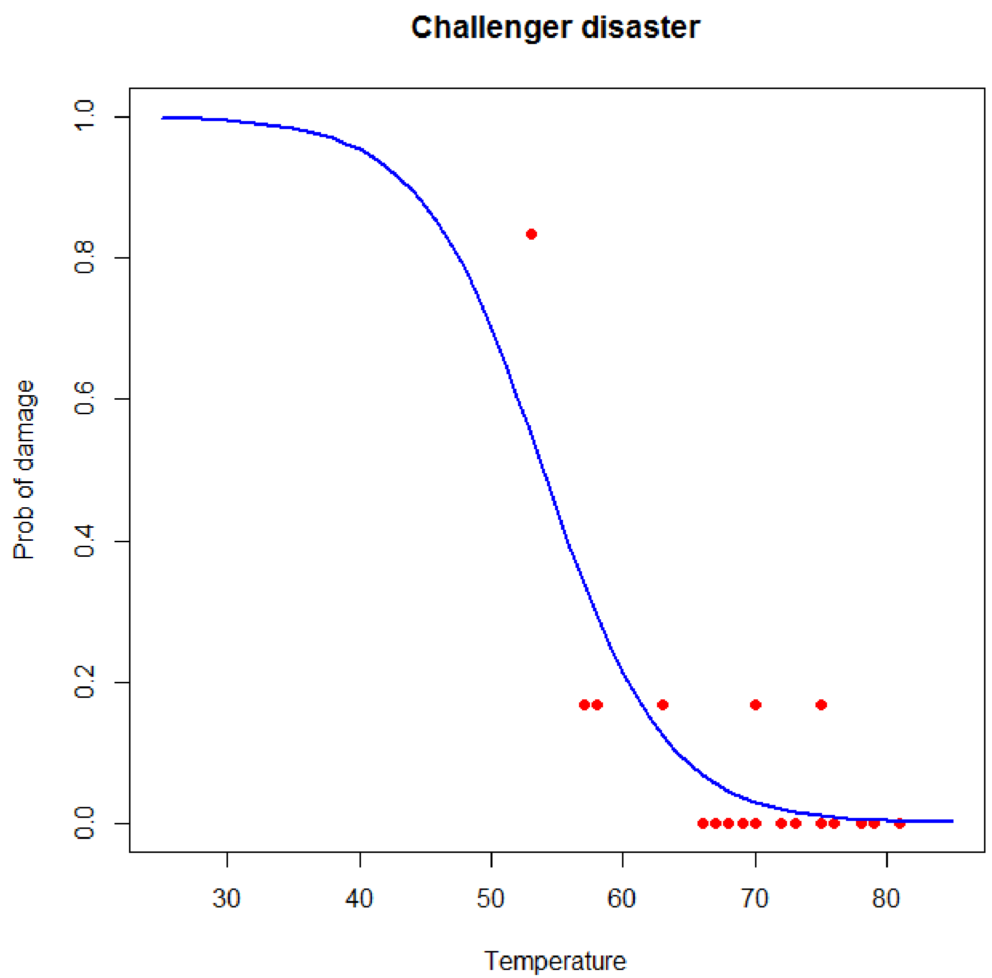
Generalized linear modeling with R



Source: Faraway (2006).

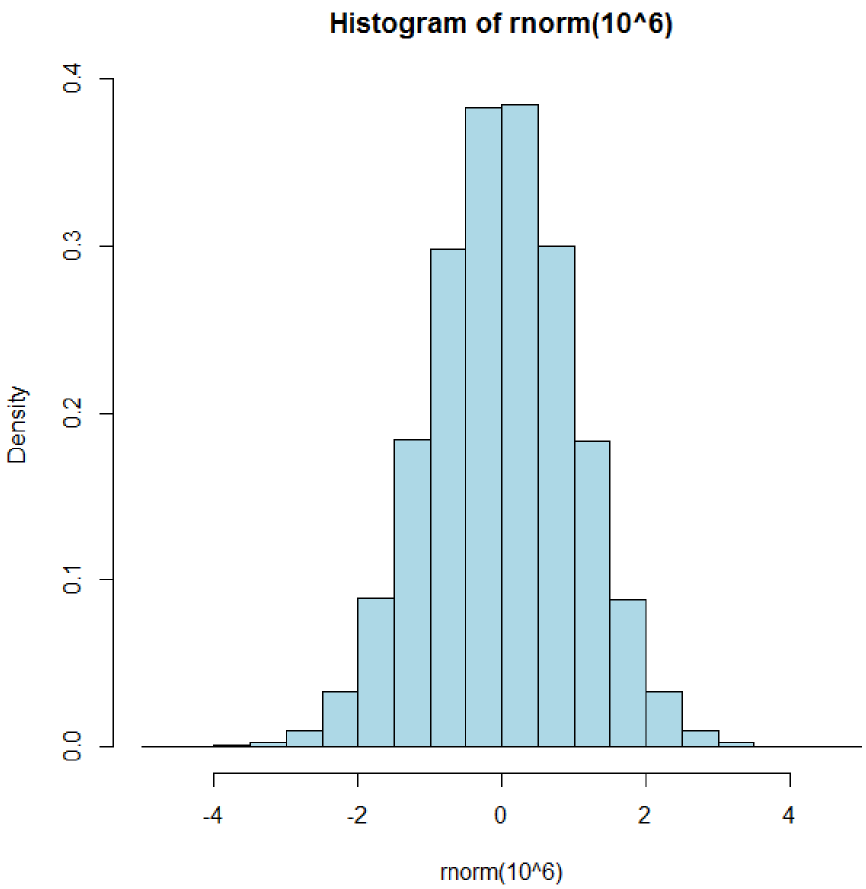
Why not linear regression?

- 0/1 data (activity, infection, presence, sex)
- proportional data (0 – 100 %)
- number of successes vs. number of failures
- count data (species richness)

Before getting started, we should think about

1. distribution of the **response**
2. systematic part (covariates)
3. link-function between expected value of the response and the systematic part

Normal distribution



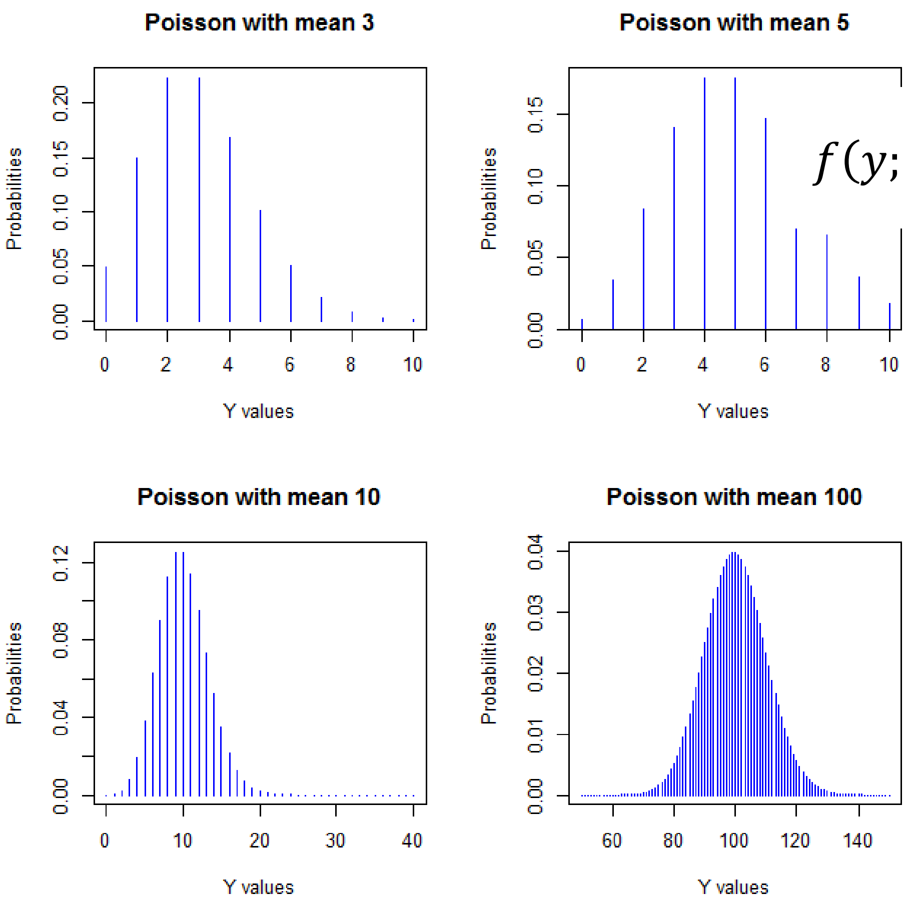
$$f(y_i; \boldsymbol{\mu}, \boldsymbol{\sigma}) = \frac{1}{\sigma\sqrt{2\mu}} * e^{\frac{(y_i \; - \mu)^2}{2\sigma^2}}$$

defined by mean

and standard deviation

Poisson distribution

seit 1558



Modified after: Zuur et al. (2009).

$$f(y;\, \boldsymbol{\mu}) = \frac{\mu^{y} * e^{-\mu}}{y!} \qquad y \geq 0, y \text{ integer}$$

mean = variance

-> larger mean values have also a larger variance

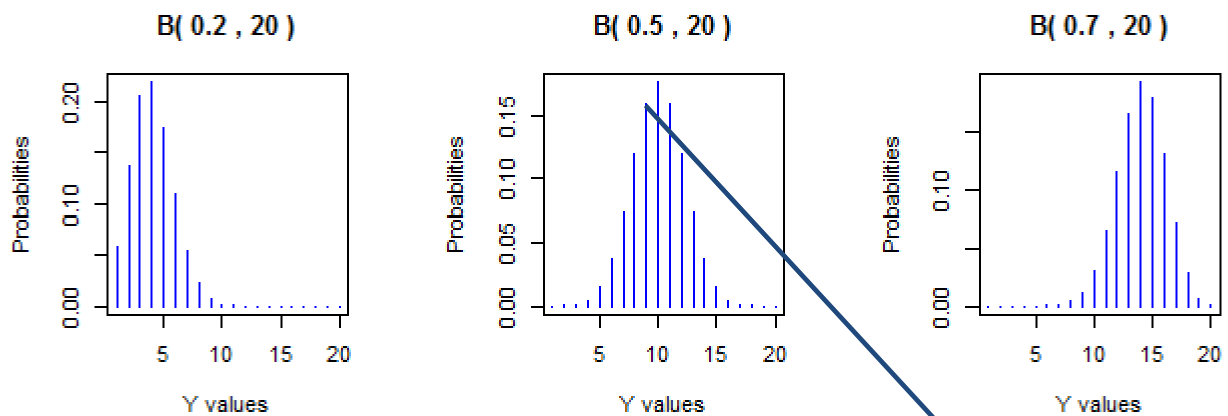
overdispersion: if the variance is larger than the mean

Bernoulli and binomial distribution

seit 1558

- Tossing a coin (head or tail)
- Bernoulli distribution = binomial distribution with N = 1

$$f(y;\,\pi) = \binom{N}{y} * \pi^{y} * (1-\pi)^{N-y}$$



Modified after: Zuur et al. (2009).

(20!/(9! * 11!))*0.5^9*(1-0.5)^11 = 0.16

Further common distributions

1. **Negative binomial distribution** -> quick and dirty solution for overdispersion
2. **Gamma distribution** -> can be used for a continous response variable that has positive values (> 0)

Predictor function

seit 1558

$$g(x_i) = \boldsymbol{\alpha} + \boldsymbol{\beta_1 X_{1i}} + \cdots + \boldsymbol{\beta_n X_{ni}}$$

Poisson predictor function

seit 1558

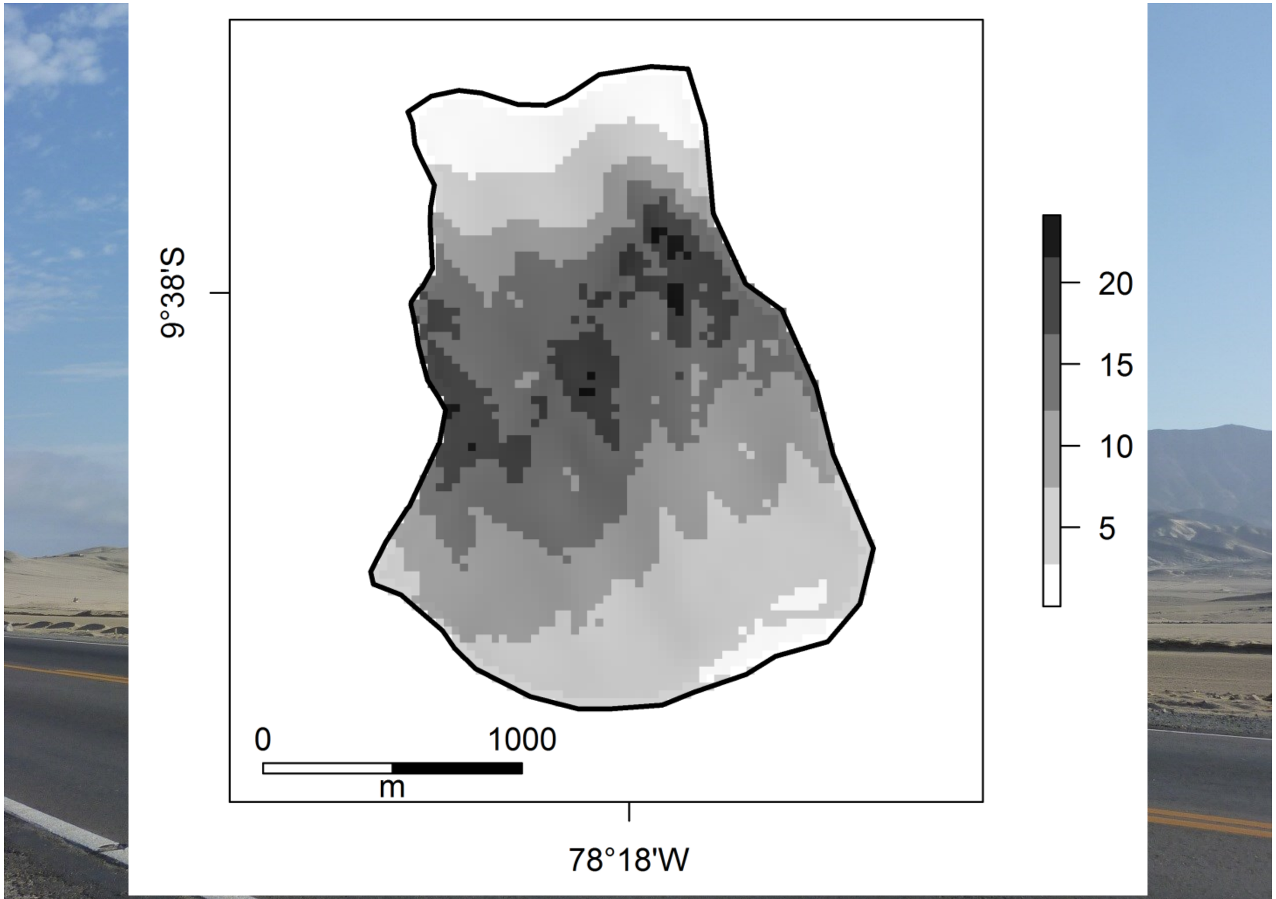
1. distribution -> increase in spread
2. density curves avoid negative realizations
3. exponential link -> no negative fitted values
4. Maximum likelihood algorithm (see Zuur 2009: 213)

$$Y_i \sim P(\mu_i) \text{ and } E[Y_i] = \mu_i = e^{\boldsymbol{\alpha}+\boldsymbol{\beta}_i \boldsymbol{X}_{1i}+\cdots+\boldsymbol{\beta}_n \boldsymbol{X}_{ni}}$$

$$\log(E[Y_i]) = \boldsymbol{\alpha} + \boldsymbol{\beta}_i \boldsymbol{X}_{1i} + \cdots + \boldsymbol{\beta}_n \boldsymbol{X}_{ni}$$

log-link

Modeling species richness

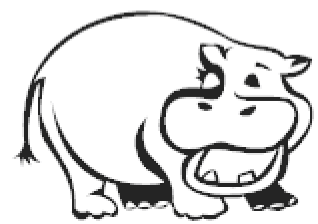


Source: Muenchow et al. (2013).

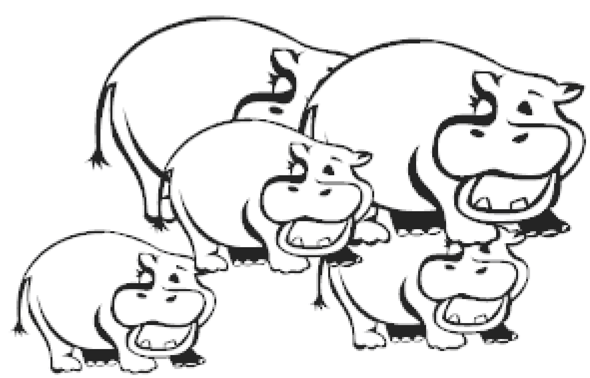
Binomial GLM

1. Express 0 and 1 as probability P
2. Apply a series of transformations on P
3. back-transform to 0-1 intervall

Absence (= 0)



I am not here, because the habitat is not good!



Here we are!

Presence (= 1)

modified after Zuur et al. (2009).

Concept of Odds

seit 1558

$$O_i = \frac{\pi_i}{1 - \pi_i}$$

with O_i= Odds for the *i*th observation

π_i = Probability of the *i*th observation

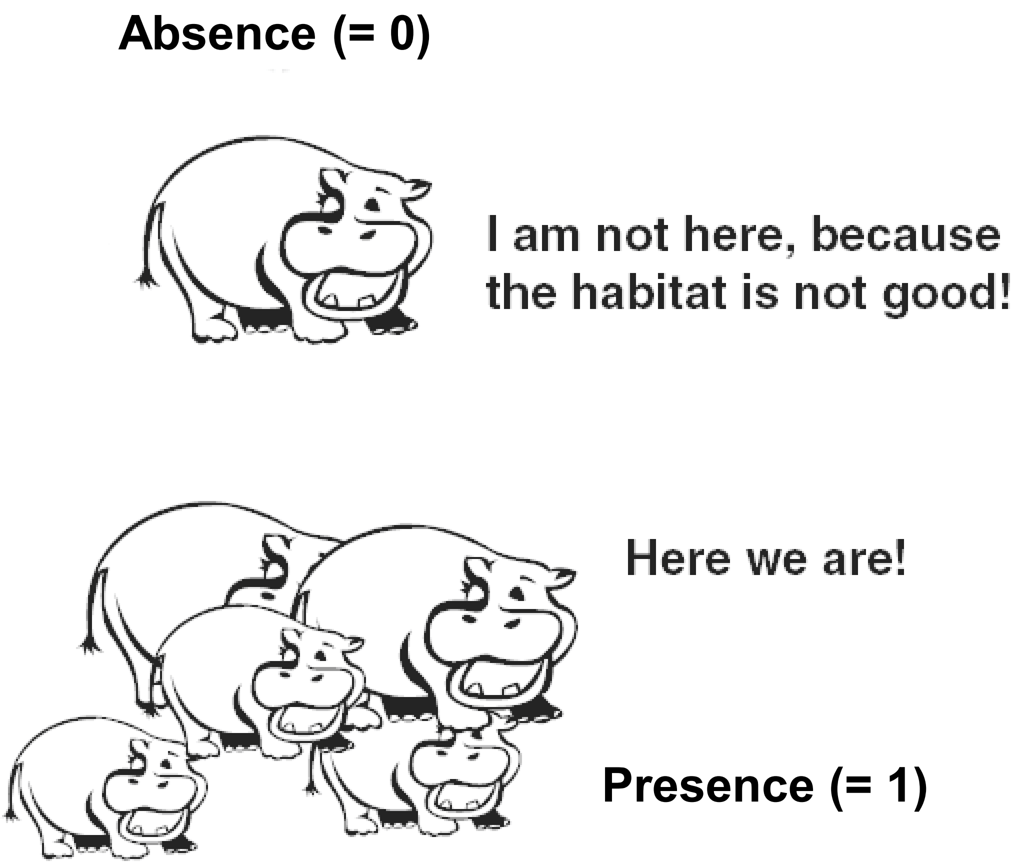
-> Value not any longer restricted to 0 and 1!!!

Concept of Odds

$$O_i = \frac{0.9}{(1-0.9)} = \mathbf{9}$$

Odds of 9 -> it is **9 times more likely** to record a hippo than not to record one.

Or: In **9 from 10** plots you will find a hippo.



modified after Zuur et al. (2009).

log-Odds as a function of predictor values

seit 1558

log-link

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta$$

with η = predictor function

-> Negative values are possible

Solving for π_i

$$\pi_i = \frac{e^{\eta}}{1 + e^{\eta}}$$

-> term lies always between 0 and 1 (back-transformation)!!!

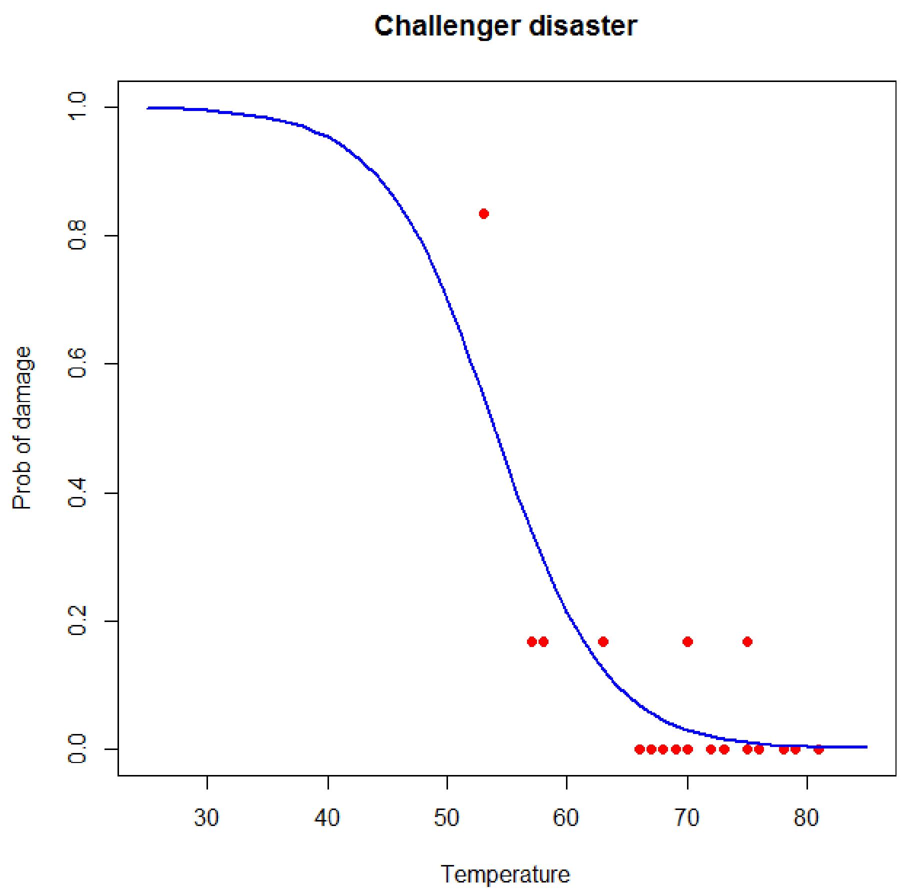
Logistic regression prediction formula

seit 1558

$$Y_i \sim B(1; P_i) \text{ and } E[Y_i] = P_i = \pi_i = \frac{e^{\boldsymbol{\alpha+\beta_i X_{1i}+\cdots+\beta_n X_{ni}}}}{1 + e^{\boldsymbol{\alpha+\beta_i X_{1i}+\cdots+\beta_n X_{ni}}}}$$

- appropriate distribution (Bernoulli)
- Maximum likelihood algorithm (see Zuur et al., 2007: 93)

Example: Challenger disaster



Source: Faraway (2006).

Modeling landslide susceptibility

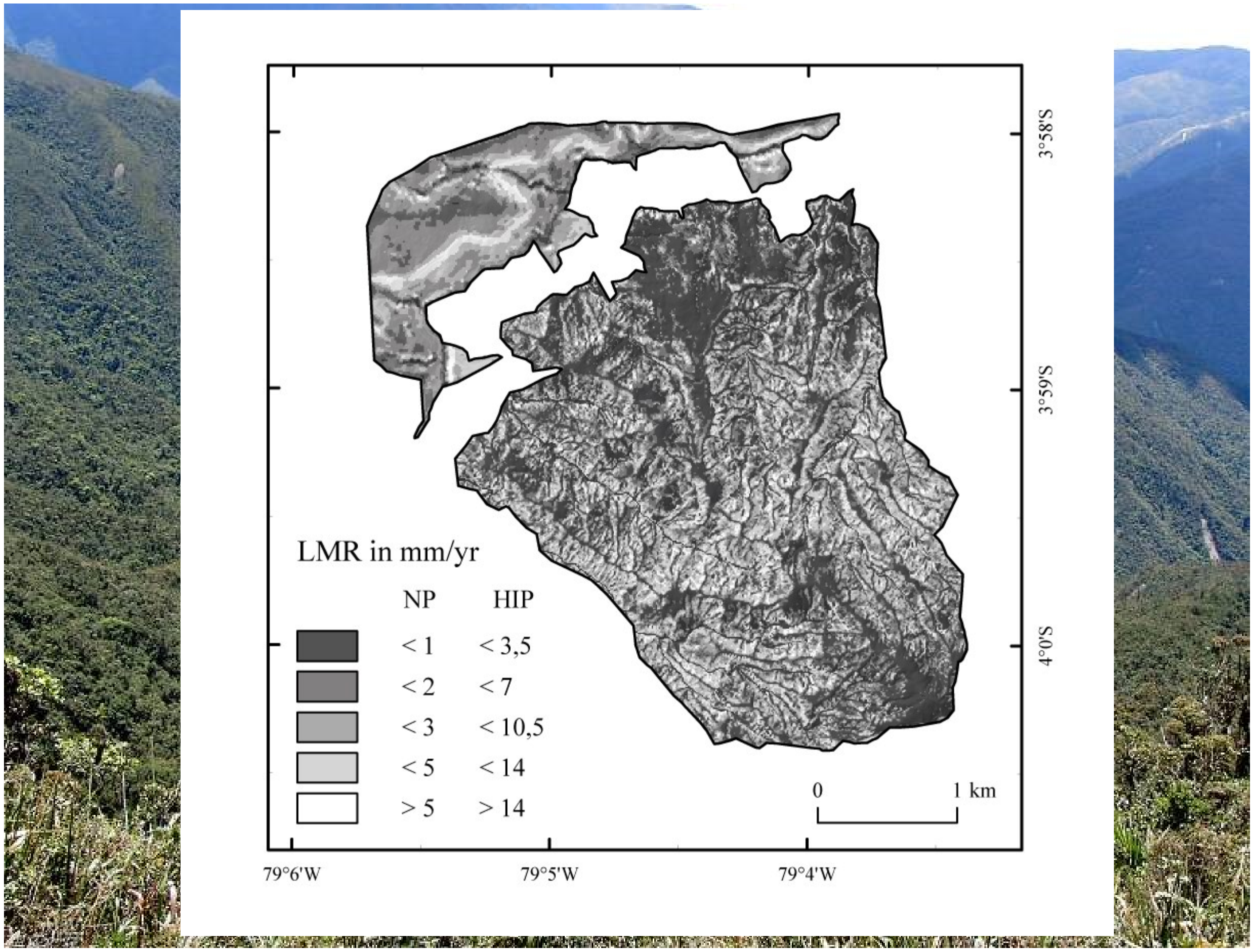


Foto: Michael Richter.

Source: Muenchow et al. (2011).

Literature

seit 1558

Crawley, M.J., 2006. Statistical computing. An introduction to data analysis using S-Plus. Wiley, Chichester.

Crawley, M.J., 2010. The R book. Wiley, Chichester.

Faraway, J.J., 2005. Linear models with R. Chapman & Hall/CRC, Boca Raton, Fla.

Faraway, J.J., 2006. Extending the linear model with R. Generalized linear, mixed effects and nonparametric regression models. Chapman & Hall/CRC, Boca Raton, Fla.

Hastie, T.J., Tibshirani, R.J., 1997. Generalized additive models. Chapman & Hall, London.

Pinheiro, J.C., Bates, D.M., 2009. Mixed-effects models in S and S-PLUS. Springer, New York.

Venables, W.N., Ripley, B.D., 2002. Modern applied statistics with S. Springer, New York, Berlin, Heidelberg.

Zuur, A.F., Ieno, E.N., Elphick, C.S., 2010. A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution 1, 3-14.

Zuur, A.F., Ieno, E.N., Smith, G.M., 2007. Analyzing ecological data. Springer, Berlin, New York.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., Smith, G.M., 2009. Mixed effects models and extensions in ecology with R. Springer, New York, NY.