Niche Modelling - Challenges

Chris Yesson

ZSL Institute of Zoology

LIVING CONSERVATION

Niche modelling - What can go wrong?

- Everything

- Data

- Analysis

- Validation

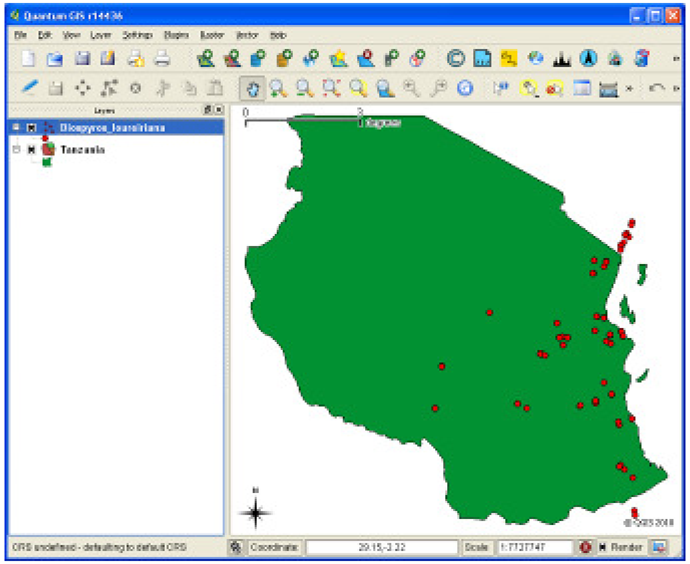
Niche modelling - What can go wrong?

- **Data**
  - can’t find enough
  - using too much
  - biased
  - incorrect
  - poor quality
- Analysis
- Validation

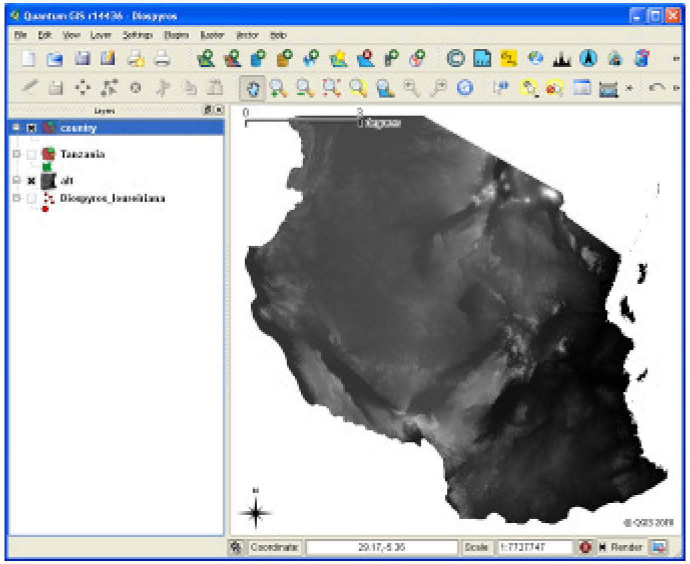
Common data sources

ZSL

- Species distribution
  - museums / herbaria (http://data.gbif.org/)
  - literature



- Environmental data
  - Global climate grids (http://www.worldclim.org/)
  - Global topographic grids (strm30)



- *More details this afternoon*

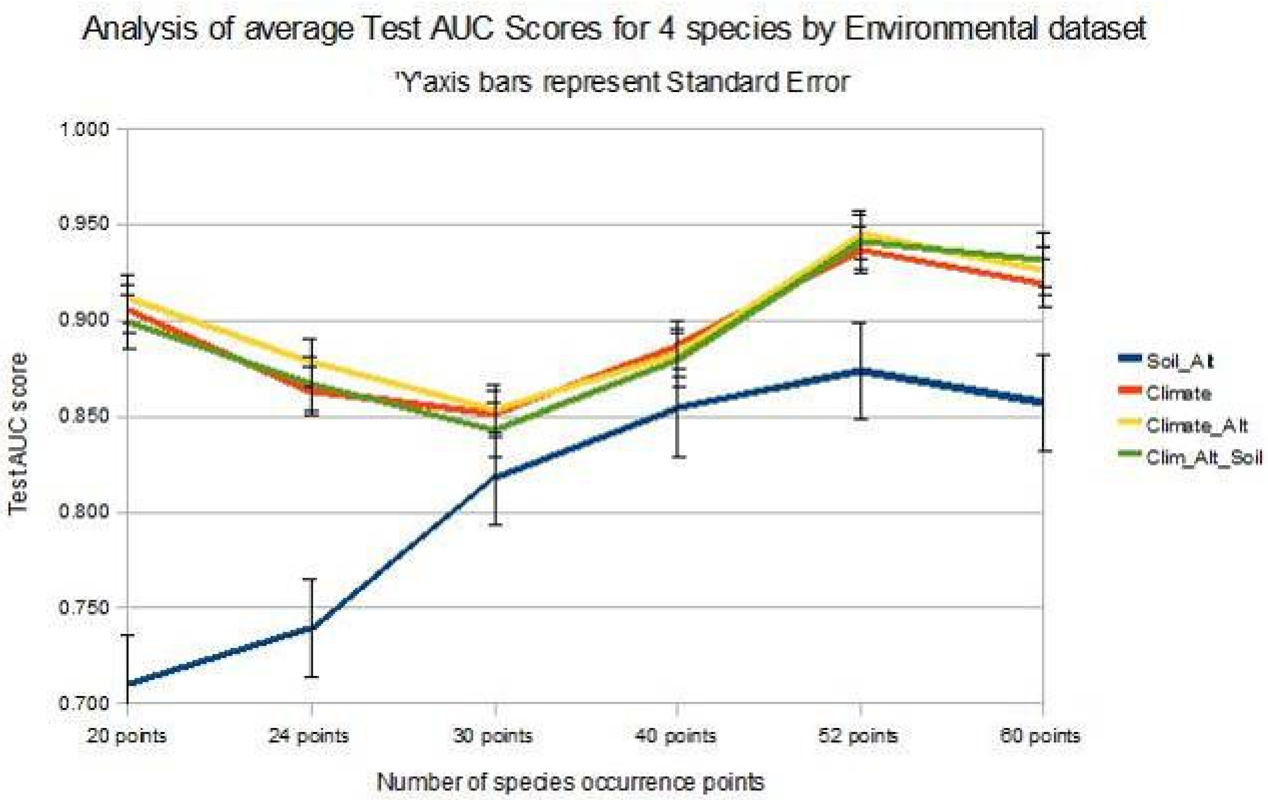
Distribution data – can't get enough?

ZSL

- Many Sources
  - –Your own collections
  - –Online museum/herbaria catalogues (GBIF)
  - –Other online catalogues (some data is not accessable via GBIF)
  - –Survey data
  - –Literature searches
  - –Specimen digitisation

Distribution data – How many points do I need?

- Some say
  - 5
  - 10
  - 20
  - 50
- But it depends on your data
- You will need solid justification for modelling with only tens of points

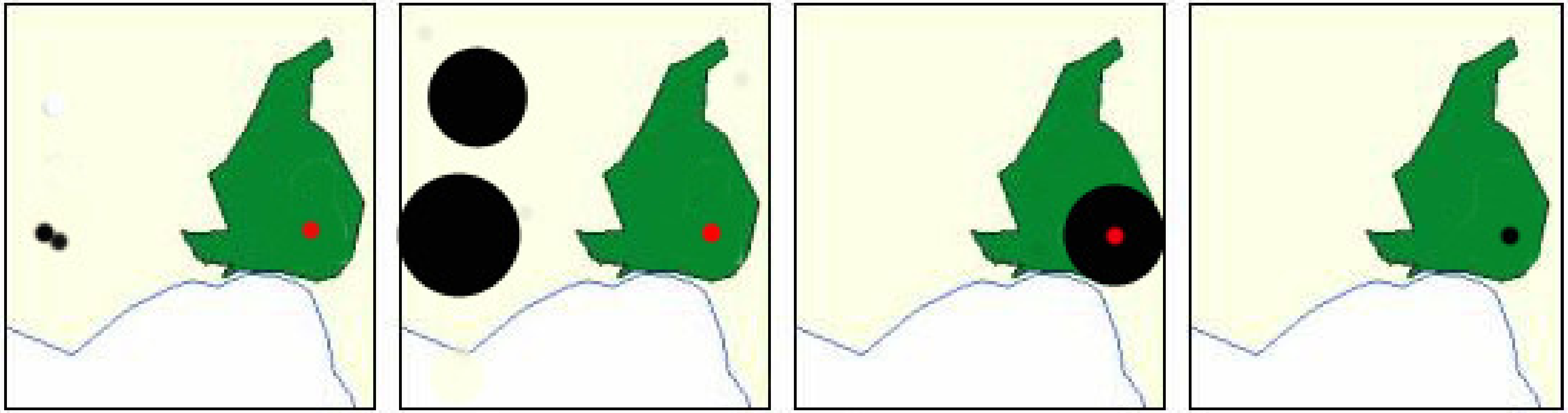


Problems with distribution data

- Sampling biases
  - regional bias
    - *Look at all that UK data, its the global centre of all diversity!*
  - environmental bias
    - *What a lot of plants grow by road-sides*
  - taxonomic bias
    - *The world is full of cute furry animals*

- Accuracy
  - geographic errors
    - *GPS or georectification?*
  - taxonomic errors
    - *Who identified that specimen?*

Accuracy vs Precision



The differences between accuracy and precision in a spatial context.

The red spots show the true location, the black spots, represent the locations as reported by a collector.

a. High precision, low accuracy.

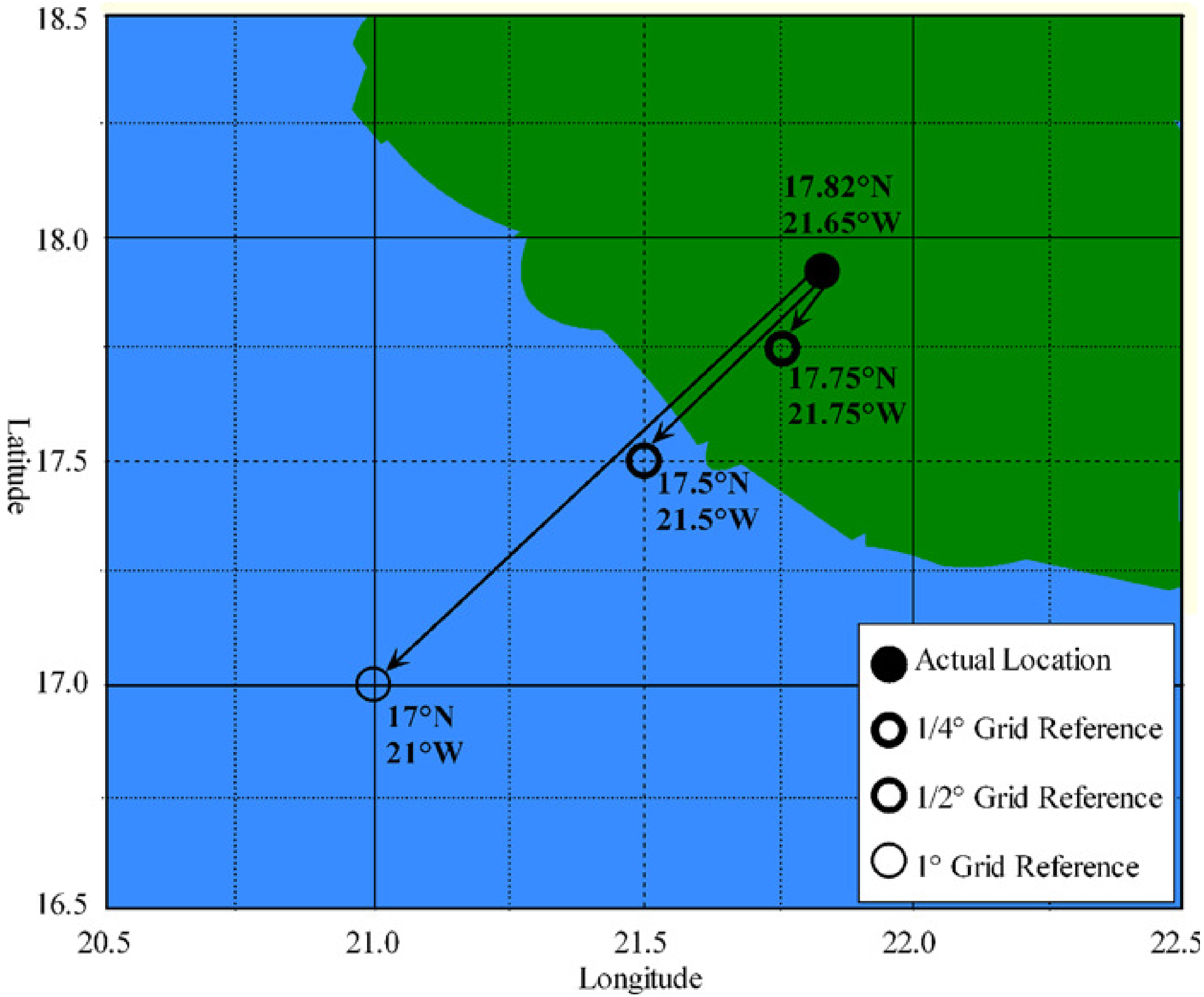
b. Low precision, low accuracy showing random error.

c. Low precision, high accuracy.

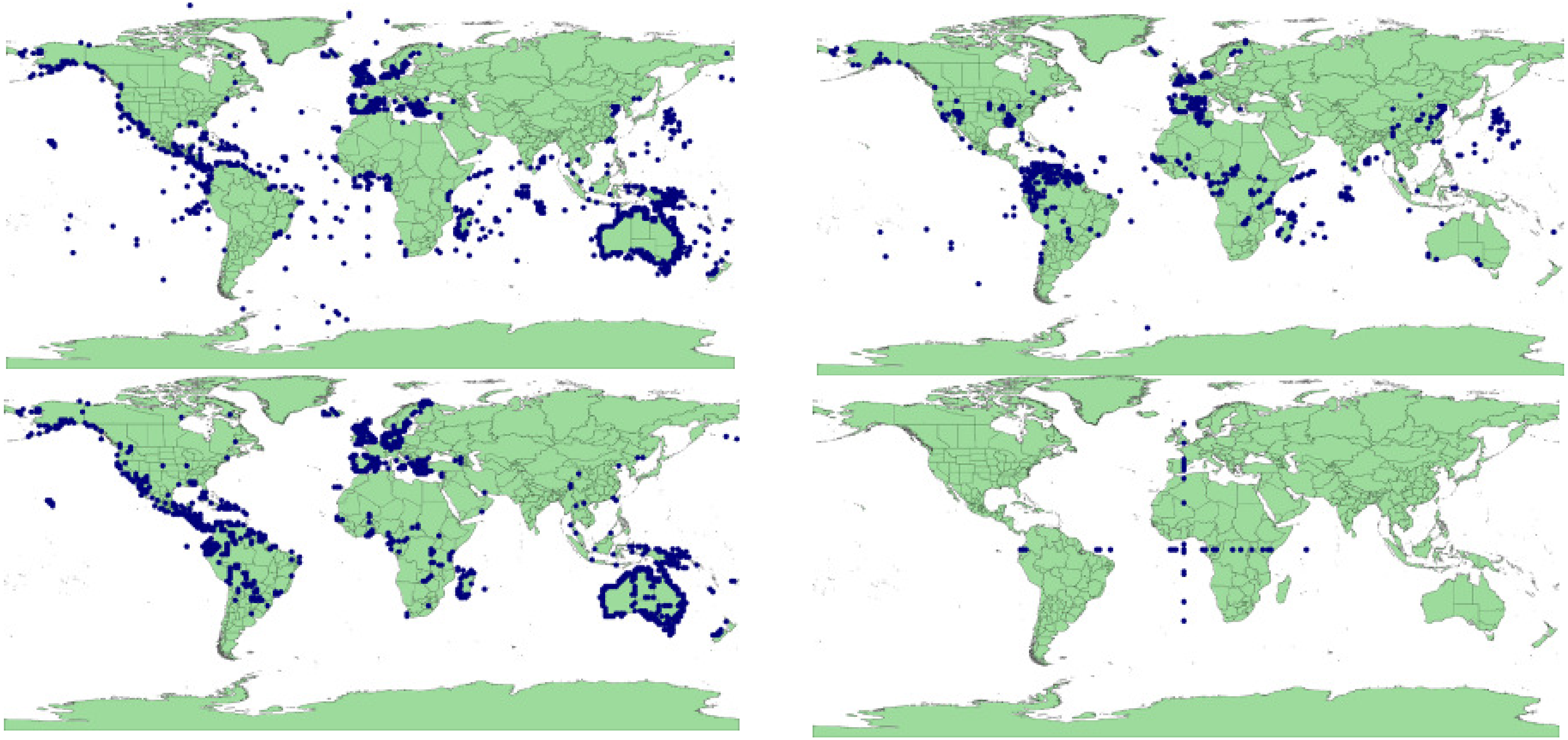
d. High precision and high accuracy

Chapman 2005 Principles of Data Quality available to download on www.gbif.org

Good data that appear bad



Basic errors

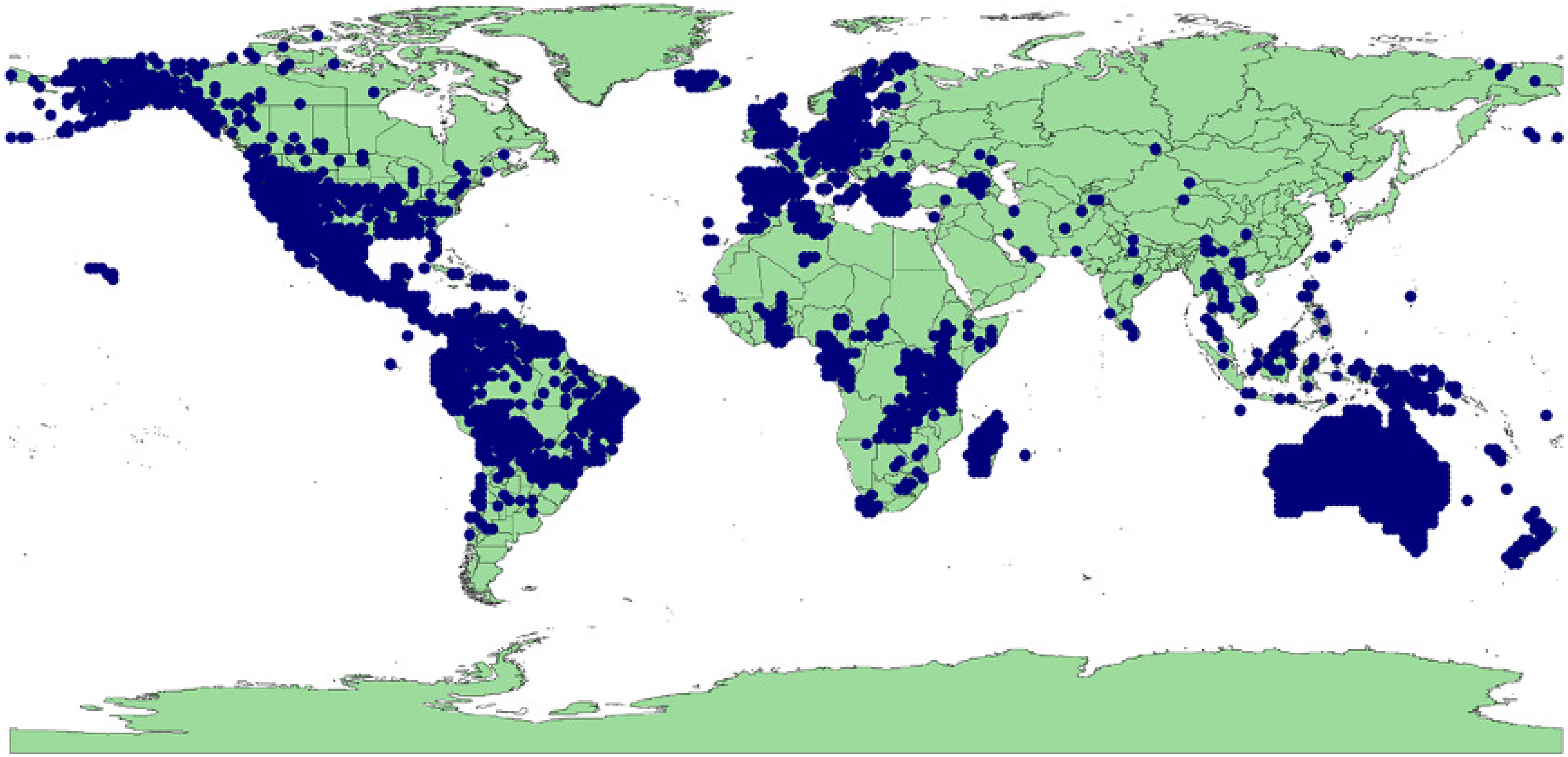


In the sea | Lat/Long reversals

Near Valid | Lat/Long zero

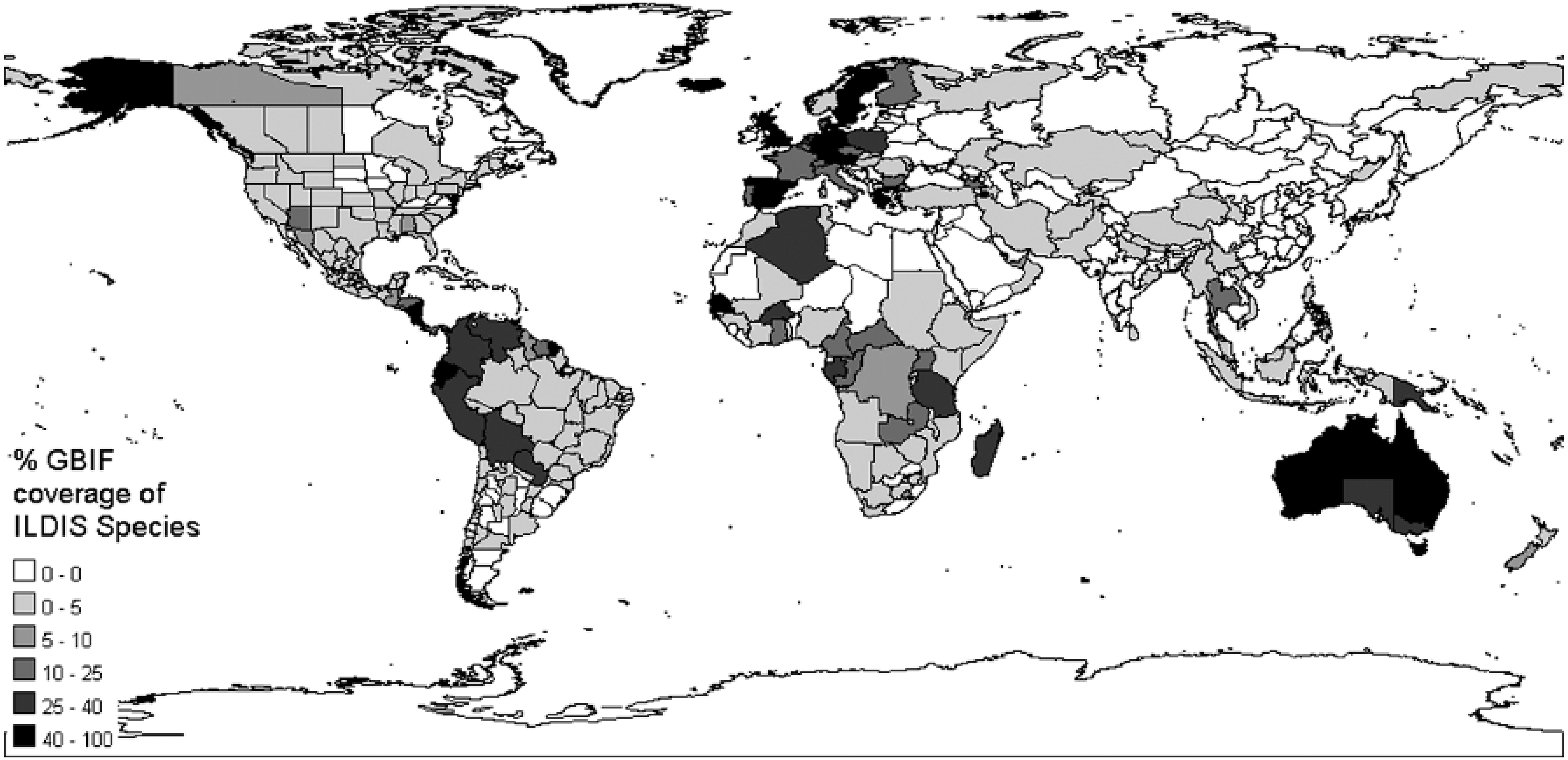
Yesson et al., (2007) How Global Is the GBIF?. *PLoS ONE.* **2:** e1124

Geographic Coverage



- Fabaceae data from GBIF showing patchy geographic coverage

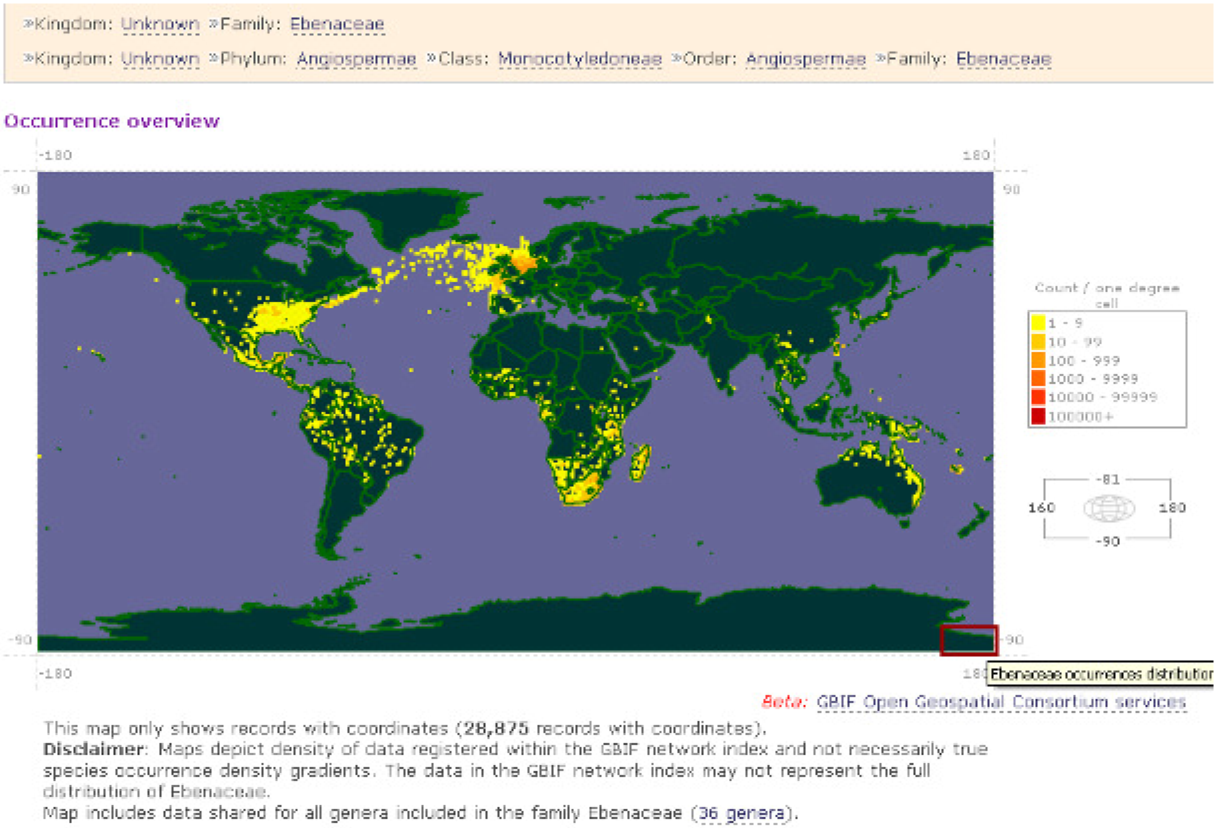
Taxon Coverage



- Global Legume coverage from GBIF data per TDWG level 4 area

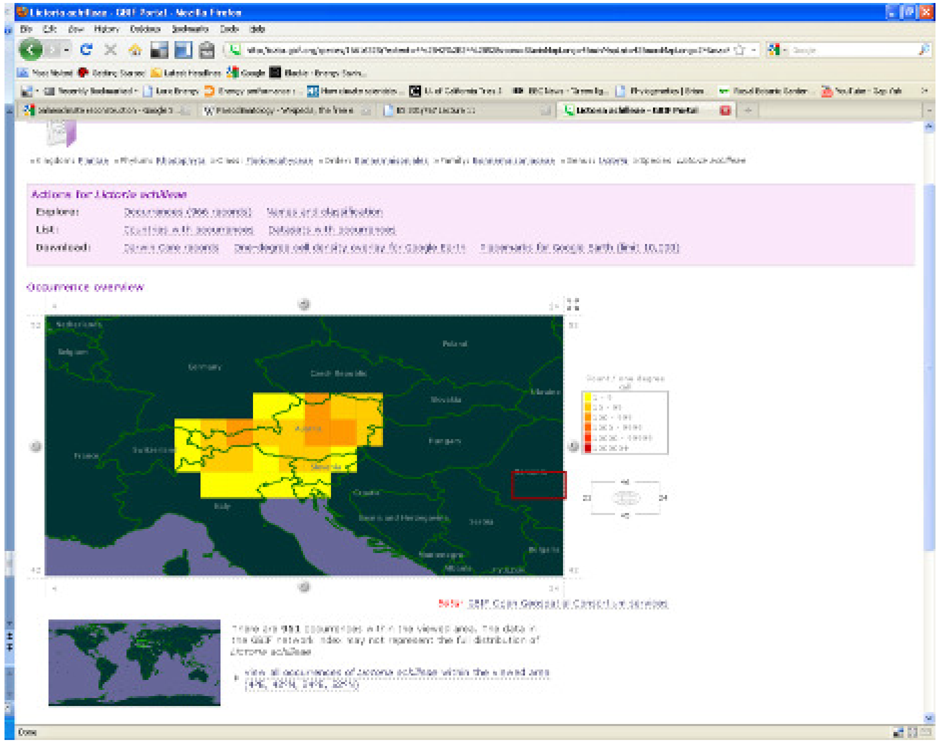
Other issues

- Taxonomy and checklists
- Misclassified data
  - –Synonymy
  - –Homonymy
  - –Misidentification



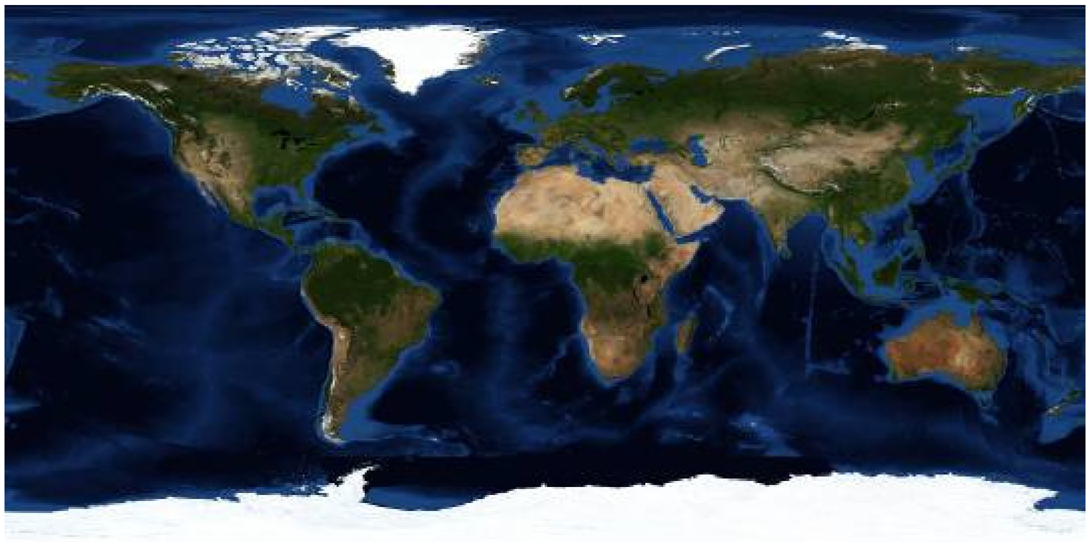
More taxonomy

- Correctly determined taxa
- Wrongly databased
- *Lictoria achillae*
  - *GBIF – listed as Rhodophyta*
  - *Source database – listed as Lepidoptera!*



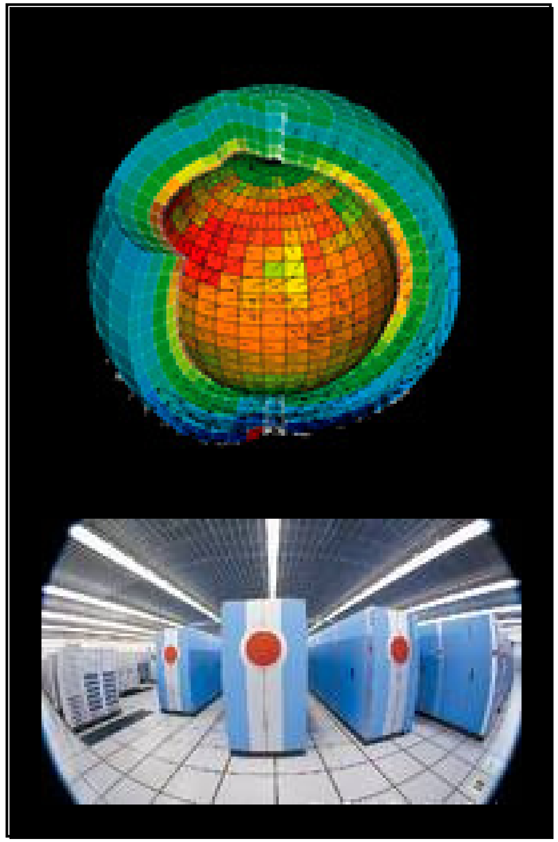
Environmental data

- Climate
- Topographic
- Classified data
- Marine



Climate data

- Modelling climate on a global scale is not trivial
- 35.86 TFLOPS
- 6.4 Mw
- 10km grid
- Resolution is improving
- Models are improving
- Computers are improving
- Global 30 arc second grids are available (http://www.worldclim.org/)



Earth Simulator (Japan)

Climate model data

- The Intergovernmental Panel on Climate Change
- IPCC1 - 1990
- IPCC2 - 1995
- IPCC3 - 2001
- IPCC4 - 2007

- IPCC provides consensus on what scientists expect to happen
- IPCC5 is on the way

ClimateGate

'ClimateGate' means - Mozilla Firefox

http://news.bbc.co.uk/1/hi/sci/tech/8388485.stm

Mobile Version | Help

BBC Home News Sport Weather iPlayer TV Radio More... Search

NEWS LIVE BBC NEWS CHANNEL

News Front Page
World
UK
England
Northern Ireland
Scotland
Wales
Business
Politics
Health
Education
Science & Environment
Technology
Entertainment
Also in the news

Video and Audio

Have Your Say
Magazine
In Pictures
Country Profiles
Special Reports

Related BBC sites
Sport
Weather
Democracy Live
Radio 1 Newsbeat
CBBC Newsround
On This Day
Editors' Blog

Page last updated at 14:56 GMT, Tuesday, 1 December 2009

E-mail this to a friend   Printable version

'Show Your Working': What 'ClimateGate' means

VIEWPOINT
Mike Hulme and Jerome Ravetz

The "ClimateGate" affair - the publication of e-mails and documents hacked or leaked from one of the world's leading climate research institutions - is being intensely debated on the web. But what does it imply for climate science? Here, Mike Hulme and Jerome Ravetz say it shows that we need a more concerted effort to explain and engage the public in understanding the processes and practices of science and scientists.

As the repercussions of ClimateGate reverberate around the virtual community of global citizens, we believe it is both important and urgent to reflect on what this moment is telling us about the practice of science in the 21st Century.

In particular, what is it telling us about the social status and perceived authority of scientific claims about climate change?

We argue that the evolving practice of science in the contemporary world must be different from the classic view of disinterested - almost robotic - humans establishing objective claims to universal truth.

Climate change policies are claimed to be grounded in scientific

"Practising scientists know that they do not simply follow a rulebook to do their science, otherwise it could be done by a robot"

THE GREEN ROOM

A weekly series of thought-provoking opinion pieces on environmental topics

Decision time
It is time for nations to end two decades of deadlock on whale conservation

- Your comments

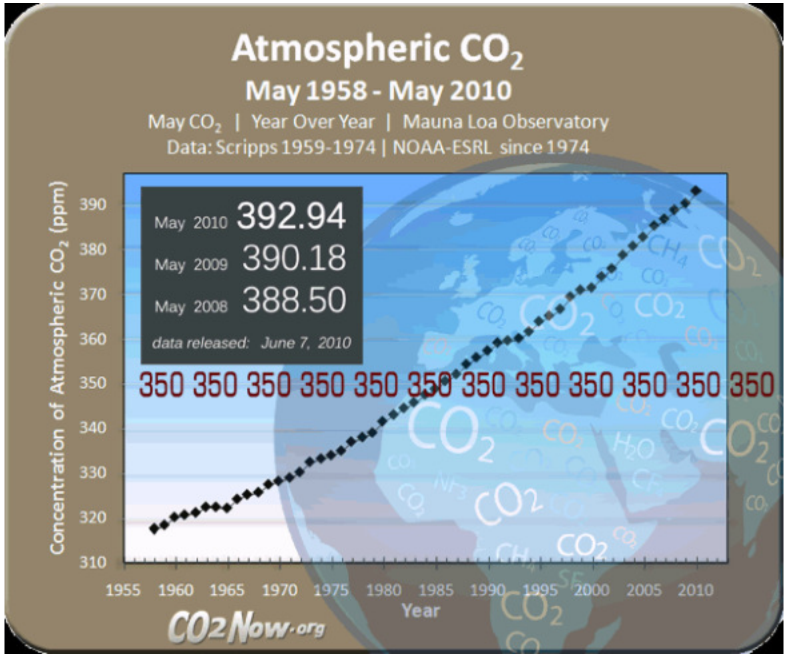
RECENT ARTICLES
- Hooking the 'pirates'
- Whales - 'resource' or 'right'?
- Oil: Choices we need to make
- A new direction for climate policy
- Seeing REDD over forest peoples
- Reviving the spirit of Rio
- EU must increase climate ambition
- Counting the cost of alien invasions
- Has CITES had its day?
- Curious case of Kiwi hedgehogs
- Restoring natural capital

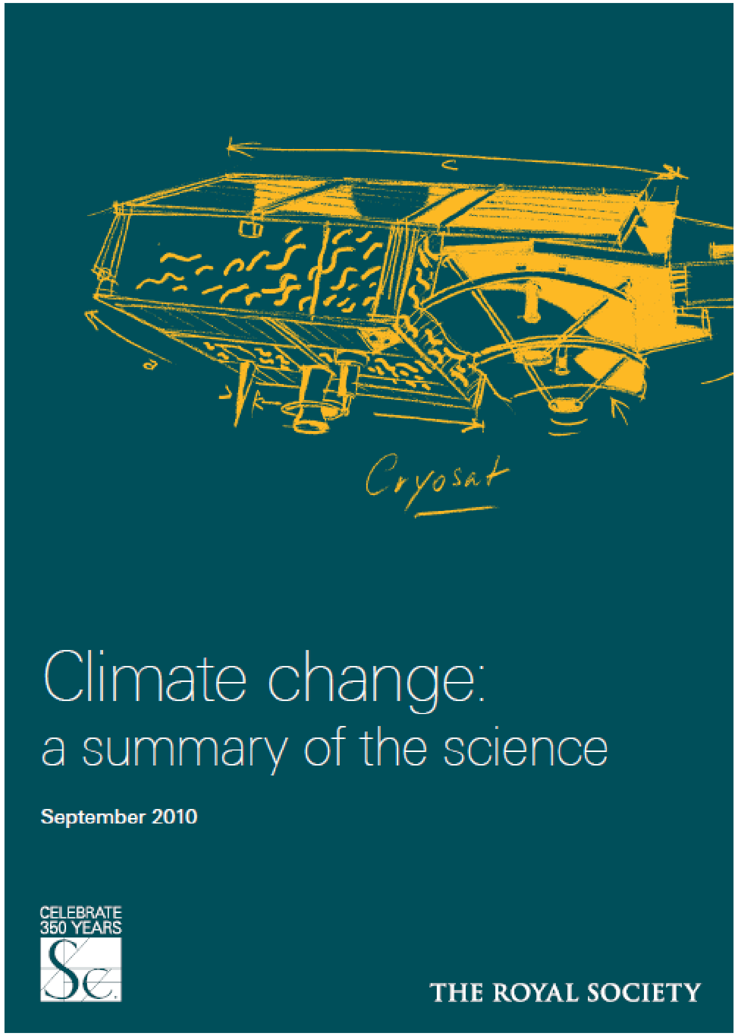
LINKS
- Copenhagen summit
- Richard Black's Earth Watch
- Earth News

Can we rely on future climate models?

ZSL

- They are models, not predictions
- Sound basis in science
  - see royal society summary 2010
- Real observations – CO_2 Now





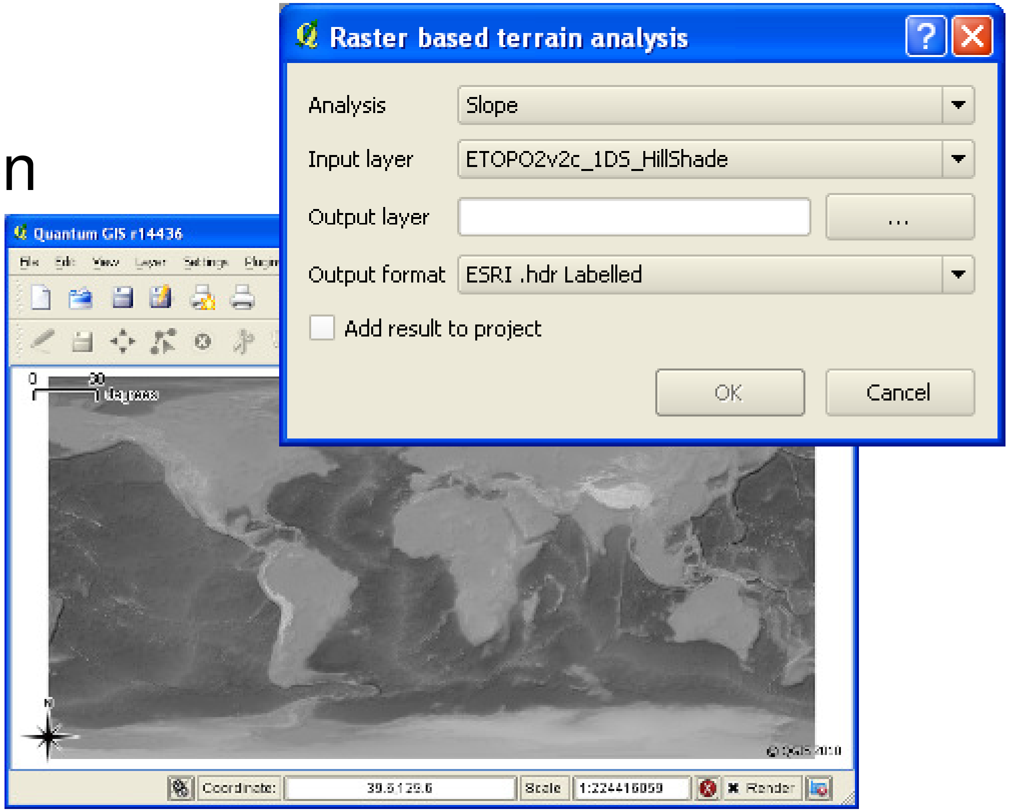
Climate data continued

- Present
  - Good quality, high resolution, includes direct observations
- Future
  - 50-100 years in the future = lots of uncertainty, low resolution, no direct observations
- Past
  - thousands-millions of years in the past = very uncertain, low resolution, some indirect observations

- *Climate datasets are made by climatologists for climatologists*

Topographic data

- Elevation/bathymetry
  – derived from satellites & local surveys
  – high resolution (1km global grids)
- Data derived from elevation
  – aspect
  – slope
  – ‘roughness’
- *Beware the correlation with other (climatic) data*



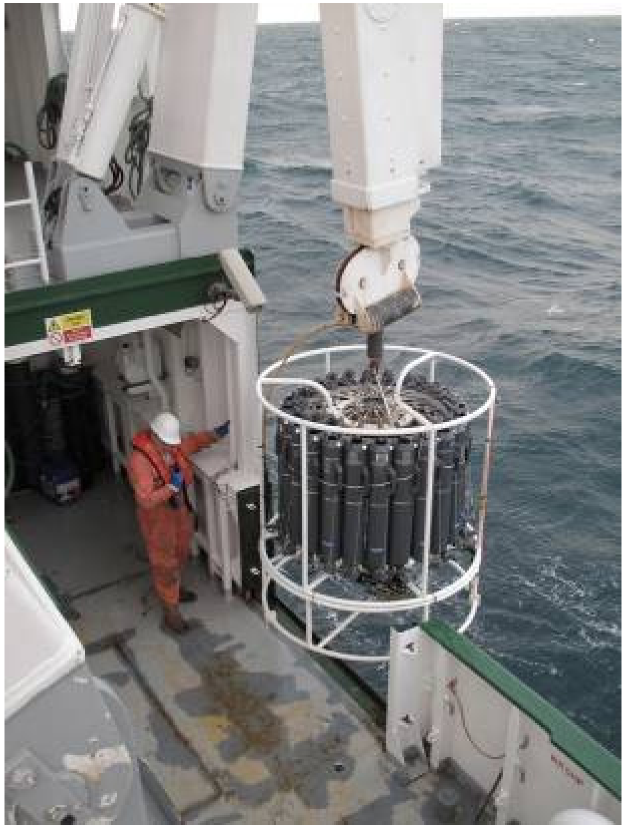
Biome & Land use classifications

- Infrastructure
  - Roads
  - Cities/towns/villages

- Land use classifications (farmland, etc)
- Soil data
- Biomes (UN data)
- *Classified data can be difficult to incorporate into models*

Marine

- Bathymetry data
  - satellite & ship soundings
  - 1km grid (really?)
  - derived data (slope aspect etc)
- Ocean chemistry
  - pH, salinity, carbon
- Primary productivity
- Currents
- Temperature



- *Lots of publicly accessible, global data sets, but generally low resolution*

Niche modelling - What can go wrong?

- Data

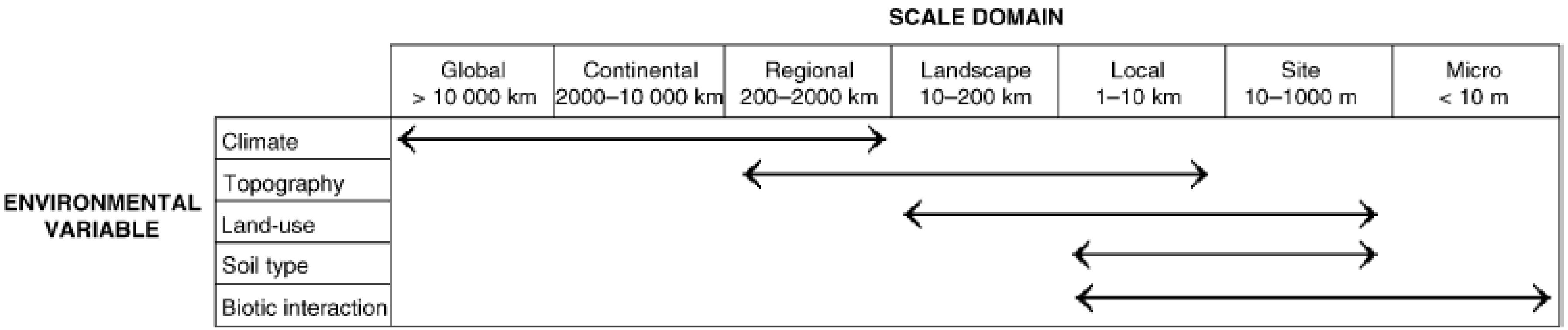
- **Analysis**
  - Overfitting
  - Scale
  - Algorithm selection

- Validation

Choose your environmental layers carefully

- Too many layers leads to models overfitting

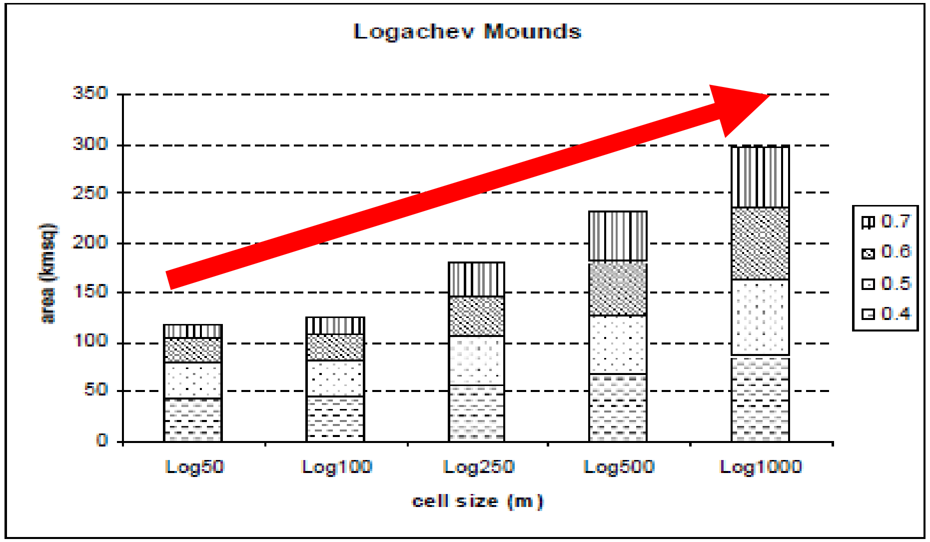
- Scale is important

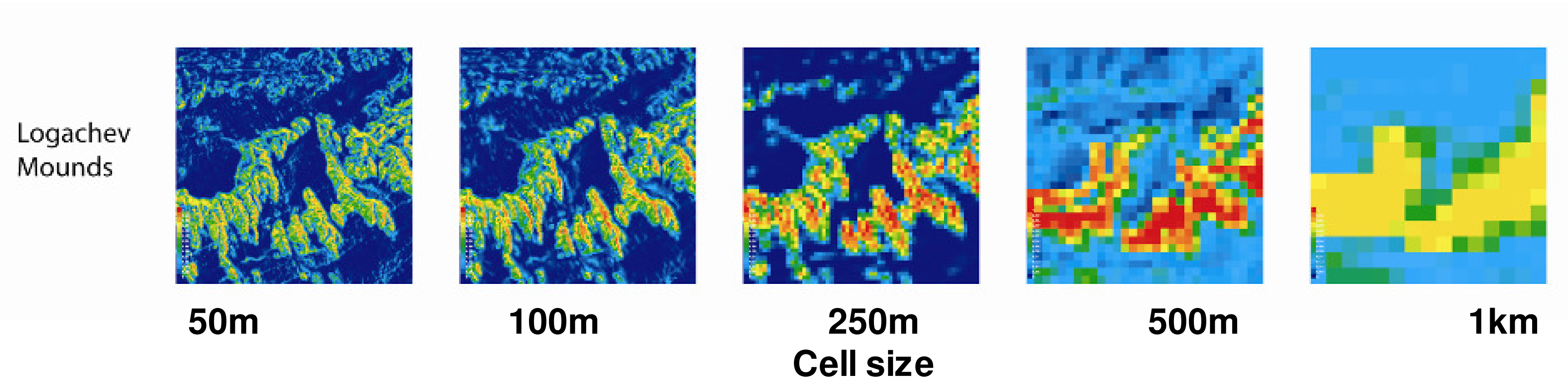


Pearson & Dawson (2003) Global Ecology & Biogeography **12:** 361–371

Pixel size experiments

- Local niche model for *L. pertusa*
- Expectation that models will overpredict at coarser resolutions





Niche Modelling algorithms

Table 4. Modelling methods implemented.

Method	Class of model, and explanation	Data[1]	Software	Std errors?[2]	Contact person
BIOCLIM	envelope model	p	DIVA-GIS	no	CG, RH
BRT	boosted decision trees	pa	R, gbm package	no	JE
BRUTO	regression, a fast implementation of a gam	pa	R and Splus, mda package	yes	JE
DK-GARP	rule sets from genetic algorithms; desktop version	pa	DesktopGarp	no	ATP
DOMAIN	multivariate distance	p	DIVA-GIS	no	CG, RH
GAM	regression: generalised additive model	pa	S-Plus, GRASP add-on	yes	AG,AL,JE
GDM	generalised dissimilarity modelling; uses community data	pacomm	Specialized program not general released; uses Arcview and Splus	no	SF
GDM-SS	generalised dissimilarity modelling; implementation for single species	pa	as for GDM	no	SF
GLM	regression; generalised linear model	pa	S-Plus, GRASP add-on	yes	AG,AL,JE
LIVES	multivariate distance	p	Specialized program not general released	no	JLi
MARS	regression; multivariate adaptive regression splines	pa	R, mda package plus new code to handle binomial responses	yes	JE, FH
MARS-COMM	as for MARS, but implemented with community data	pacomm	as for MARS	yes	JE
MARS-INT	as or MARS; interactions allowed	pa	as for MARS	yes	JE
MAXENT	maximum entropy	pa	Maxent	no	SP
MAXENT-T	maximum entropy with threshold features	pa	Maxent	no	SP
OM-GARP	rule sets derived with genetic algorithms; open modeller version	pa	new version of GARP not yet available	no	ATP

Elith J, Graham CH, Anderson RP*, et al.* (2006)
Novel methods improve prediction of species' distributions from occurrence data.
Ecography **29**, 129-151.

Algorithms in brief

- Simple – Poor performance - Understandable

- Complex – high performance – Black box

Bioclim

Environmental Distance

CSM

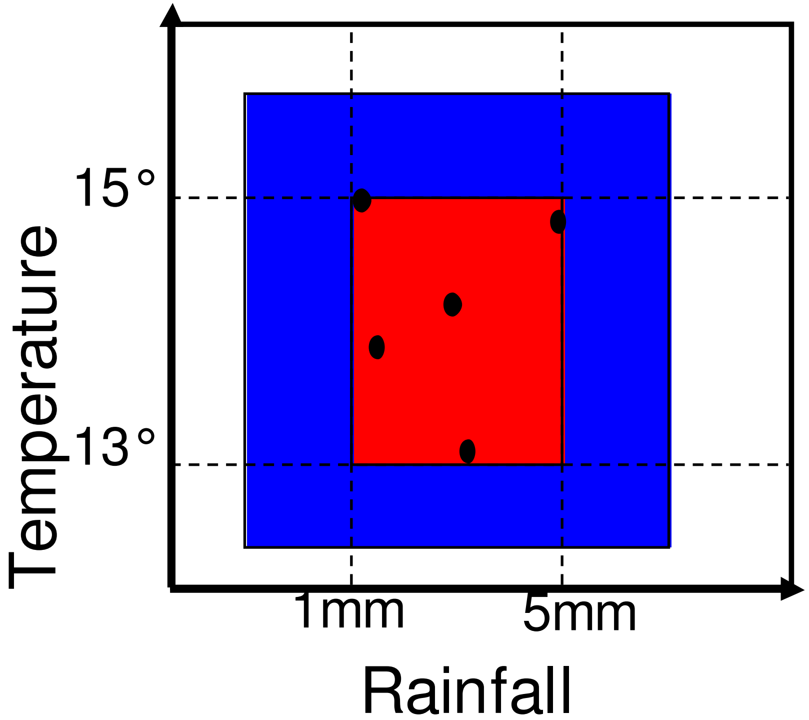
ENFA

GARP

Maxent

The Bioclim method (Nix 1986)

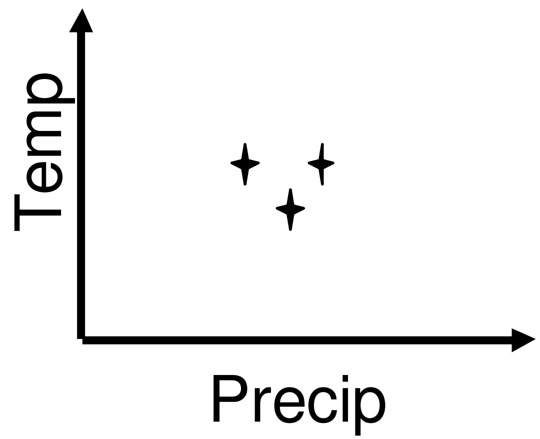
- Find Min, Max, Mean, Standard Deviation
- Core area = within 1 std.dev of mean
- Marginal = within observed range
- Overlap for multiple layers
- Methodology (Busby, 1991) has 300+ citations
- Performs poorly in comparisons with other methods

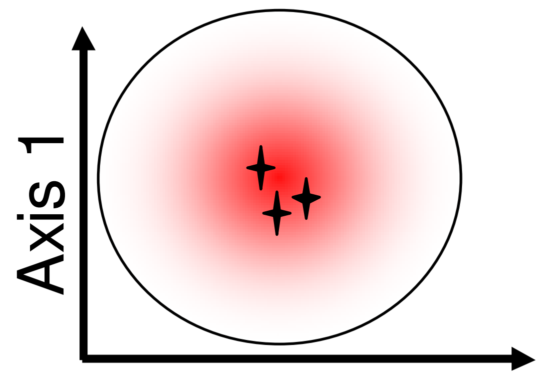


Nix, H. A. (1986) *A Biogeographic Analysis of Australian Elapid Snakes*. Australian Flora and Fauna Series Number 7: Atlas of Elapid snakes of Australia (ed. by R. Longmore), pp. 4-15. Australian Government Publishing Service, Canberra.

Distance methods

- Rescale and re-orient axes
- Account for correlation
- Some are modified PCAs
- CSM, ENFA, Environmental distance in openModeller



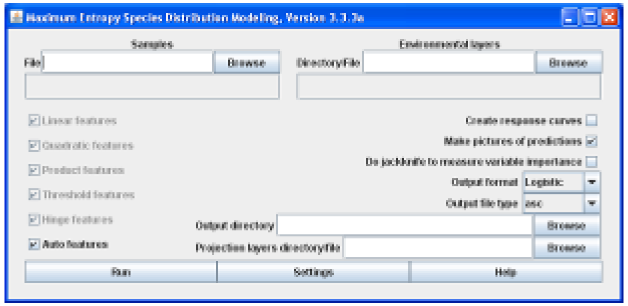


GARP

- **GARP** – *genetic algorithm* based calculation gives probability density model based on frequency of model iterations
- Originally software in the desktop garp software
- New improved version available in openModeller
- Original methodology (Stockwell & Peters 1999) has 500+ citations
- Criticised by many as a black box

Maximum entropy

- *maximum entropy algorithm* uses machine learning to best identify the link between actual distribution pints and a set of given variables
- Very popular method (700+ citations for Phillips et al, 2006)
- Criticised for overfitting
- Dedicated Maxent software maintained and updated by the original developers



- … or openModeller implementation

Phillips, S. J.; Anderson, R. P. & Schapire, R. E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling.* **190:** 231-259

Niche modelling - What can go wrong?

- Data
- Analysis
- Validation
  - Kappa and ROC/AUC
  - Absence data

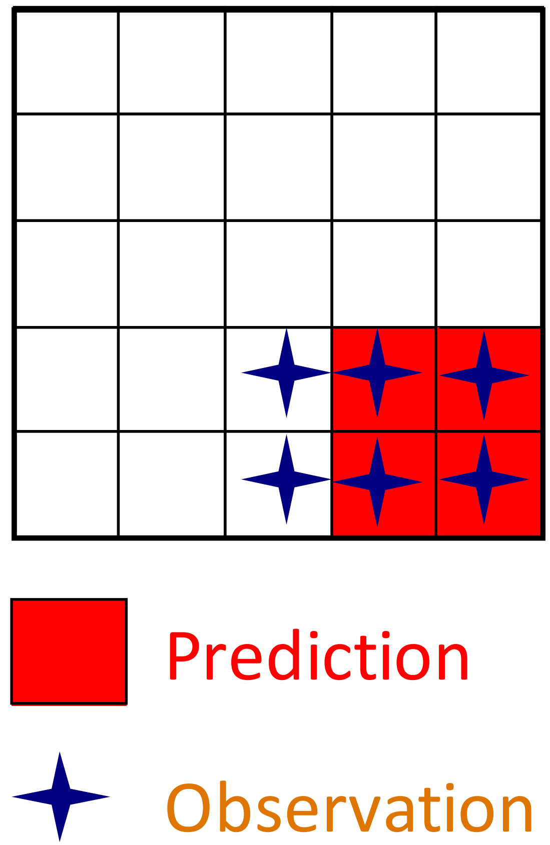
How good is a niche model?

- Niche models are based on observations.
- Models are evaluated empirically.
- Two popular approaches:
  - Kappa
    - score for presence/absence models
  - Area Under the Curve (AUC)
    - Score for multi-value models

Prediction errors - Underprediction

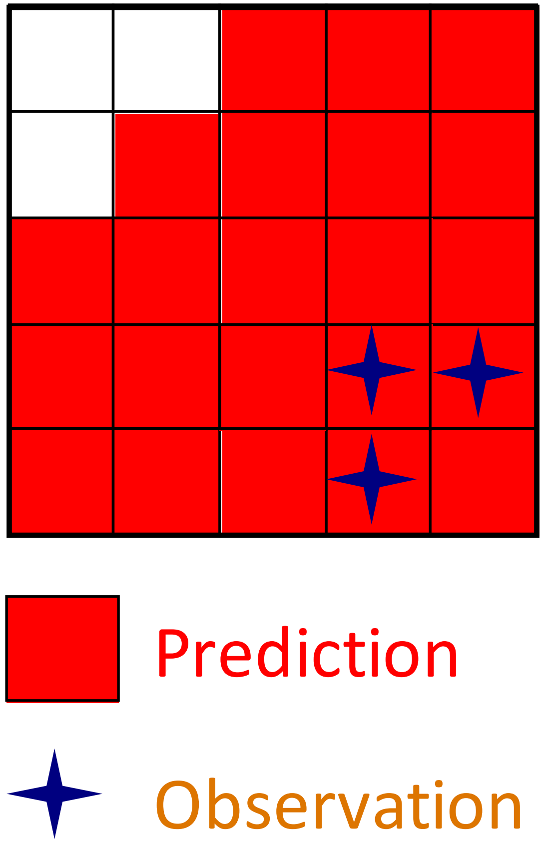
ZSL

- Underprediction leads to real distribution points which are outside the predicted area
- Such points are termed false negatives
- These can be avoided by widening the area of prediction



Prediction errors - Overprediction

- Overprediction is the selection of areas which do not contain your species
- Such areas are termed false positives



Confusion Matrix

		Observation + present	 - absent
Prediction	+ present	+ Observation + Prediction	- Observation + Prediction *Overprediction*
	- absent	+ Observation - Prediction *Underprediction*	- Observation - Prediction

Kappa (κ) – A single value for model accuracy

ZSL

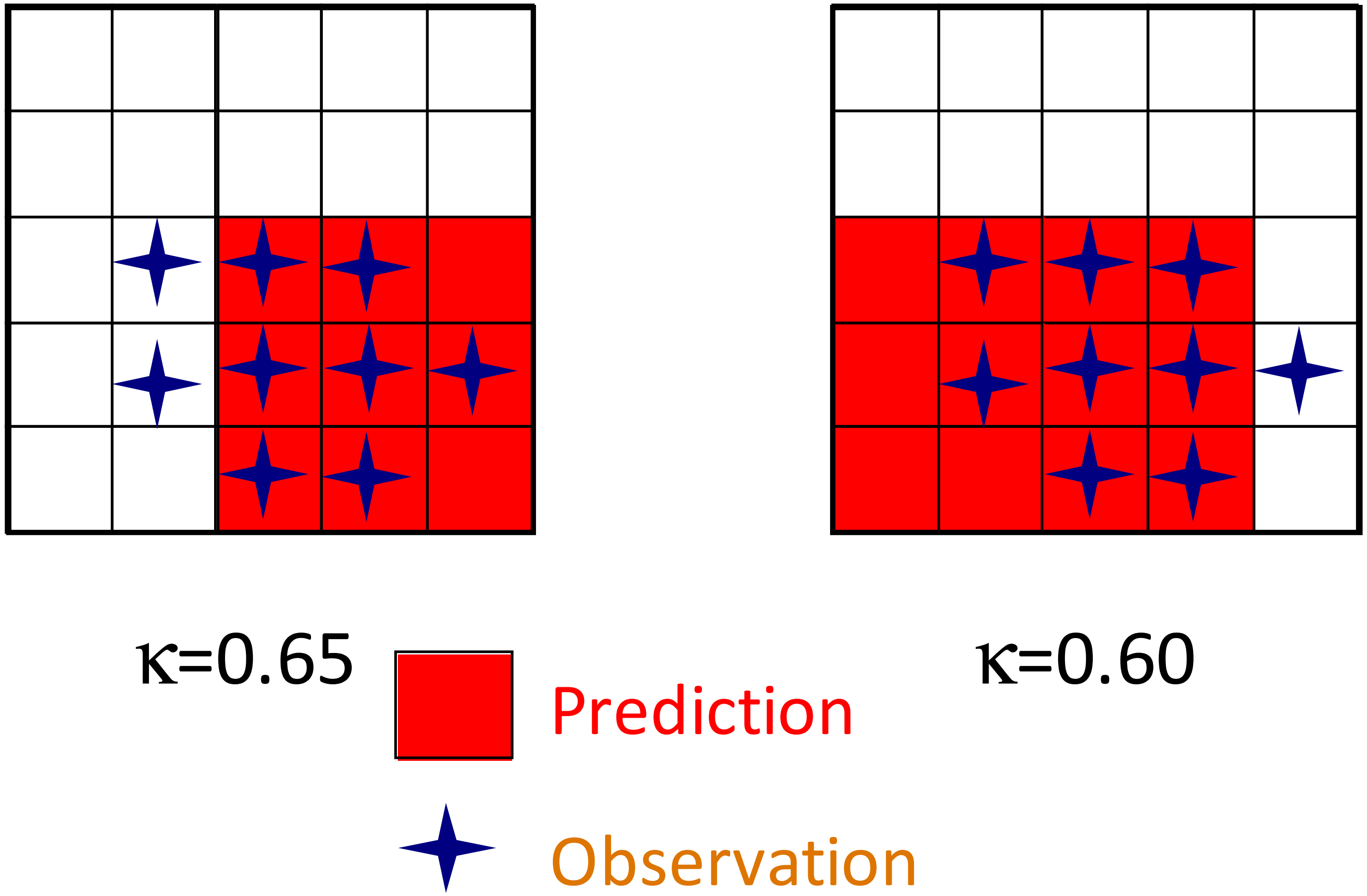
		Observation +	Observation −	Total
Prediction	+	A	B	A+B
Prediction	-	C	D	C+D
Total		A+C	B+D	N=A+B+ C+D

$$\kappa = \frac{(A+D) - E/N}{N - E/N}$$

where

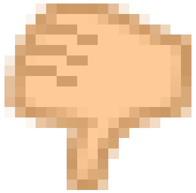
$$E = (A+C)(A+B) + (B+D)(C+D)$$

Kappa (κ) - Examples



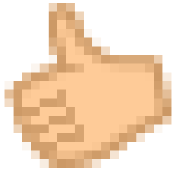
Kappa (κ) – rule of thumb

- $\kappa \leq 0.4$ is *poor*



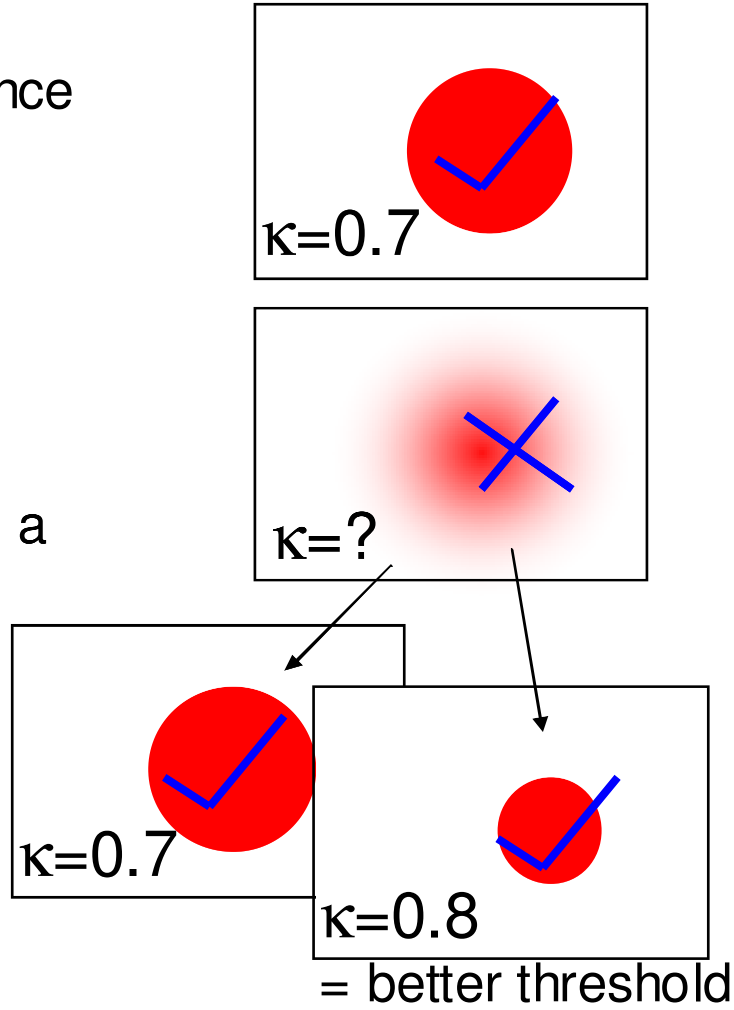
- $0.4 < \kappa < 0.75$ is *good*

- $\kappa \geq 0.75$ is *excellent*



Validation for multi-value models

- Kappa assumes that models predict presence p=1 or absence p=0
- Generally models are scored to provide probability of presence $0 \leq p \leq 1$
- Calculating kappa requires the selection of a threshold
  - $p \leq$ Threshold implies absence
  - P > Threshold implies presence
- The threshold can be chosen to maximise kappa



Sensitivity & Specificity

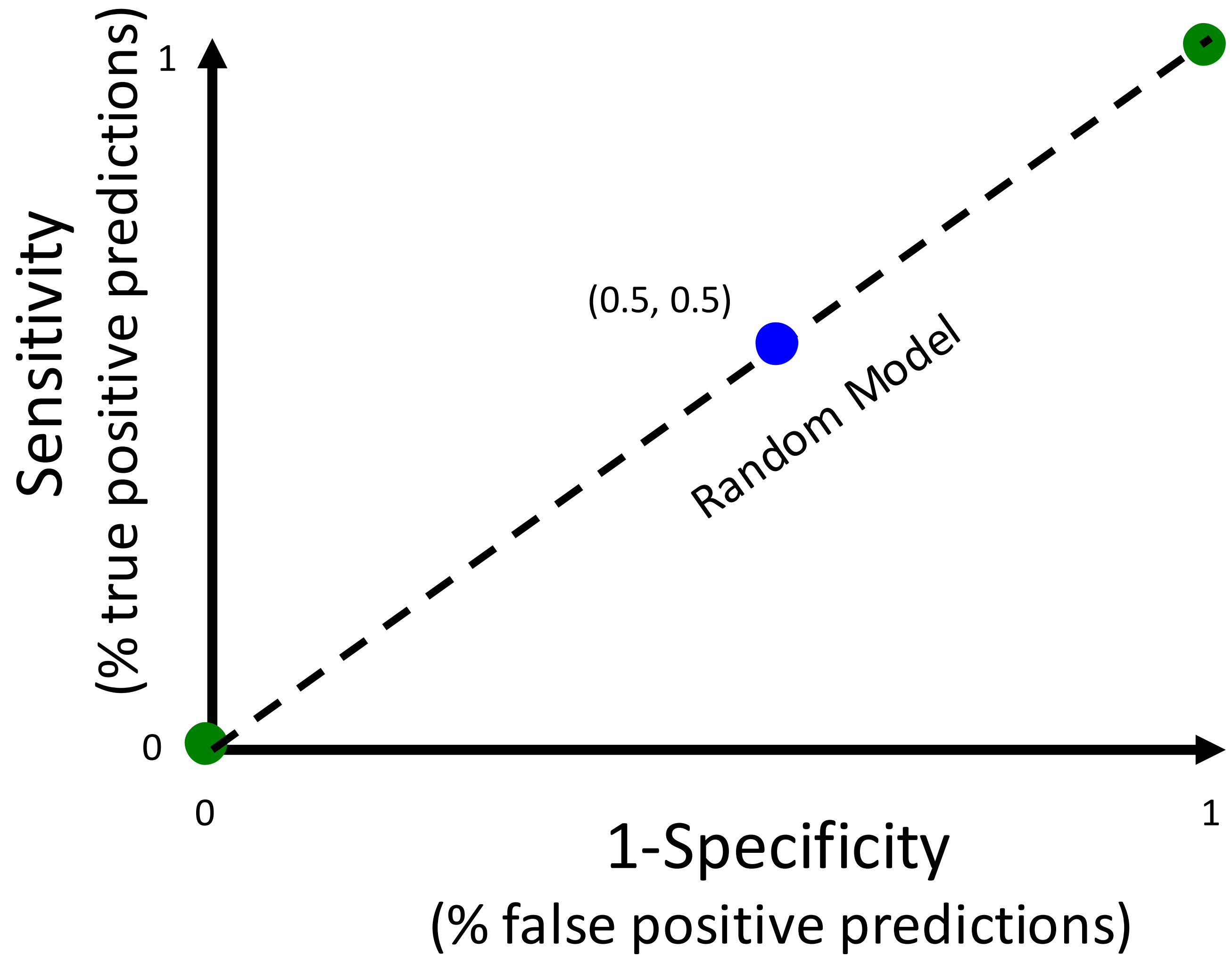
- Sensitivity
  - Proportion of true positive predictions
  - Sensitivity = A / (A+C)

- Specificity
  - Proportion of false positive predictions
  - Specificity = B / (B+D)

- Plotting these values is informative

		Observation +	Observation −
Prediction	+	A	B
	-	C	D

Receiver Operating Characteristic - ROC plots



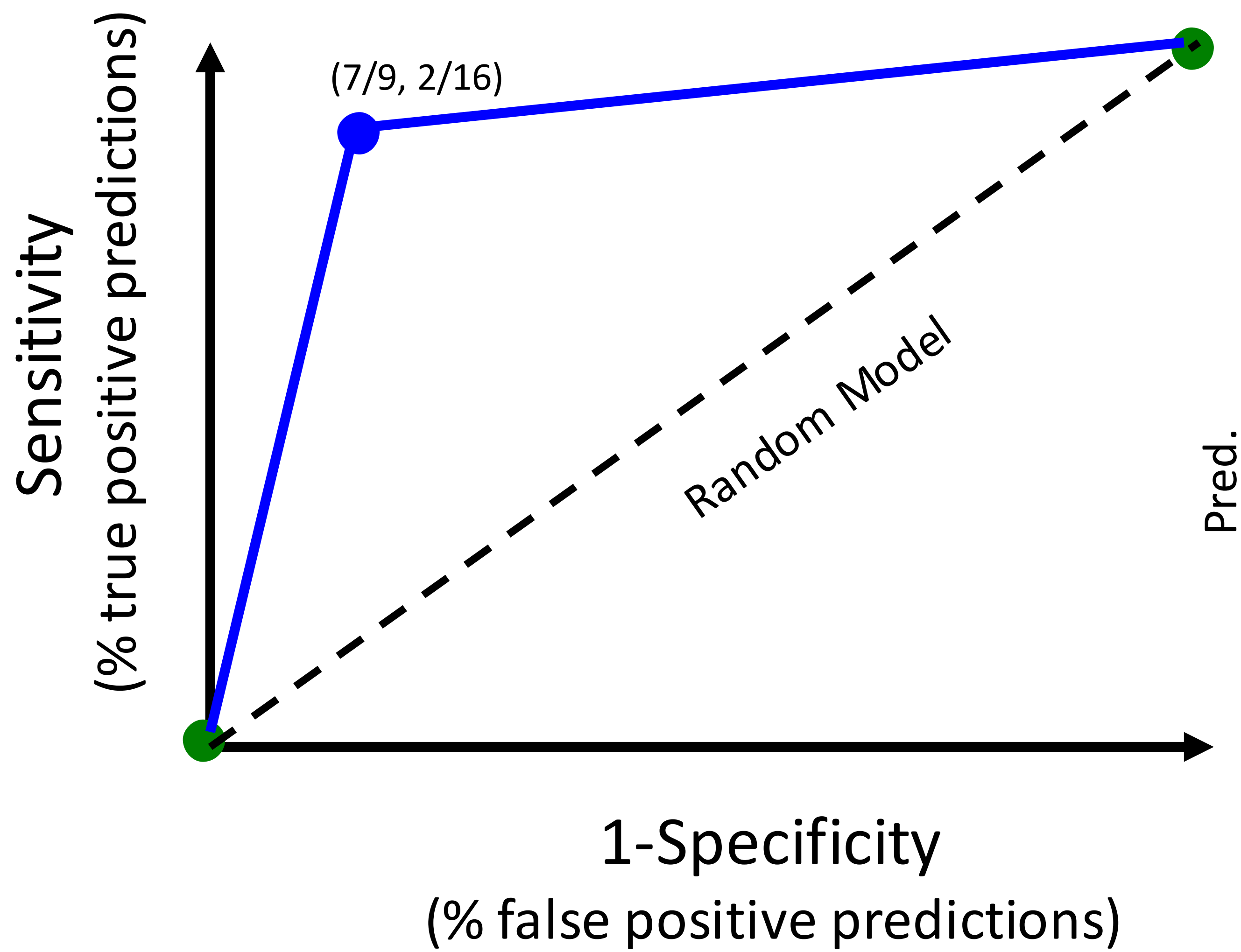
Confusion Matrix

Obs.

Pred.	+	-
+	A =5	B =5
-	C =5	D=5

Sensitivity = 5/10
Specificity = 5/10

ROC plots

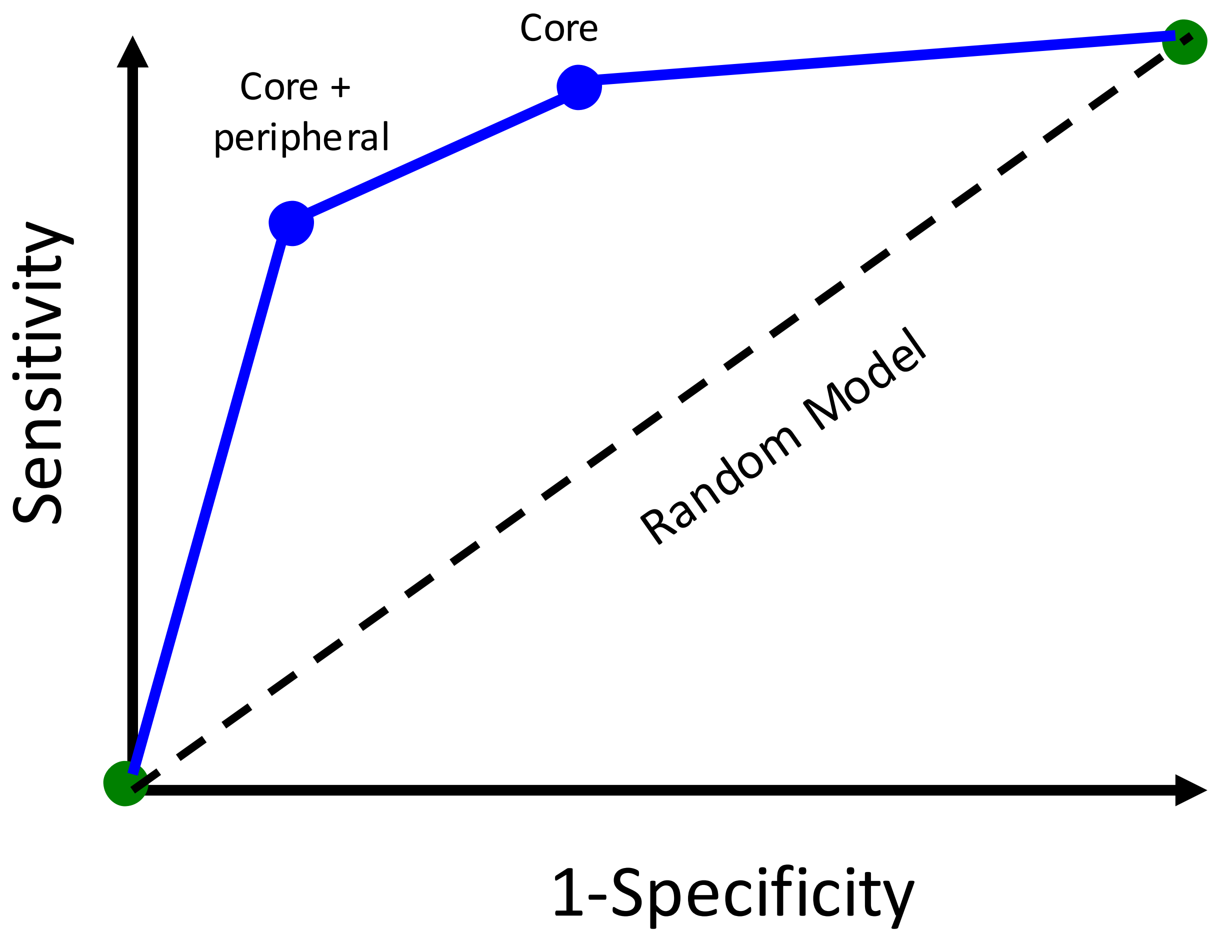


Confusion Matrix

		Obs. +	Obs. -
Pred.	+	A =7	B =2
	-	C =2	D =14

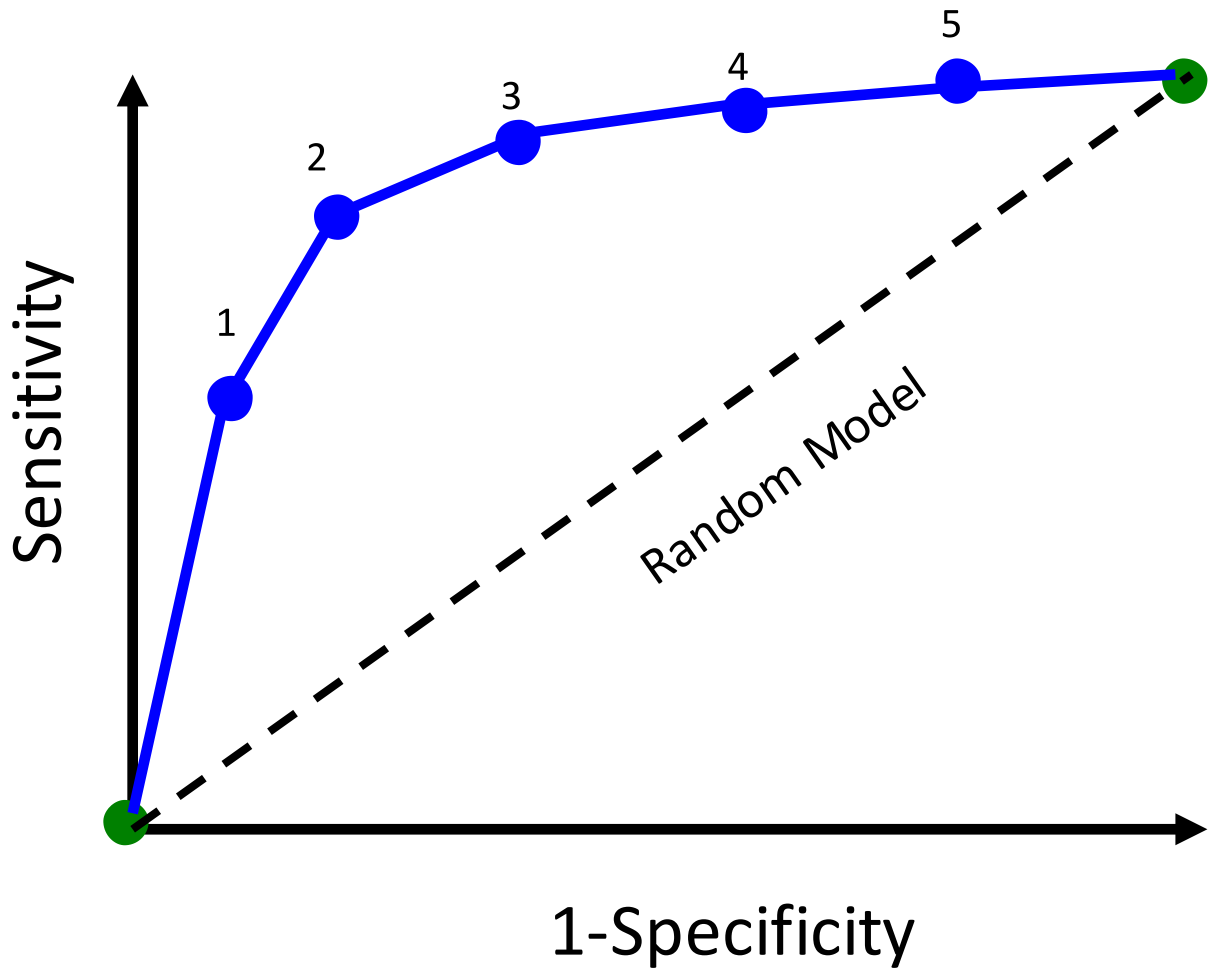
Sensitivity = 7/9
Specificity = 14/16

ROC plots – Bioclim models

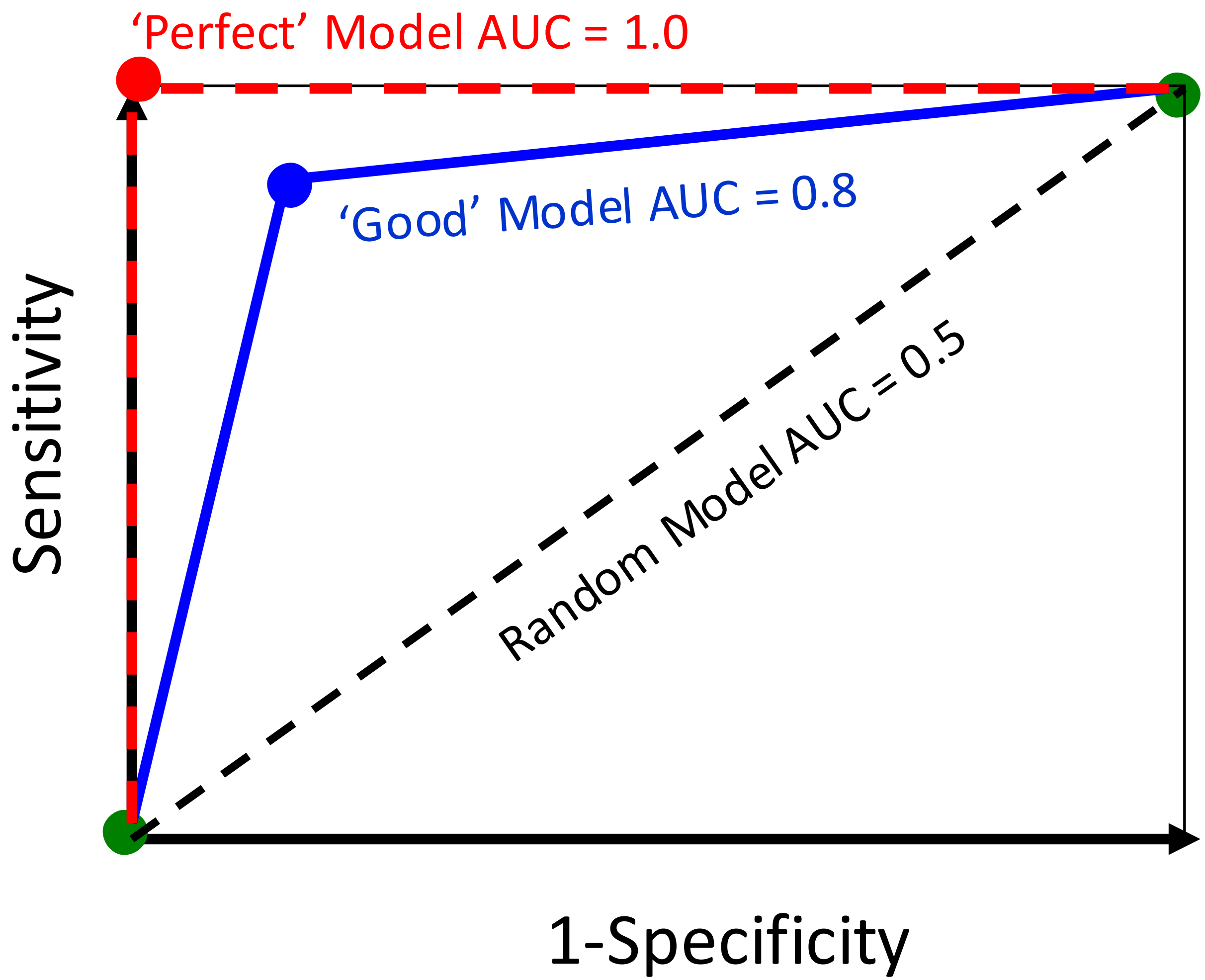


ROC plots – multi-value models

ZSL

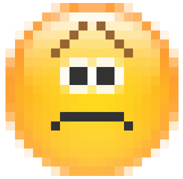


AUC – Area under a ROC plot

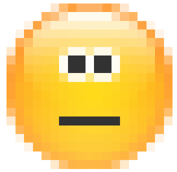


AUC – rule of thumb

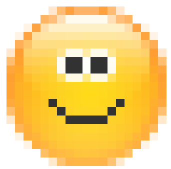
- AUC ≤ 0.6 *fail*



- 0.6 < AUC < 0.7 is *poor*



- 0.7 < AUC < 0.8 is *fair*
- 0.8 < AUC < 0.9 is *good*



- AUC ≥ 0.90 is *excellent*

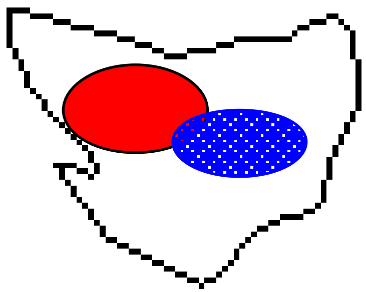
Kappa or AUC?

- Kappa can be used to determine the optimal threshold value
- … but Kappa is dependent on the choice of threshold
- Kappa can be sensitive to absolute numbers of positives in the validation set

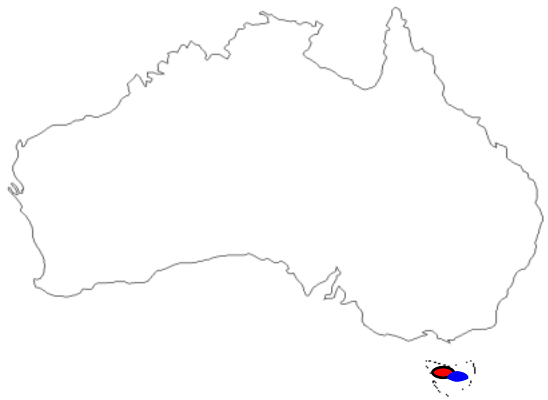
- AUC is independent of threshold
- AUC is generally preferred to Kappa
- But AUC is heavily criticised too, particularly where pseudo-absence data is used

Absence data

- Most model evaluation requires ‘validation’ data
- This should be independent of the data used to build the model
- It should include where species do and DO NOT occur
- Absence data is very difficult to obtain
- Picking the wrong background influences model evaluation



Different?



Similar?

Validation in practice

ZSL

- Validation data must be independent of model-building data
- Usually distribution data is randomly partitioned
  - 30% for validation
  - 70% for model building
- Validation requires absence data
  - Usually not available
  - Can we assume absence for areas with no presence data?
  - Pseudo absence if often used (a random selection of areas with no presence data)

References

The main reference describing validation

- Fielding, A. H. & Bell, J. F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*. **24**, 38-49.

Example using validation to compare models

- Elith, J., *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography.* **29**, 129-151.

Problems - summary

- Garbage in – Garbage out
  - *Check your data quality*
- Sampling biases
  - *all data is biased*
- Scale
- Uncertainty
- High validation scores = good model?

FAQ

ZSL

- Have I got enough data?
  - *No*
- Model output = distribution, right?
  - *No*
- I ran one model with the latest algorithm, am I done?
  - *No, its best to run multiple models with multiple algorithms to get a better understanding of the data*
- Where can I get environmental data from?
  - *… that's tricky*

IFAQ (infrequently asked questions)

- How is my data biased?
  - *all data is biased, understanding how it useful*
- Am I using too many environmental layers?
  - *many layers cause over-fitting*
- How has my choice of algorithm & parameters influenced my results?
  - *try different algorithms & parameters to find out*
- Can I believe the validation statistics?
  - *multiple validation methodologies all with 'good' results is best*

Happy modelling

- Slides, practical and data are available online
- http://www.zsl.org/science/ioz-staff-students/dr-chris-yesson/qgis-workshop-tanzania-nov-2010,1468,AR.html