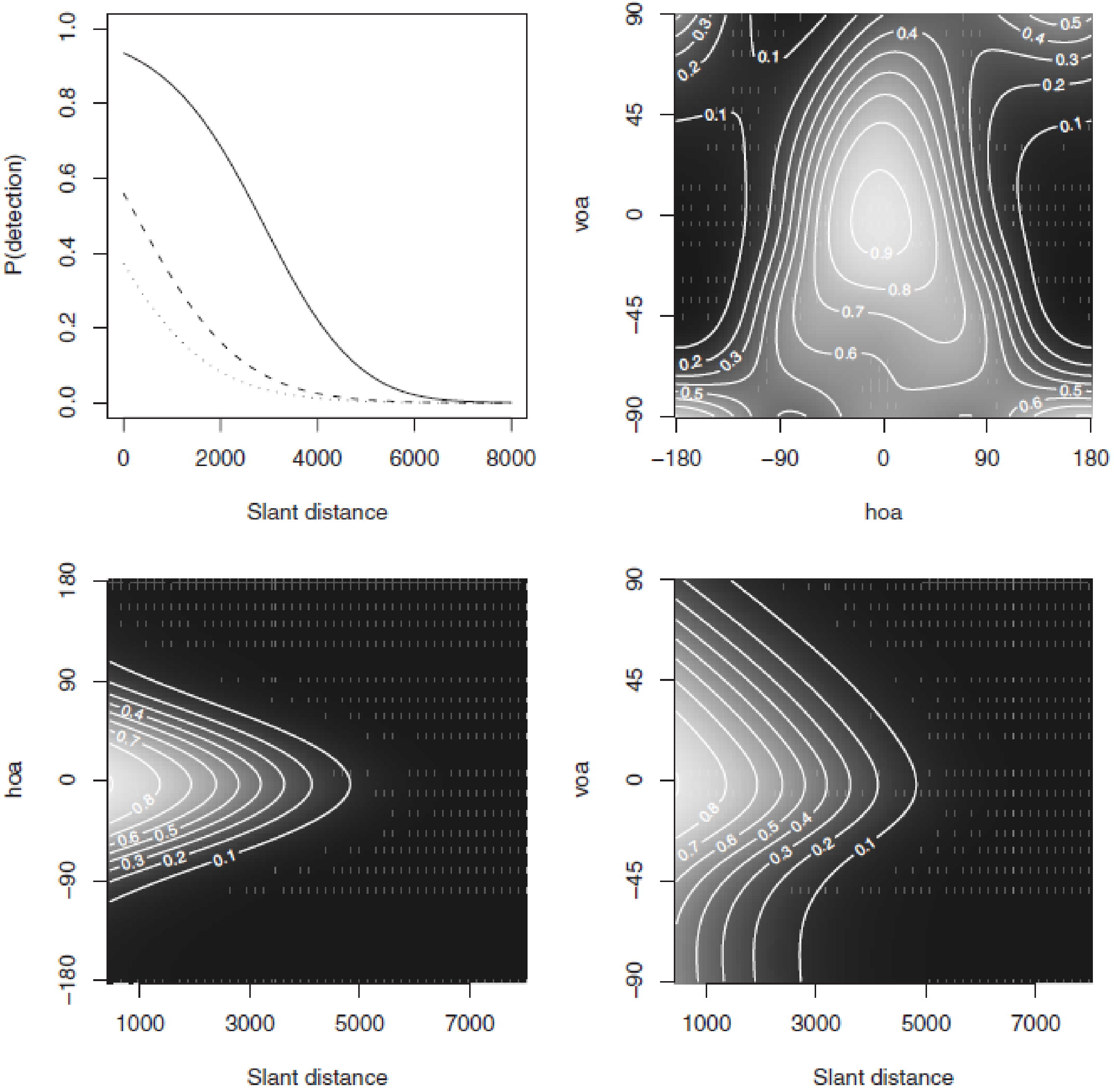
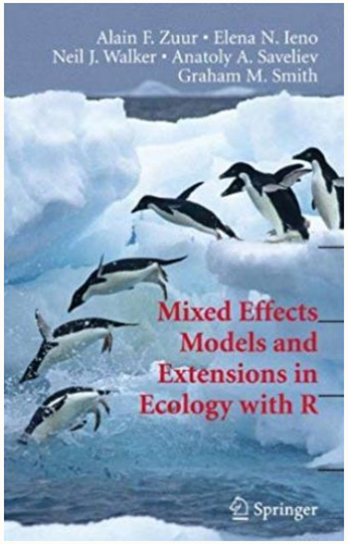


FIG. 3. The estimated detection function. Partial plots (on the response scale) of the fitted smooths for a binomial GAM model with slant distance and a 2D smooth of *hoa* and *voa*. For the top left plot, the off-axis angles are fixed at 0, 45, and 90° (respectively the solid, dashed, and dotted lines). Remaining plots are two-dimensional representations of the smooths, where black and white represent respectively an estimated probability of detection of 0 and 1. Distance (top right panel) and angle not shown (bottom panels) are fixed respectively at 0 m and 0 °.

Marques, T. A.; Thomas, L.; Ward, J.; DiMarzio, N. & Tyack, P. L. 2009 Estimating cetacean population density using fixed passive acoustic sensors: an example with Blainville's beaked whales *The Journal of the Acoustical Society of America* **125**:1982-1994

OUTRAS CLASSES DE MODELOS DE REGRESSÃO potencialmente uteis em ecologia

Outros modelos de regressão potencialmente uteis em ecologia


Alain F. Zuur • Elena N. Ieno
Neil J. Walker • Anatoly A. Saveliev
Graham M. Smith

Mixed Effects Models and Extensions in Ecology with R

Springer

Modelos para dados com excesso de zeros –
Um facto comum em contagens ecológicas

- Hurdle models – a single type of zeros
- Zero inflated models – more than a type of zeros

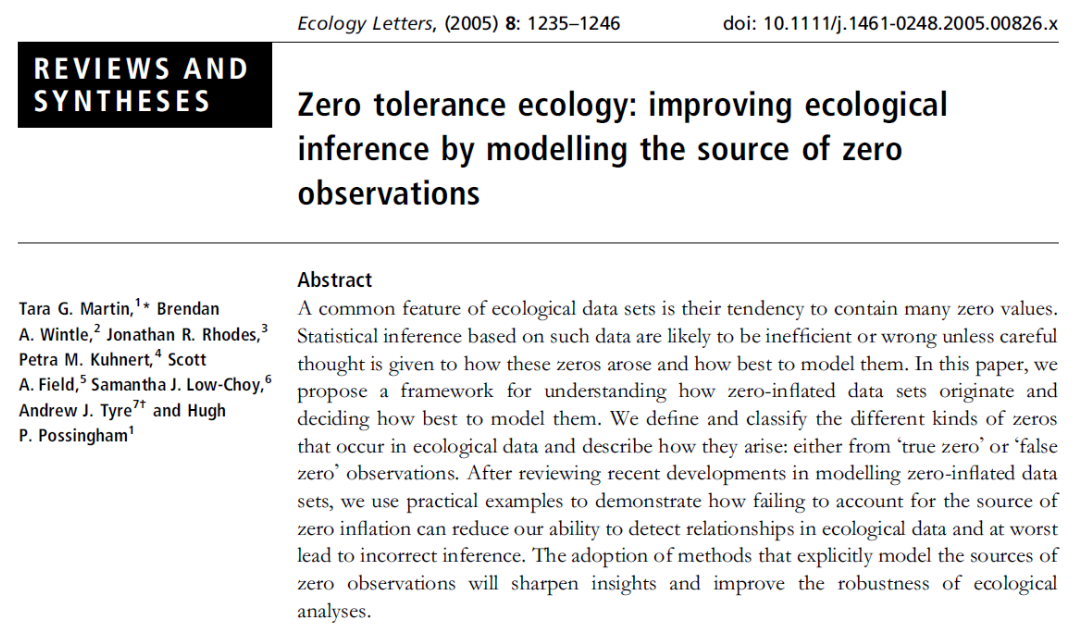
Modelos truncados (e.g. em 0)

Modelos mistos (com efeitos fixos e aleatórios: LMM, GLMM, GAMM) – permitem adicionar diferentes estruturas de correlação (ou seja não independência), e.g.

1. Correlação dentro de grupos
2. Correlação temporal
3. Correlação espacial

Generalized Estimating Equations (GEE's) – alternativa para modelos mistos, em que modelamos apenas o valor médio e a variância da variável resposta, não a sua distribuição

Ficam apenas com os nomes e alguns exemplos de aplicação prática, para poderem voltar a eles quando um dia durante o vosso trabalho perceberem que precisam deles!



Ecology Letters, (2005) **8:** 1235–1246 doi: 10.1111/j.1461-0248.2005.00826.x

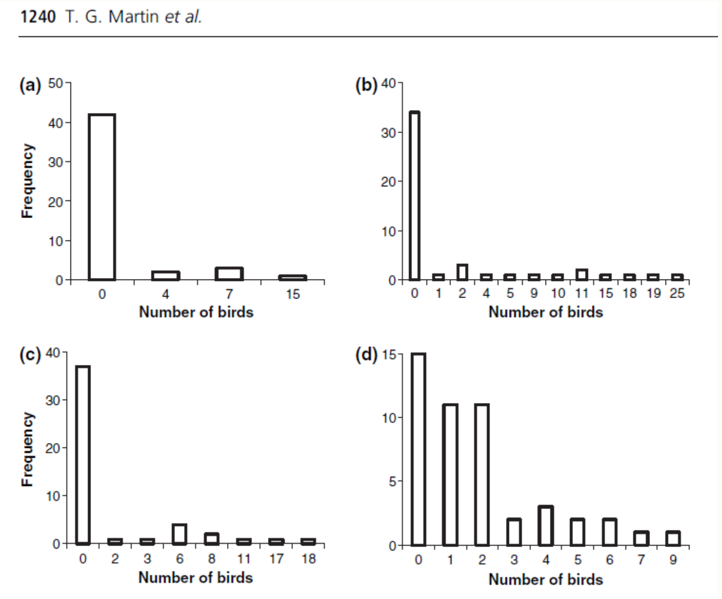
REVIEWS AND SYNTHESES

Zero tolerance ecology: improving ecological inference by modelling the source of zero observations

Tara G. Martin,[1]* Brendan A. Wintle,[2] Jonathan R. Rhodes,[3] Petra M. Kuhnert,[4] Scott A. Field,[5] Samantha J. Low-Choy,[6] Andrew J. Tyre[7]† and Hugh P. Possingham[1]

Abstract

A common feature of ecological data sets is their tendency to contain many zero values. Statistical inference based on such data are likely to be inefficient or wrong unless careful thought is given to how these zeros arose and how best to model them. In this paper, we propose a framework for understanding how zero-inflated data sets originate and deciding how best to model them. We define and classify the different kinds of zeros that occur in ecological data and describe how they arise: either from 'true zero' or 'false zero' observations. After reviewing recent developments in modelling zero-inflated data sets, we use practical examples to demonstrate how failing to account for the source of zero inflation can reduce our ability to detect relationships in ecological data and at worst lead to incorrect inference. The adoption of methods that explicitly model the sources of zero observations will sharpen insights and improve the robustness of ecological analyses.



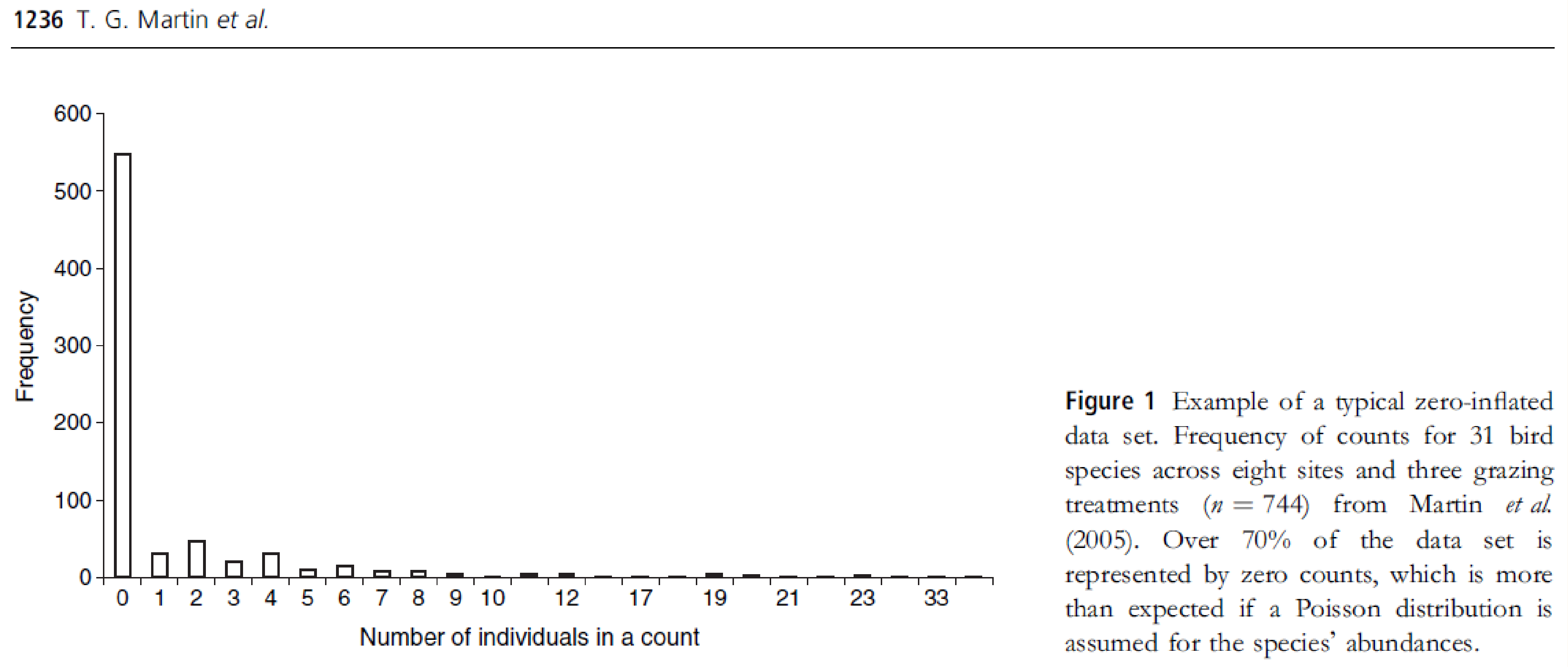


Figure 1 Example of a typical zero-inflated data set. Frequency of counts for 31 bird species across eight sites and three grazing treatments ($n = 744$) from Martin *et al.* (2005). Over 70% of the data set is represented by zero counts, which is more than expected if a Poisson distribution is assumed for the species' abundances.

Estimating group size from acoustical footprint to improve Blainville's beaked whale abundance estimation

Patrícia A. Jorge[a], Tiago A. Marques[b,*], Helena Mouriño[a], Len Thomas[c], David J. Moretti[d], Karin Dolan[d], Diane Claridge[e], Charlotte Dunn[e]

Under re-review in Applied Acoustics

To model the group size as a function of the acoustic footprint we considered the class of generalized linear models. Generalized additive models models (Wood, 2006) were attempted to explore non-linear relations, but no evidence for these was found and hence these were discarded. Because group size is a strictly positive number, we only considered zero-truncated GLMs. We considered the known group size for our sample of 51 dives as a Poisson response and the available covariates as explanatory variables. We considered Akaike's Information Criteria (AIC) for model selection, and performed visual inspection of qq-plots and residual plots for absolute goodness-of-fit. Models were implemented in the R software (R Core Team, 2017), with the help of package VGAM (Yee, 2010).

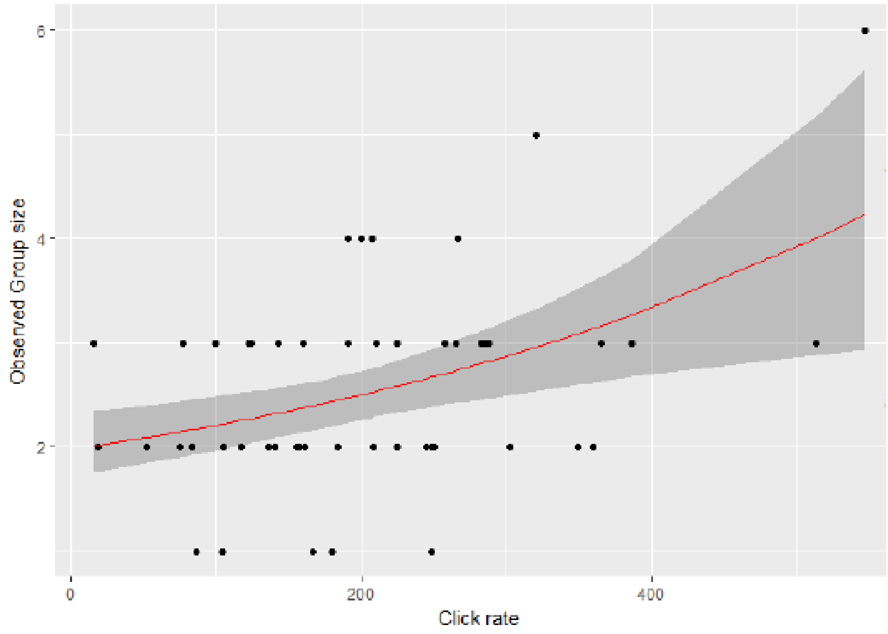


Figure 1: Observed group sizes and corresponding click rate (black dots), along with the modelled relationship (red line), and the model's bootstrap 95% percentile interval for the mean group size (grey area).

Review

Generalized linear mixed models: a practical guide for ecology and evolution

Benjamin M. Bolker[1], Mollie E. Brooks[1], Connie J. Clark[1], Shane W. Geange[2], John R. Poulsen[1], M. Henry H. Stevens[3] and Jada-Simone S. White[1]

“… Despite the availability of accurate techniques for estimating GLMM parameters in simple cases, complex GLMMs are challenging to fit and statistical inference such as hypothesis testing remains difficult…”

“…GLMMs are surprisingly challenging to use even for statisticians. Although several software packages can handle GLMMs (Table 1), few ecologists and evolutionary biologists are aware of the range of options or of the possible pitfalls.…”

https://biologyforfun.wordpress.com/2014/03/12/generalized-linear-mixed-models-in-ecology-and-in-r/

OTTER RESPONSE TO LARGE DAMS CONSTRUCTION 69

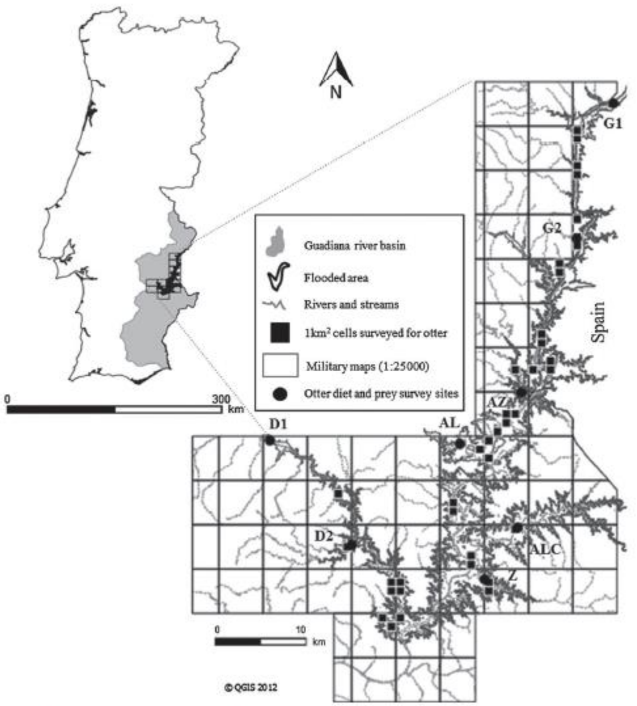
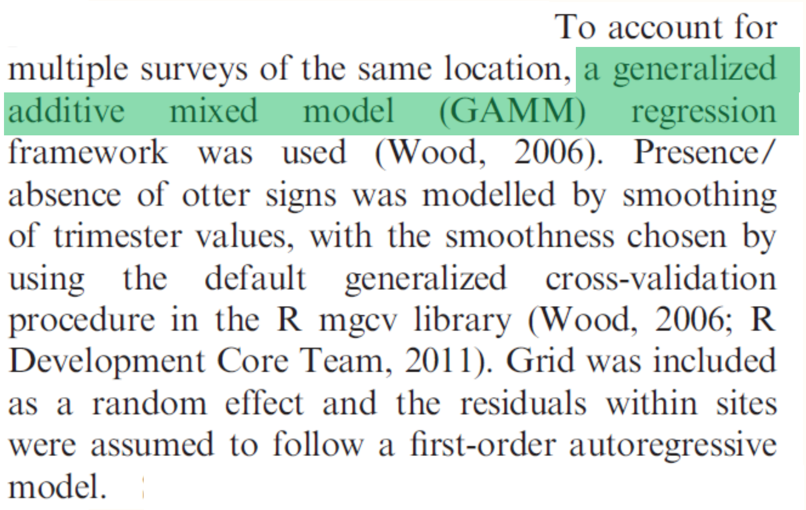


Figure 1. Location of the Alqueva Reservoir in southern Portugal, showing the 25 km^2 and the 1 km^2 (black squares) otter survey grid c circles) where otter diet and prey were assessed. Two sites (G1 and G2) were located in the main Guadiana River. Other sites were locat Guadiana River: the Azevel stream (AZ), Álamo stream (AL), Degebe stream (D1 and D2), Alcarrache stream (ALC) and Ze



To account for multiple surveys of the same location, a generalized additive mixed model (GAMM) regression framework was used (Wood, 2006). Presence/absence of otter signs was modelled by smoothing of trimester values, with the smoothness chosen by using the default generalized cross-validation procedure in the R mgcv library (Wood, 2006; R Development Core Team, 2011). Grid was included as a random effect and the residuals within sites were assumed to follow a first-order autoregressive model.

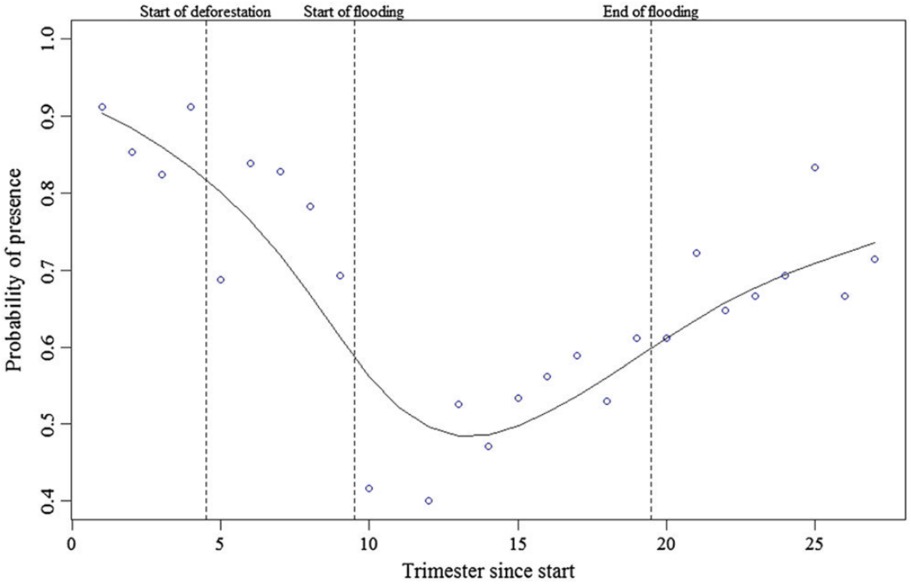


Figure 3. Probability of otter presence as a function of trimester in the flooded area of the Alqueva Reservoir. Data are represented by points and fitted model by a black line.

Pedroso, N. M.; Marques, T. A. & Santos-Reis, M. 2014 The response of otters to environmental changes imposed by the construction of large dams *Aquatic Conservation: Marine And Freshwater Ecosystems* **24**: 66–80
Já usado na Aula 3, slide 70.

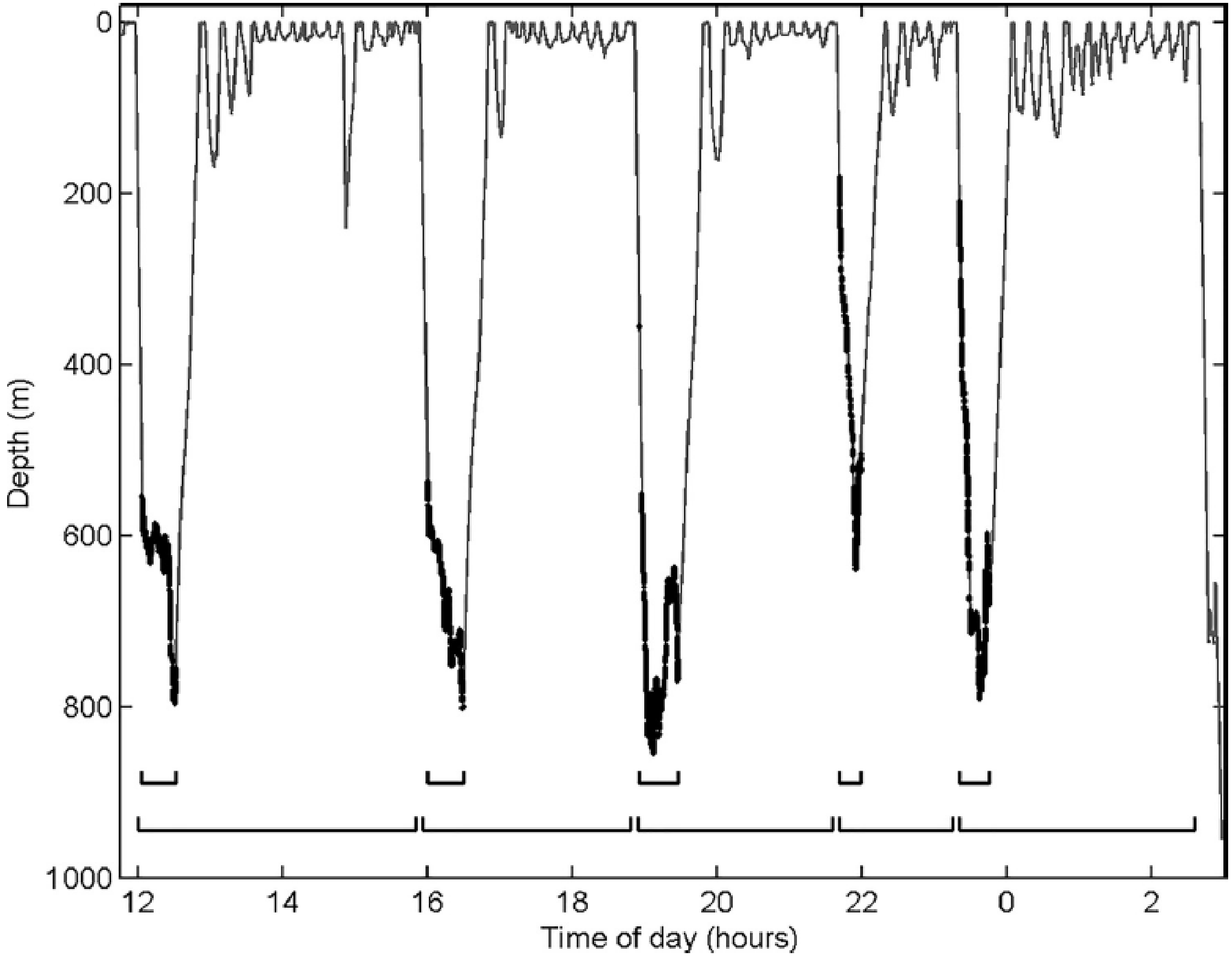


FIG. 1. Example dive profile of a Blainville's beaked whale tagged in the waters adjacent to El Hierro, Canary Islands. Bold sections indicate the presence of foraging clicks. Shorter, upper markers delineate vocal periods, while lower, longer markers indicate the lengths of individual dive cycles. The final dive featured tag detachment and was not analyzed.

Warren, V. E.; Marques, T. A.; Harris, D.; Tyack, P. L.; Thomas, L.; de Soto, N. A.; Hickmott, L. & Johnson, M. P. 2017 Spatio-temporal variation in click production rates of beaked whales: implications for passive acoustic density estimation The Journal of the Acoustical Society of America 141: 1962-1974

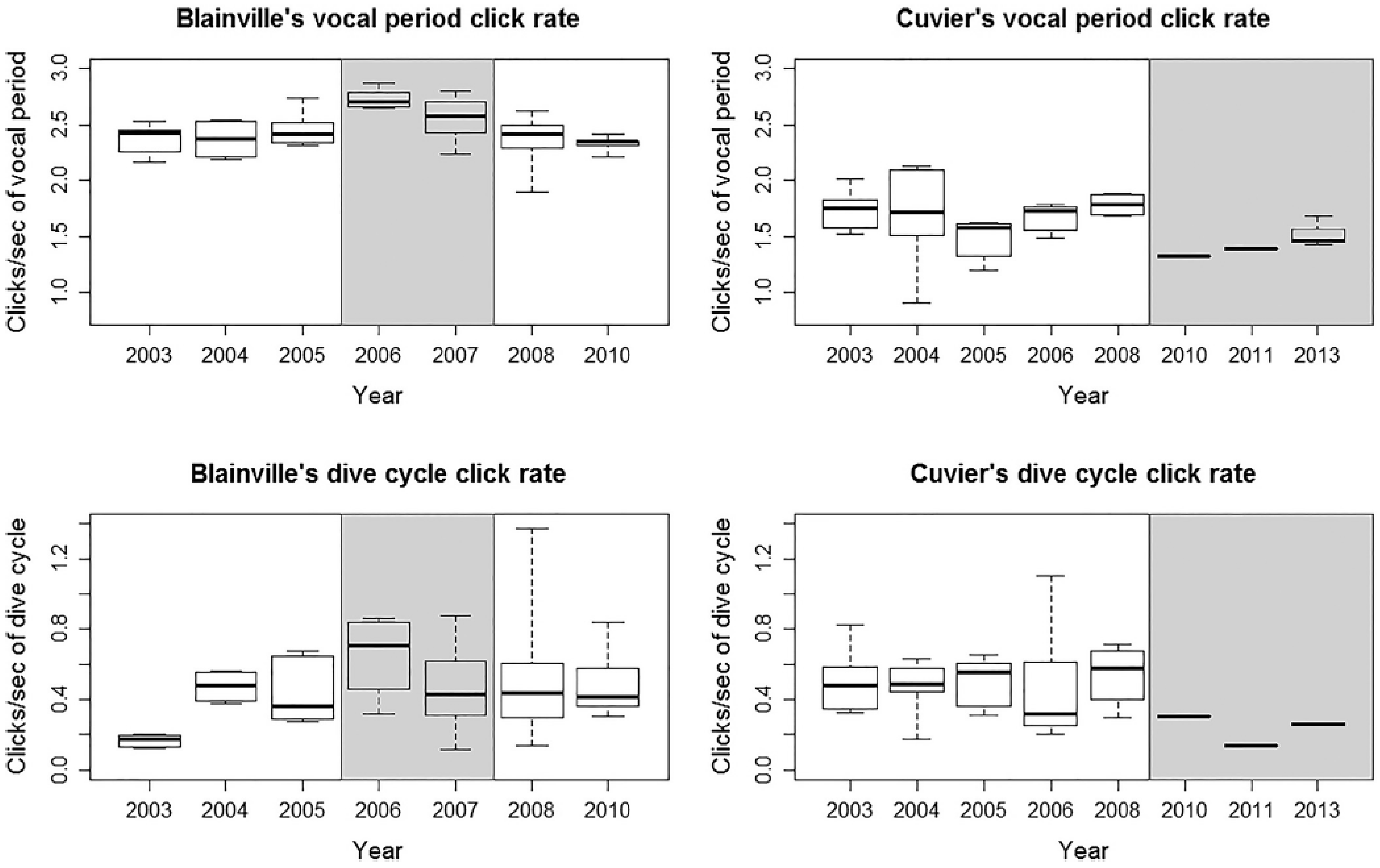
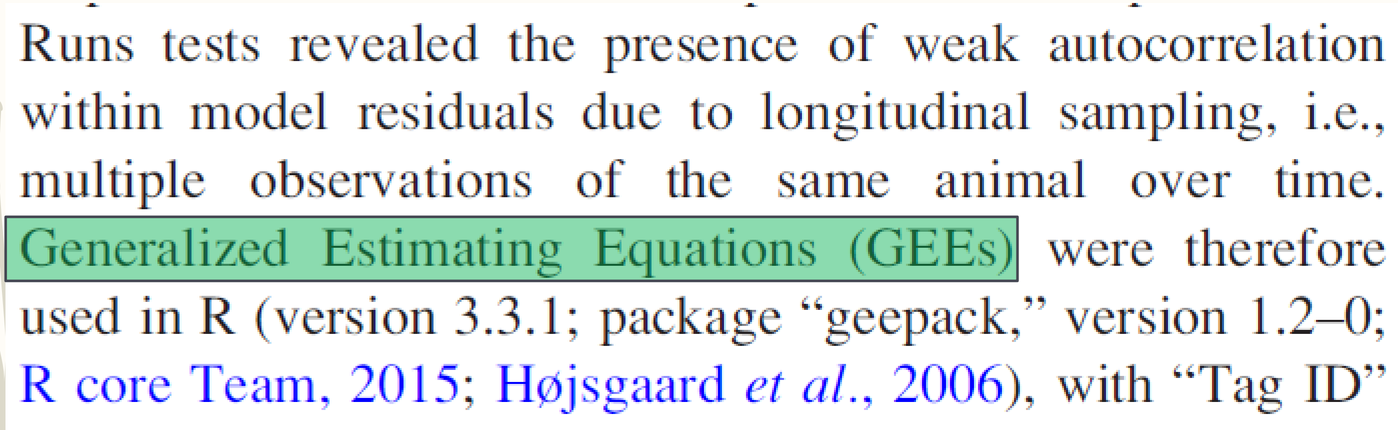


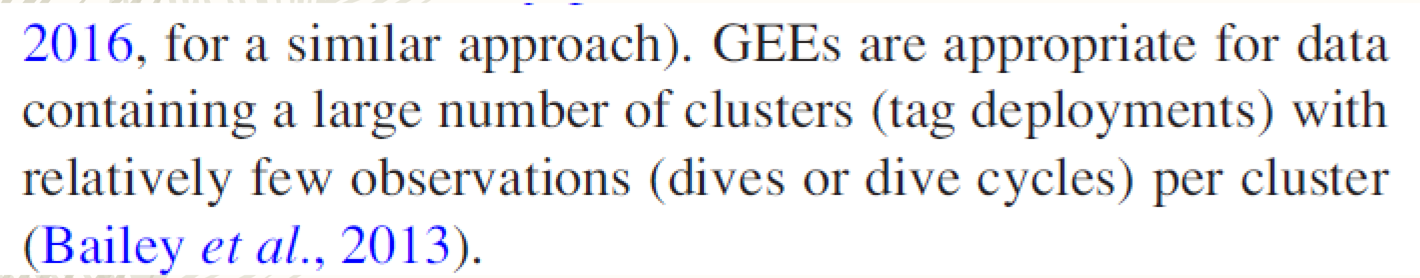
FIG. 4. Inter-annual variation in vocal period and dive cycle click production rates for Blainville's (left) and Cuvier's (right) beaked whales. Box plots consist of median, interquartile range and maximum/minimum extremes. In the Blainville's data, boxes in white areas represent animals tagged in El Hierro and boxes in grey areas (2006 and 2007) indicate tags deployed in the Bahamas. In the Cuvier's plots, boxes in the white area represent Liguria, and boxes in the grey area (2010, 2011, and 2013) are southern California deployments. See Table I for respective sample sizes. Y axes scales differ between vocal period plots (upper) and dive cycle plots (lower).

Warren, V. E.; Marques, T. A.; Harris, D.; Tyack, P. L.; Thomas, L.; de Soto, N. A.; Hickmott, L. & Johnson, M. P. 2017 Spatio-temporal variation in click production rates of beaked whales: implications for passive acoustic density estimation The Journal of the Acoustical Society of America 141: 1962-1974



Runs tests revealed the presence of weak autocorrelation within model residuals due to longitudinal sampling, i.e., multiple observations of the same animal over time. Generalized Estimating Equations (GEEs) were therefore used in R (version 3.3.1; package "geepack," version 1.2–0; R core Team, 2015; Højsgaard *et al.*, 2006), with "Tag ID"

…

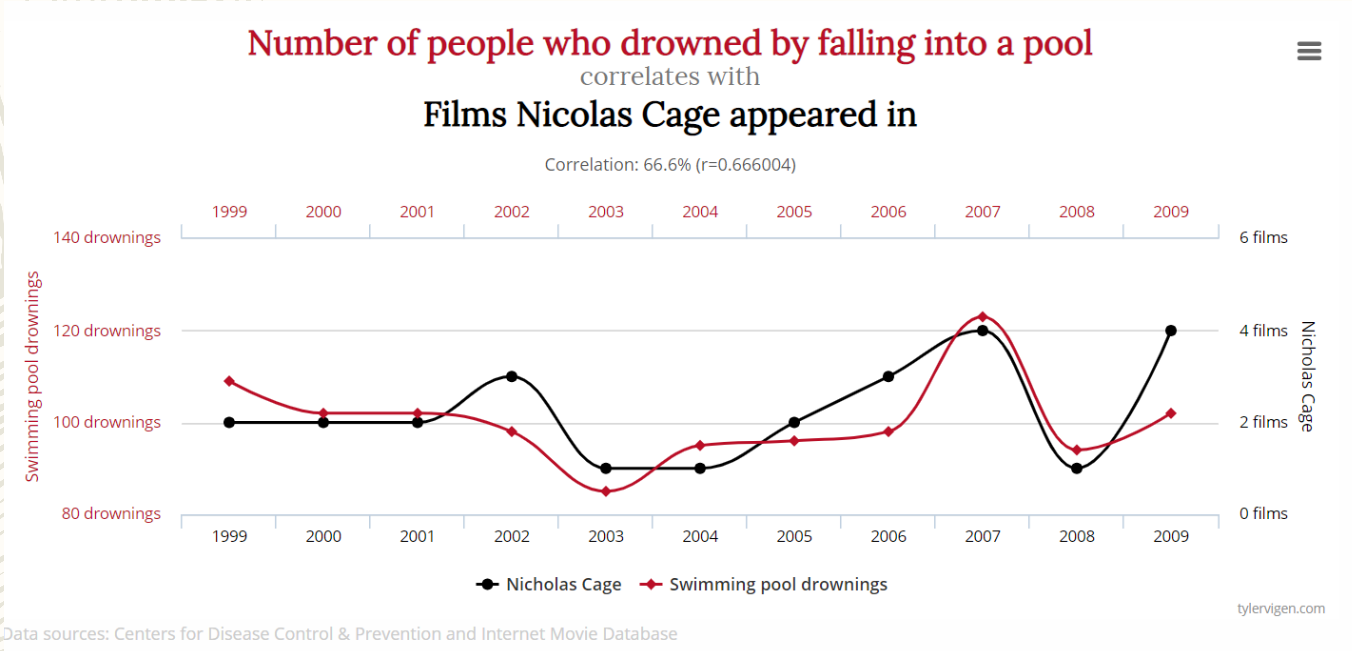


2016, for a similar approach). GEEs are appropriate for data containing a large number of clusters (tag deployments) with relatively few observations (dives or dive cycles) per cluster (Bailey *et al.*, 2013).

Warren, V. E.; Marques, T. A.; Harris, D.; Tyack, P. L.; Thomas, L.; de Soto, N. A.; Hickmott, L. & Johnson, M. P. 2017 Spatio-temporal variation in click production rates of beaked whales: implications for passive acoustic density estimation The Journal of the Acoustical Society of America 141: 1962-1974

ecologia numérica

correlação

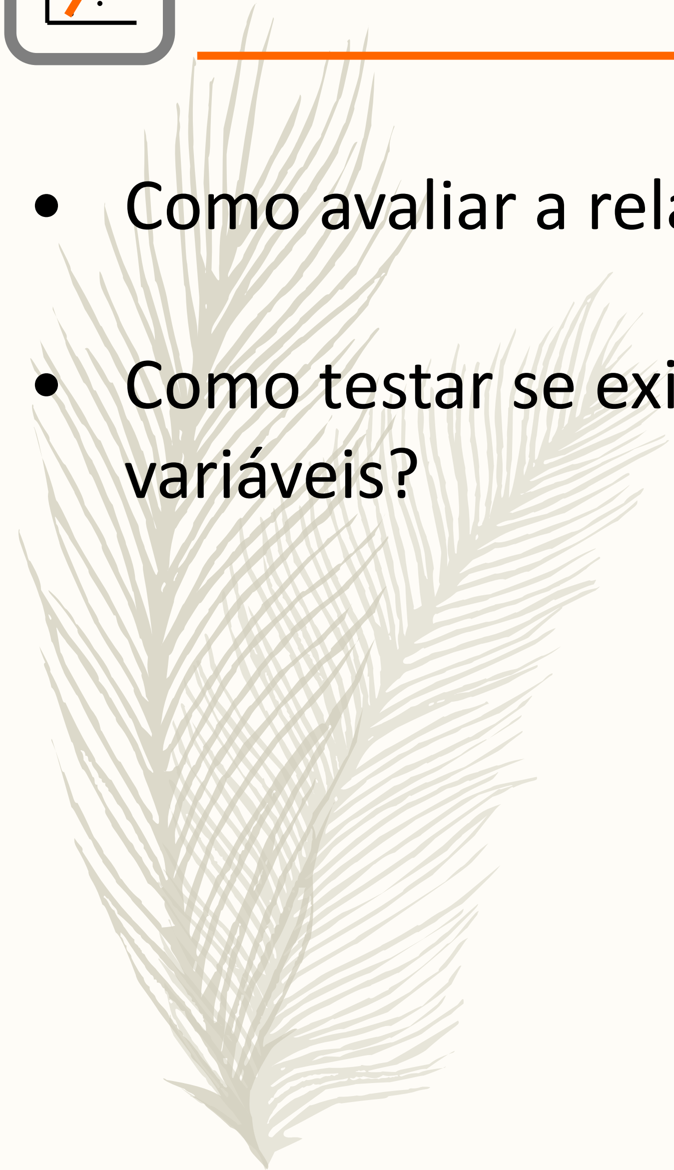


Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in

Correlation: 66.6% (r=0.666004)

Data sources: Centers for Disease Control & Prevention and Internet Movie Database

correlação



- Como avaliar a relação entre variáveis (contínuas)?
- Como testar se existe uma relação forte entre variáveis?

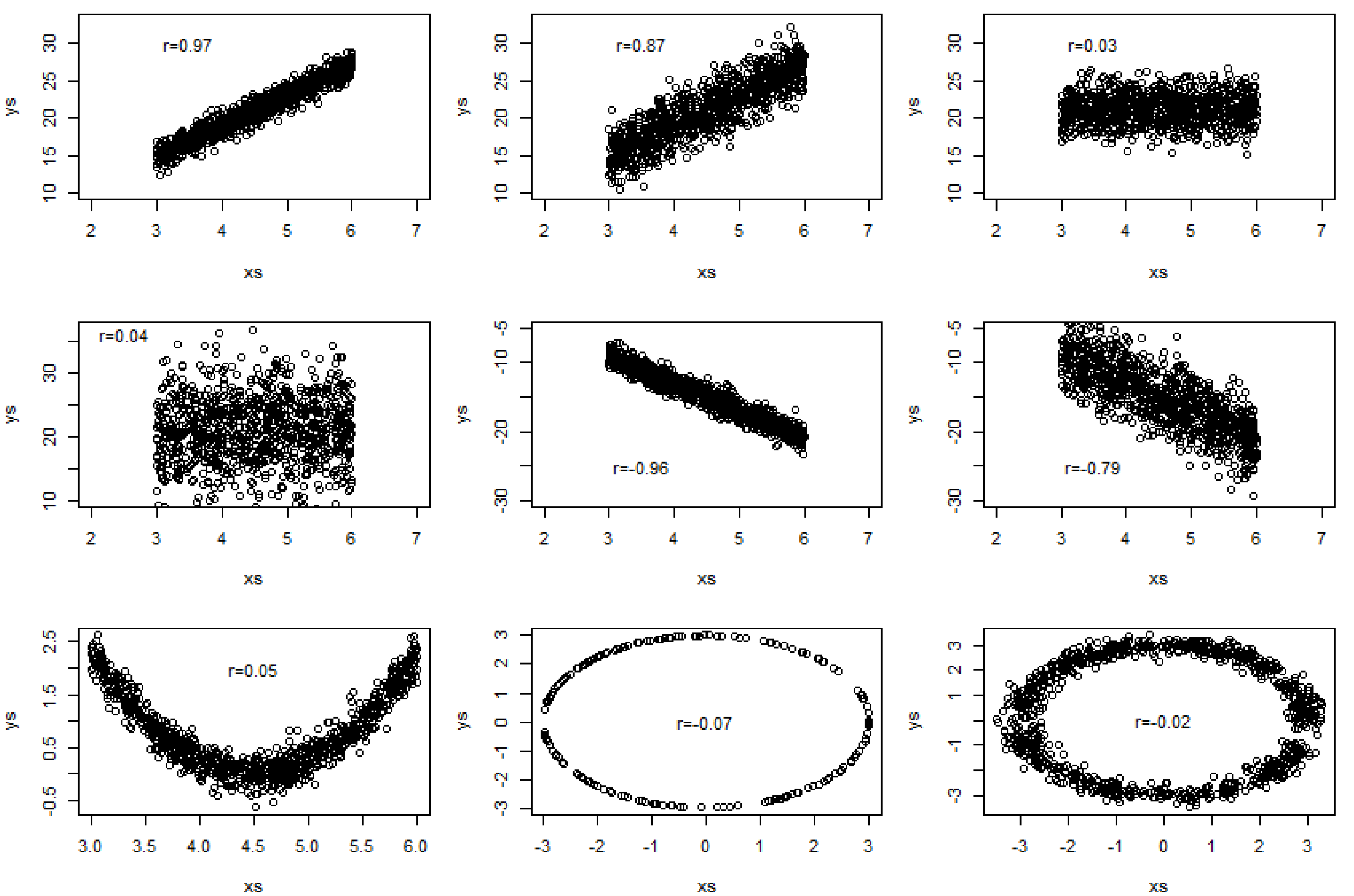
Correlação simples

Expressa a variação conjunta entre duas variáveis, numa situação em que nenhuma é necessariamente dependente da outra.

r – coeficiente de correlação de Pearson

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

Correlação (linear) simples



Testes de hipóteses na análise de correlação

Testes t

Hipótese:

H0: $\rho=0$
H1: $\rho \neq 0$

Estatística de teste:

$$\boxed{t = \frac{r}{s_r}}$$

onde

$$\boxed{s_r = \sqrt{\frac{1-r^2}{n-2}}}$$

Valor crítico:

$$\boxed{t_{\alpha(2),n-2}}$$

Critério de decisão:

Rejeitar H0 se: $\boxed{|t| > t_{\alpha(2),n-2}}$

Não rejeitar H0 caso contrário

Testes de hipóteses na análise de correlação

Testes F

Hipótese:

H0: $\rho=0$
H1: $\rho \neq 0$

Estatística de teste:

$$\boxed{F=\frac{1+|r|}{1-|r|}}$$

Valor crítico:

$$\boxed{F_{\alpha,n-2,n-2}}$$

Critério de decisão:

Rejeitar H0 se: $\boxed{F > F_{\alpha,n-2,n-2}}$

Não rejeitar H0 caso contrário

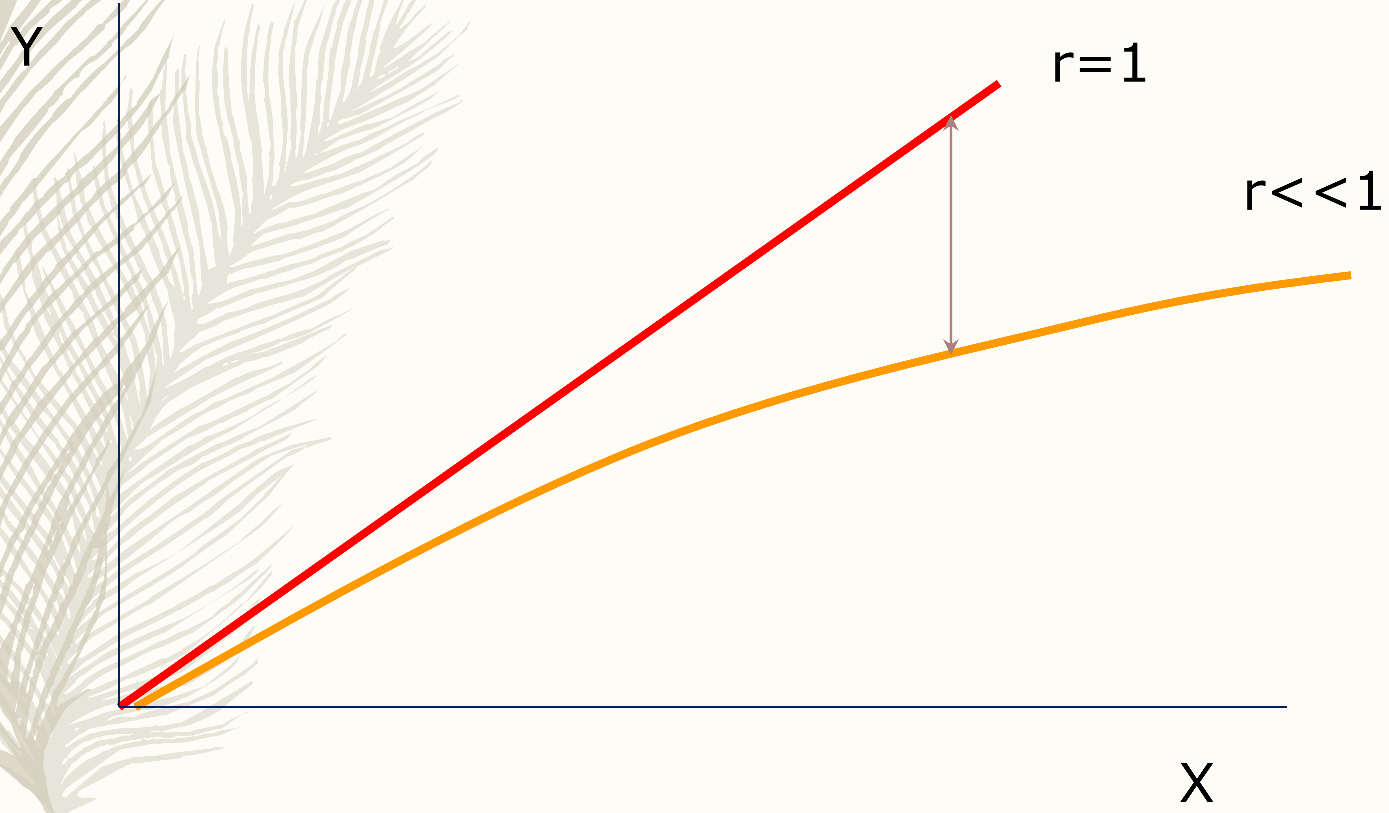
Correlação de Pearson (r ou ρ)

Pressupostos:

- A amostra é proveniente de populações normais com variâncias iguais (distribuição normal bivariada).

- Apesar dos teste F e t (que são equivalentes) serem relativamente robustos a violações moderadas dos pressupostos, o coeficiente de correlação pode não traduzir da melhor forma a relação entre as duas variáveis.

Correlação de Pearson



Correlação não-paramétrica

- Expressa a variação conjunta de duas variáveis, mas utiliza as ordens dos valores para obter uma estimativa de r.

r_S – coeficiente de correlação de Spearman

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

onde d_i=(*ordem de* X_i – *ordem de* Y_i)

Teste ao coeficiente de correlação de Spearman

Hipótese:

H0: $\rho=0$

H1: $\rho \neq 0$

$$r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n^3 - n}$$

Valor crítico:

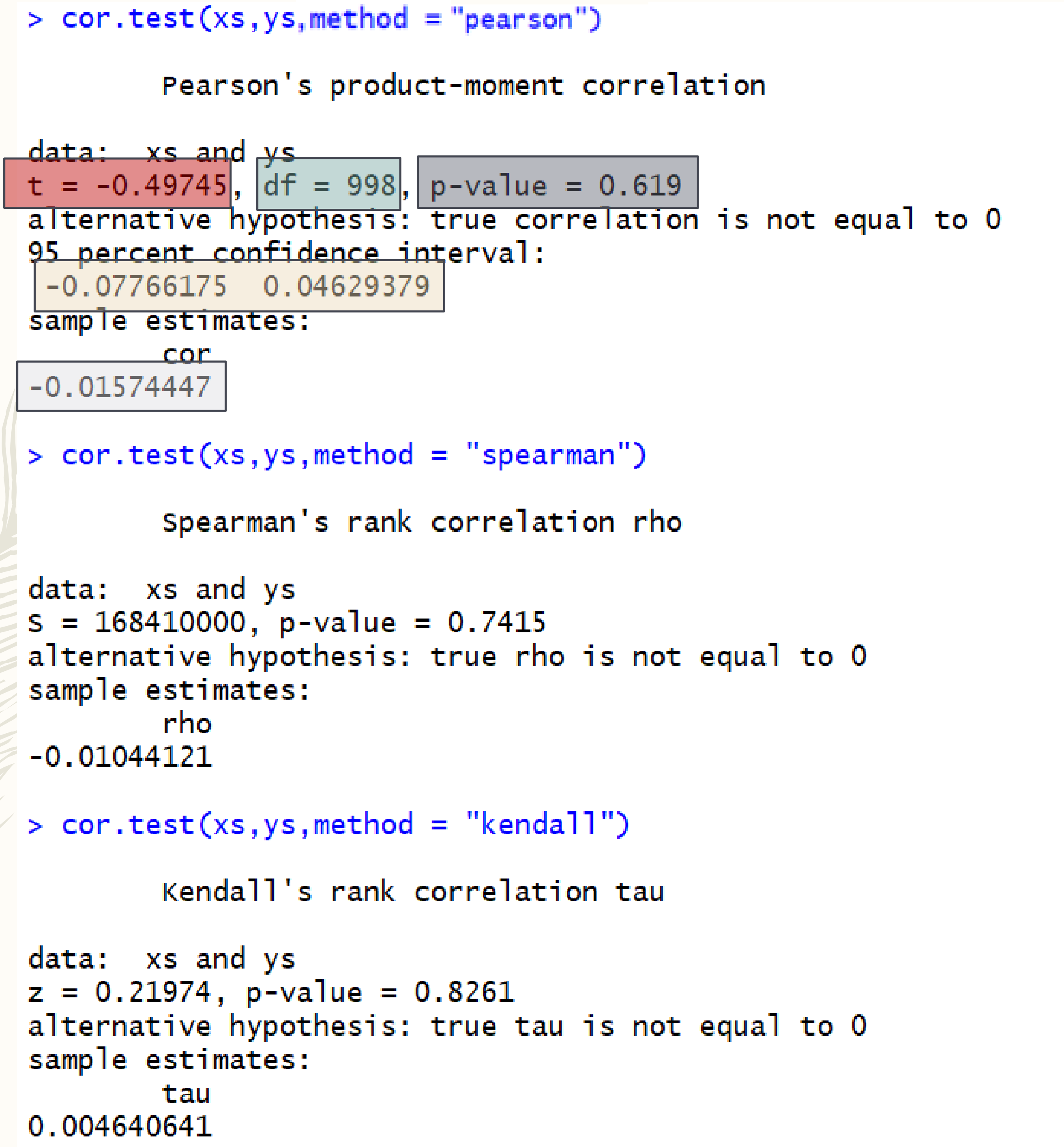
$$r_{S\,\alpha(2),n}$$

Critério de decisão:

Rejeitar H0 se: $r_S > r_{S\,\alpha(2),n}$

Não rejeitar H0 caso contrário

$H0:\rho=0$ vs. $H0:\rho\neq0$



```
> cor.test(xs,ys,method = "pearson")

        Pearson's product-moment correlation

data:  xs and ys
t = -0.49745, df = 998, p-value = 0.619
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.07766175  0.04629379
sample estimates:
        cor
-0.01574447

> cor.test(xs,ys,method = "spearman")

        Spearman's rank correlation rho

data:  xs and ys
S = 168410000, p-value = 0.7415
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01044121

> cor.test(xs,ys,method = "kendall")

        Kendall's rank correlation tau

data:  xs and ys
z = 0.21974, p-value = 0.8261
alternative hypothesis: true tau is not equal to 0
sample estimates:
        tau
0.004640641
```