Instituto Nacional de Investigación en **Salud Pública INSPI**

Workshop

Vector distribution modeling in Geographic Information System (GIS)

Quito, 19th july 2016

EPISIG

Plataforma integrada de epidemiología, geomática, bioinformática y bioestadística. del INSPI

Plan

	19th july 2016
H 1	TerrSet introduction (map composition) Species presence points
H2 – H3	Climatological time series analysis (deseason, PCA, STA)
H4 – H5	Maxent application
H6	Result analysis for decision support

Summary

In order to locate risk zones for disease vectors, the use of species distribution models is becoming widely recognized. This workshop is aimed to give a practical example of how to prepare input data for the model, how to set modeling parameters and how to interpret the results obtained.

- The first part of the workshop will apply geospatial mathematical operations to climatological time series (deseason, PCA, STA) for use in the Maxent model.
- After the structuration of presence points of dengue vector in the GIS, the second part of the workshop will focus on the Maxent model, searching to optimize its configuration parameters in function of validation results.
- The workshop will conclude applying GIS operators to the relative probability maps for use as decision support by the public health authorities.

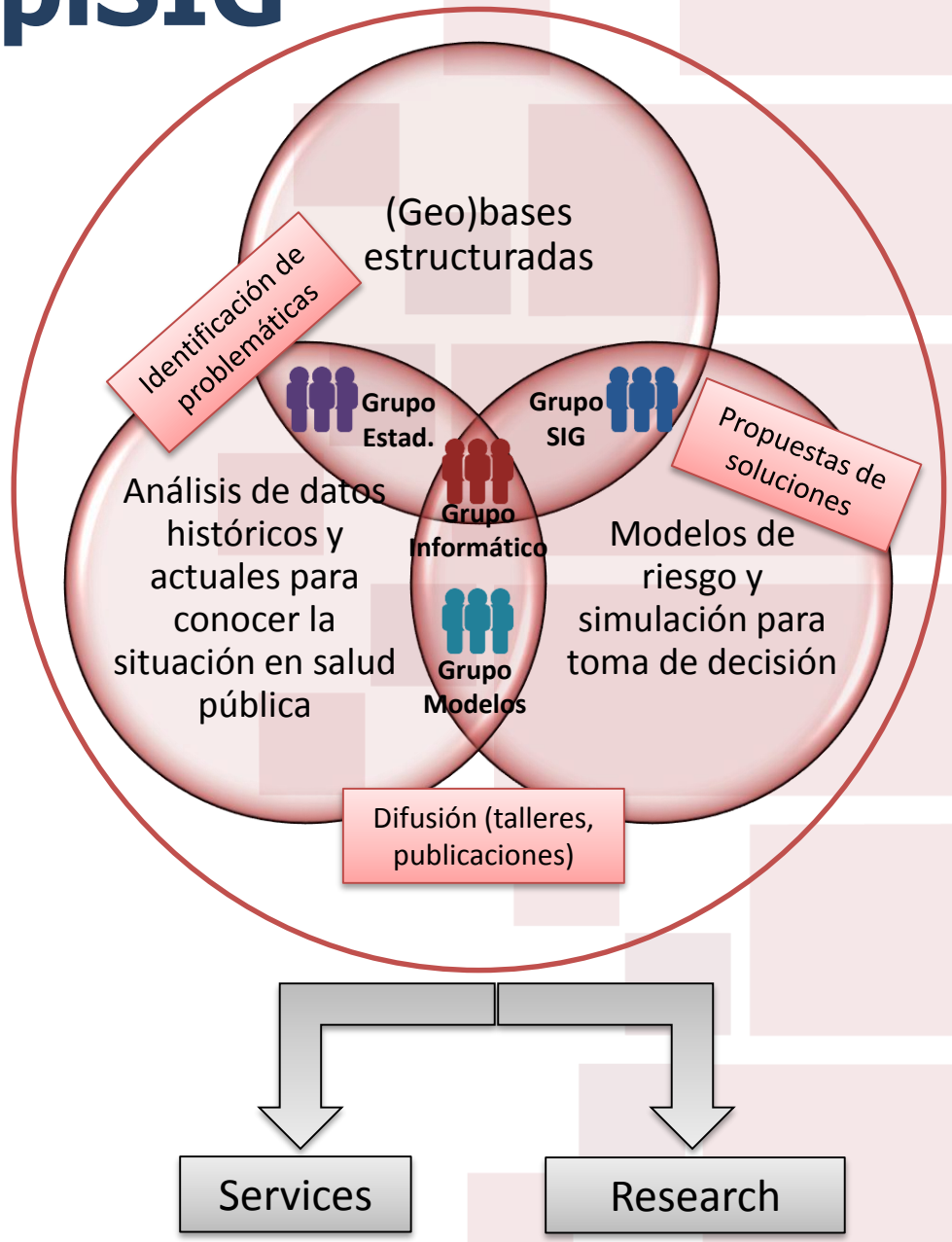
Workshop introduction

- **Purpose of the exercise** : using the example of the mosquito *Aedes aegypti*, illustrate the preparation of environmental input data and the application of the maximum entropy method to model the spatial distribution of the vector in a Geographic Information System (GIS)
- **Mathematical concepts used** :
  - Time series analysis : Deseasoning, Principal Component Analysis, T-mode, S-mode
  - Maximum entropy

Plataforma EpiSIG

The EpiSIG core utility is in charge of:

Compiling, structuring, **analyzing, modeling** and presenting results related to public, particularly to provide information and tolos to the health authorities (MSP), in order to support decision taking.



INSPI-Clark Labs agreement

- INSPI-EpiSIG = one of the 19 Resource Centers of Clark Labs in the world

  https://clarklabs.org/

– Benefits :
  - INSPI is provided with a TerrSet server license
  - Impulse to be the reference in Ecuador for spatio-temporal modeling using TerrSet, mainly apply to health related problems

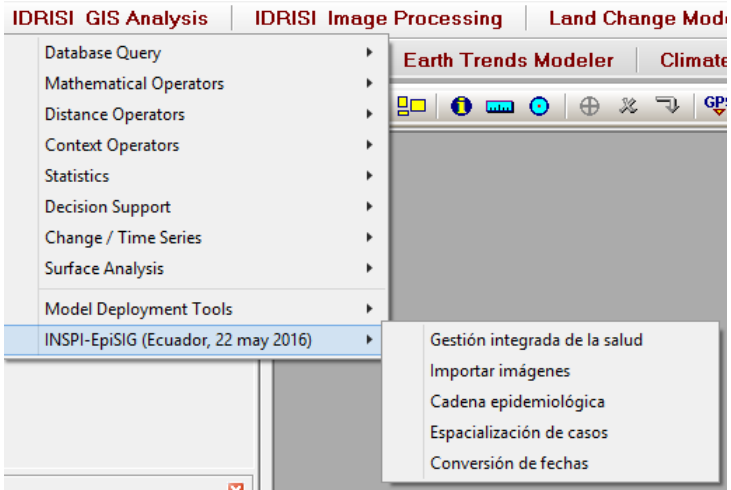
CLARK LABS

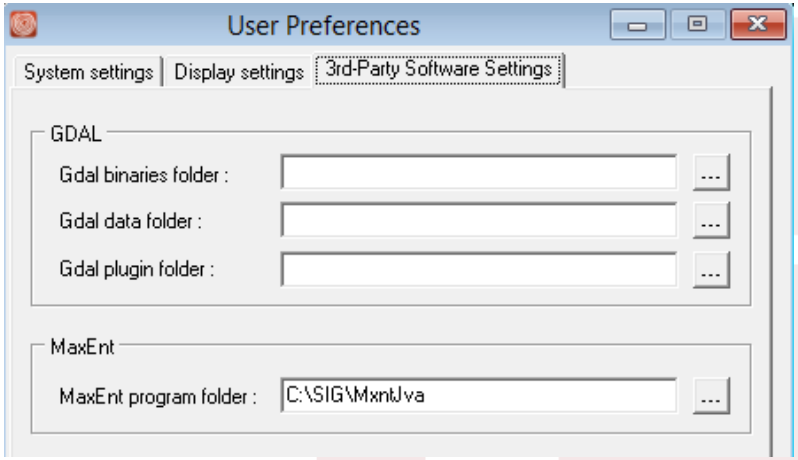
TerrSet characteristics

- More dedicated to raster images processing
- Basic modules (horizontal)
- Integrated modules (vertical)
- Allow adding our own modules

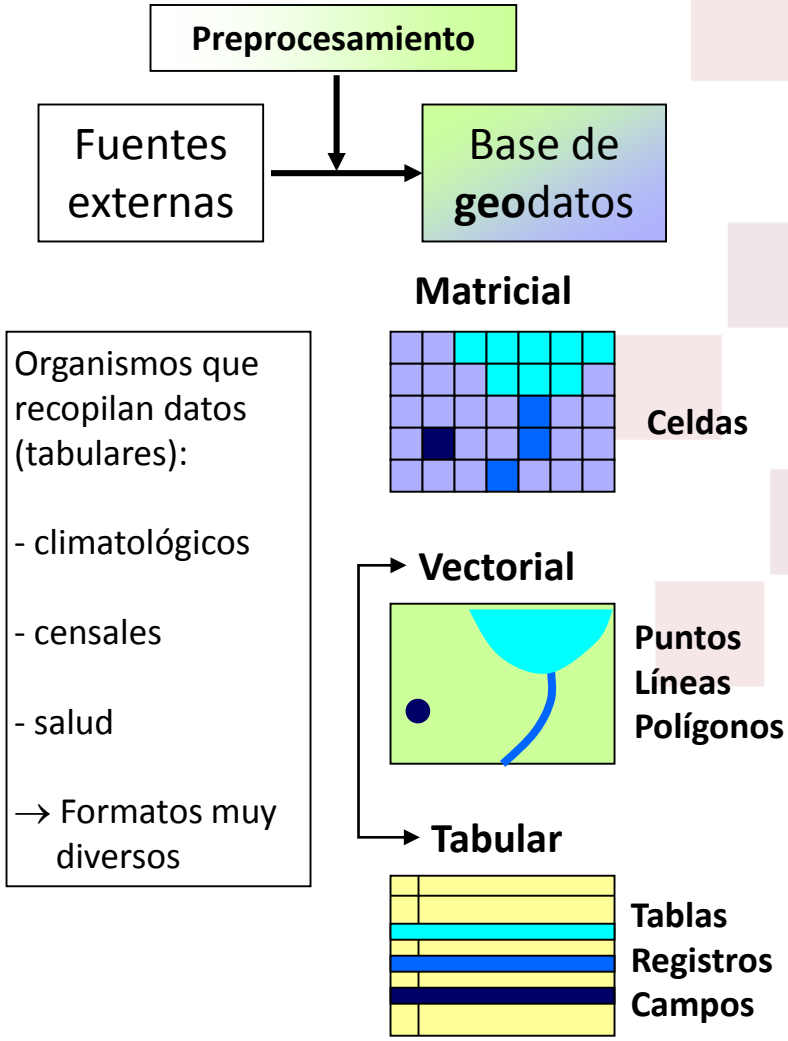
Product: TerrSet
Version: 18.21

Note : in Regional Settings, make sure to use the point as decimal symbol





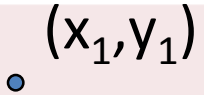
GIS formats


Preprocesamiento
Fuentes externas
Base de geodatos
Organismos que recopilan datos (tabulares):
- climatológicos
- censales
- salud
→ Formatos muy diversos
Matricial
Celdas
Vectorial
Puntos
Líneas
Polígonos
Tabular
Tablas
Registros
Campos

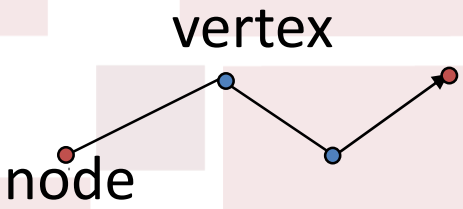
Vector objectsjetos vectoriales

Spatial **vector** entities are of 3 types :

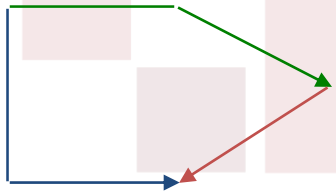
Point – pair of coordinates x e y



Line- point list

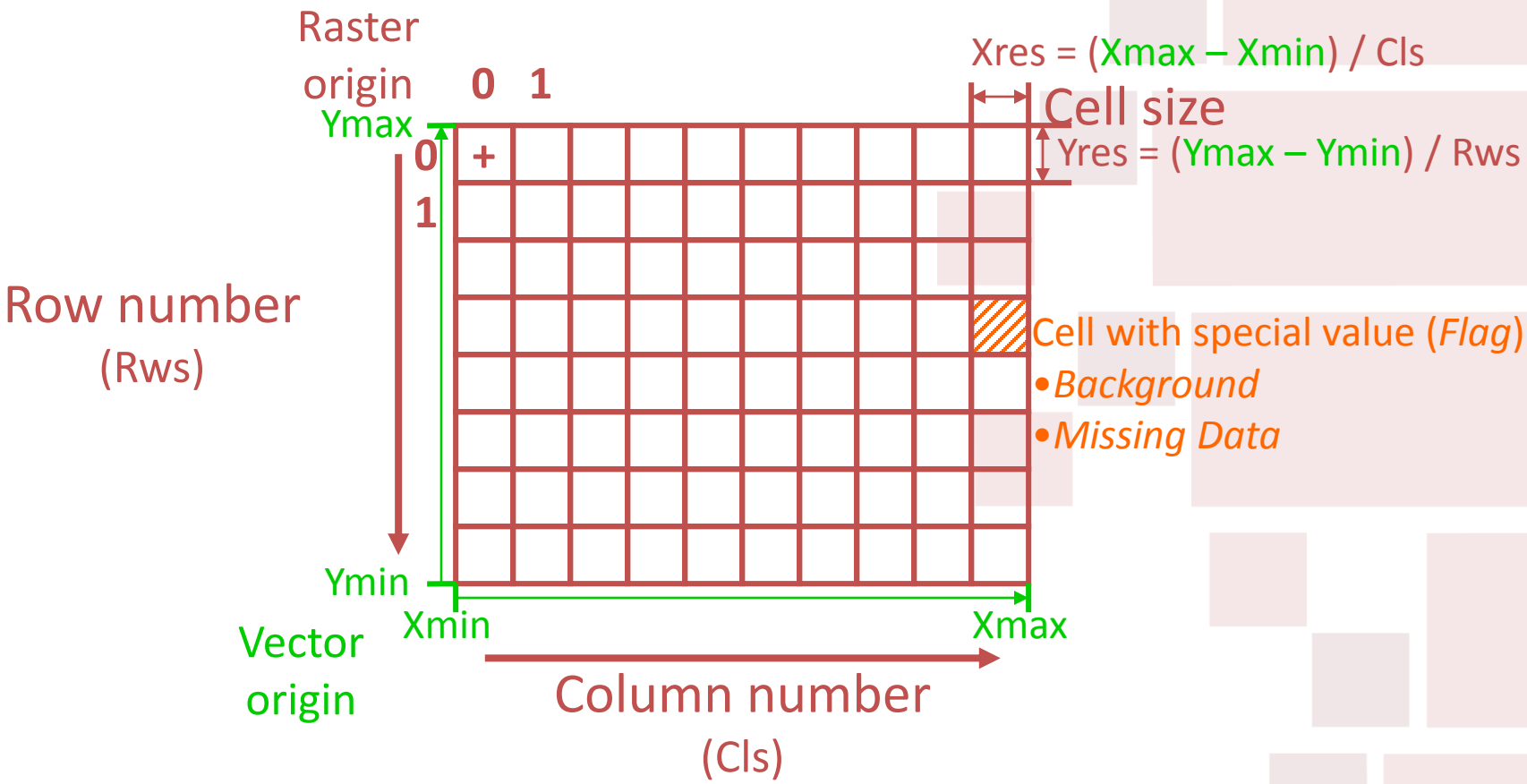


Polygon – closed line set

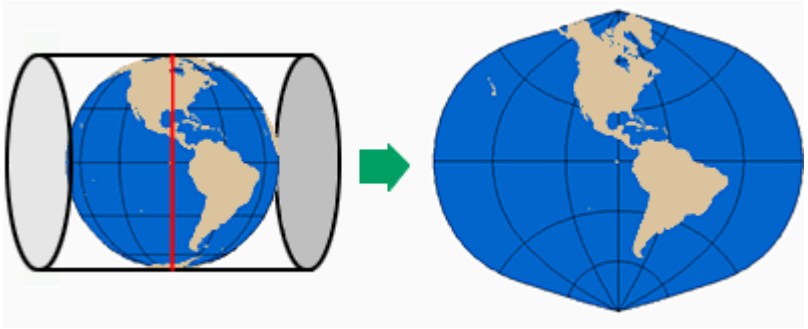


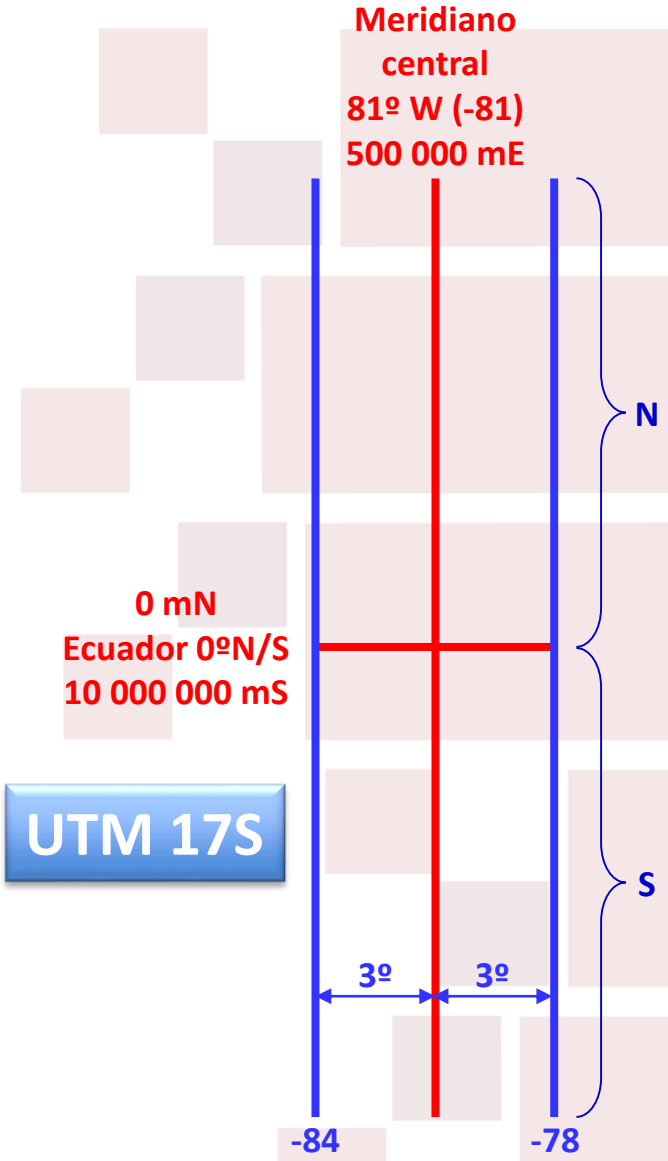
Raster format

Spatial **raster** entity is based on the **cell** or **pixel** unit.



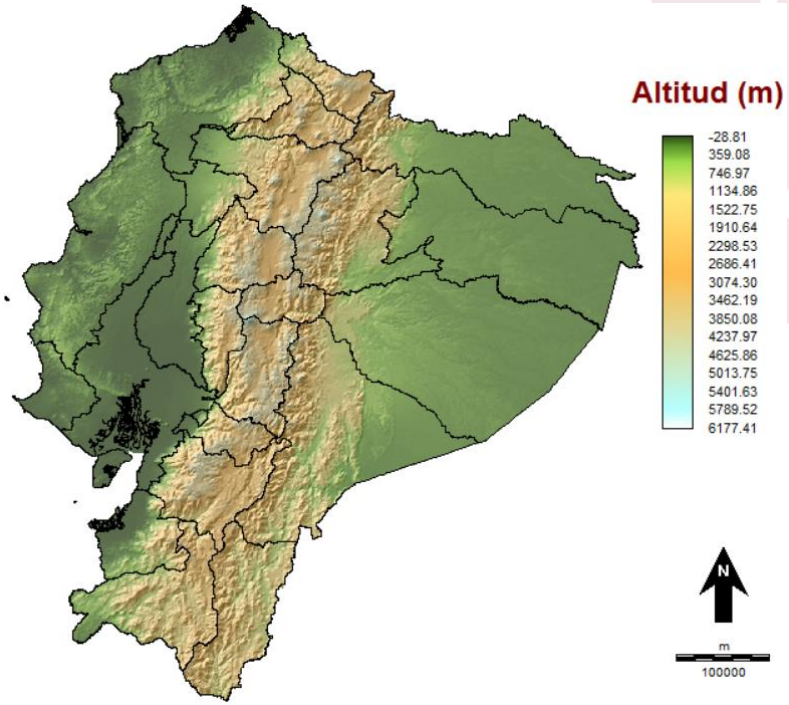
Reference system used in Ecuador





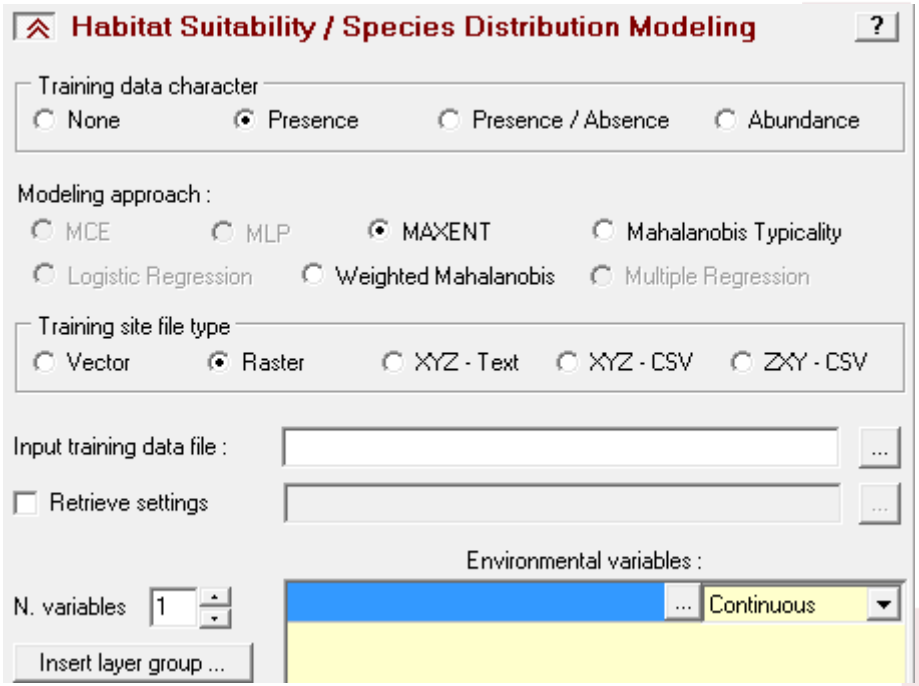
Familiarization with TerrSet

- Exercise : prepare a map composition of the elevation of Ecuador



Species distribution modeling

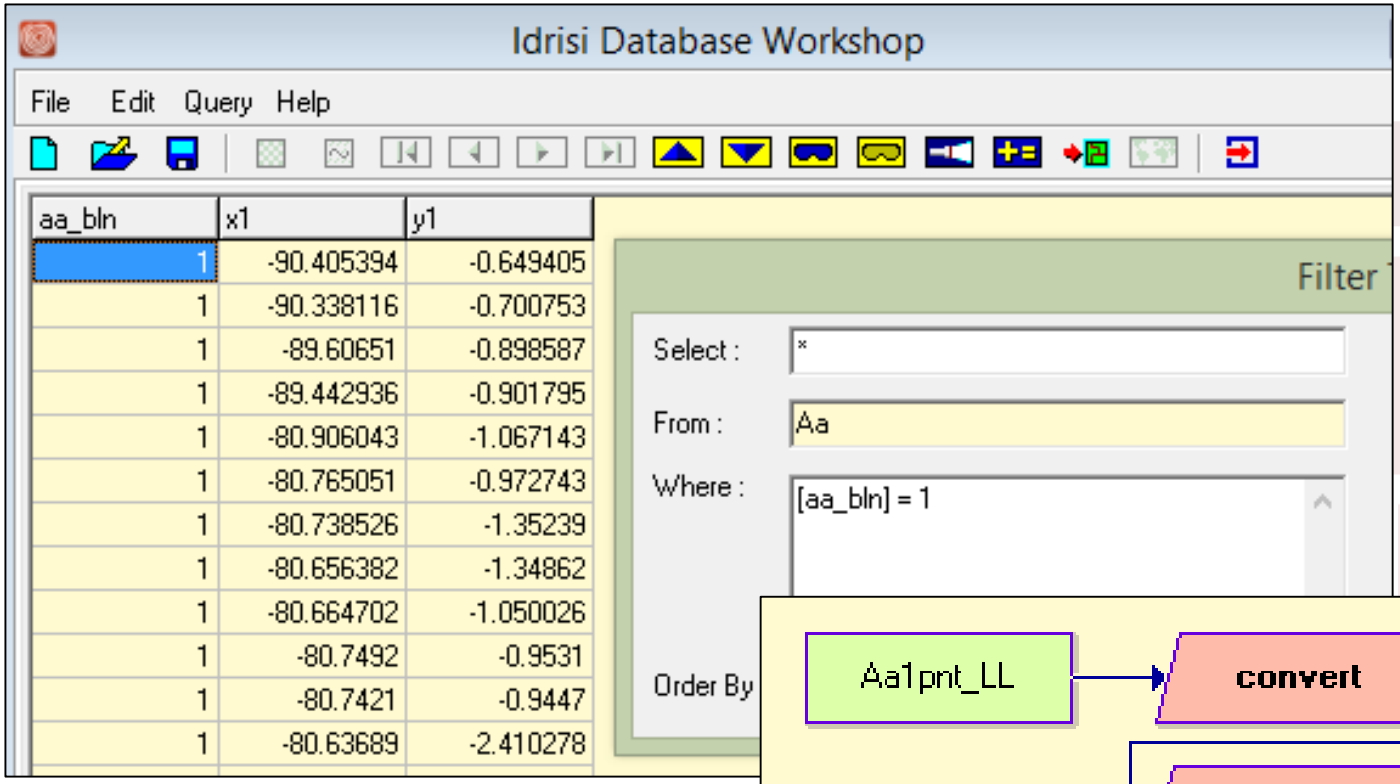
Type of collected data	Available method
None	Multi-criteria evaluation (MCE)
Presence (1)	Mahalanobis Typicality Weighted Mahalanobis **Maximum entropy**
Presence / Absence (1 / 0)	Logistic regresion (Multi-Layer Perceptron)
Abundance (0,1, … , n)	Multiple regression

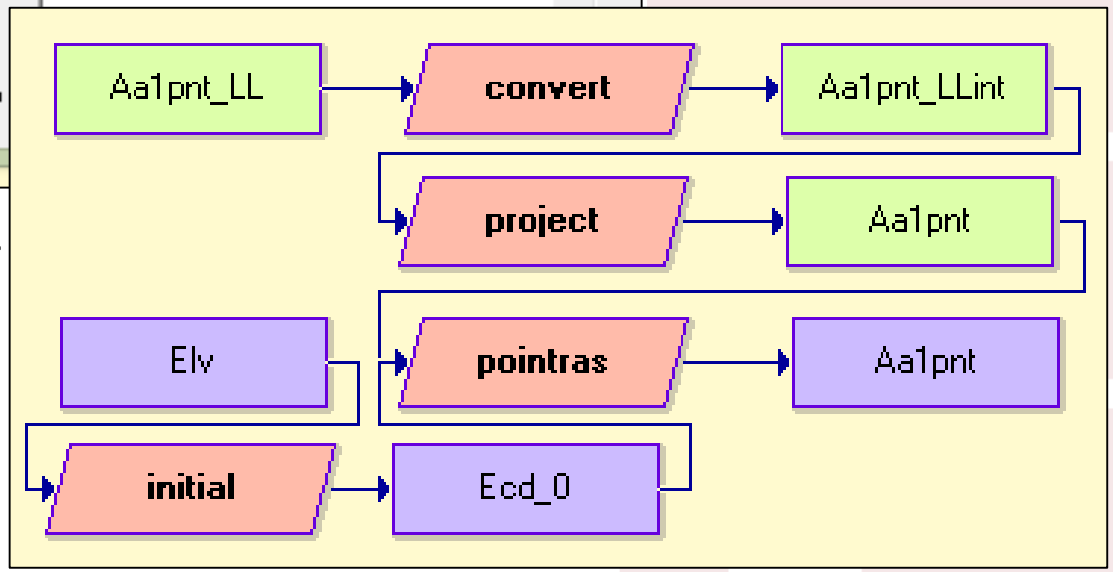


Input images

Explicative variable	Measure unit	Time	Time unit	Spatial resolution	Source
Elevation SRTM	m (above sea level)	2000	1 mission	1” (~30 m)	http://earth explorer.usgs.gov
Precipitation TRMM	Intensity mm/h → mm	2010-2015	Monthly	0.25° ~25 km	ftp://disc2.nascom.nasa.gov /ftp/data/s4pa//TRMM_L3/ TRMM_3B43/
Temperature LST MODIS	Daily mean 50×(°C+273.15)→°C	2010-2015	Monthly	0.05° 5 km	http://e4ftl01.cr.usgs.gov/MOLT/MOD11C3.005/
Land cover MODIS	Vegetation index VI × 10000 →[-1 1]	2010-2015	8-days	250 m	http://e4ftl01.cr.usgs.gov/MOLT/MOD13Q1.005/
Population INEC	Density hab/km^2	2010	1 census	Census tract	http://www.ecuadorencigras .Gob.ec/banco-de-información/
Overcrowding INEC	Person / bedroom	2010	1 census	Census tract	http://www.ecuadorencigras .Gob.ec/banco-de-información/

Presence point preparation

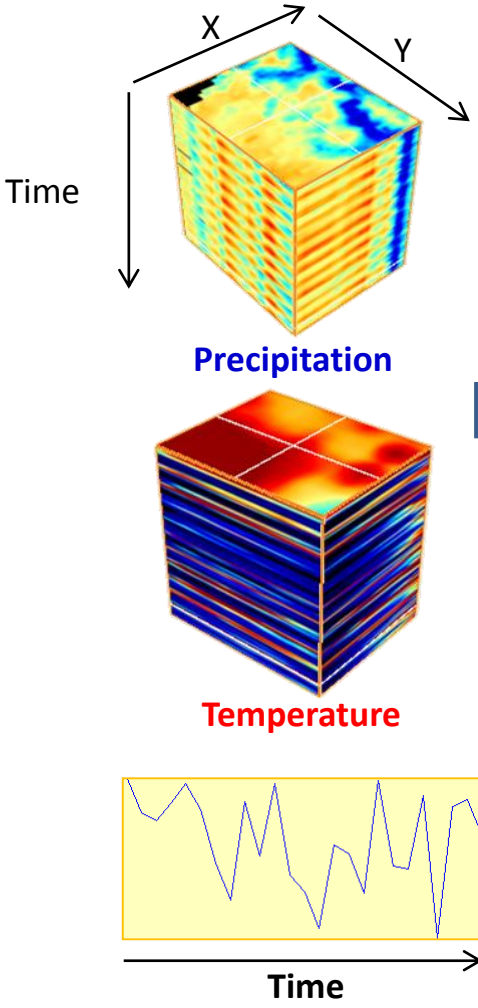




Note: this sample of points are only for exercise purpose (incomplete set that might have some error in location).

Spatio-temporal analysis

Space and time cubes of climatologic variables



- SeriesTrend Analysis
- STA (Seasonal Trend Analysis)
- PCA (Principal Components Analysis) / EOF
- EOT (Empirical Orthogonal Teleconnections)
- CCA (Canonical Correlation Analysis)
- Fourier PCA Spectral Analysis
- Linear Modeling

Dependent series : Entomologic cube

Independent variable type

Image series Index series

Independent series : Lag : Series name :

0

N Series : 1

Climatic cubes

Remove series

Slope(s) and intercept

R R2

Adjusted R2 Partial R

Create residual series

Apply mask :

Output prefix :

Run

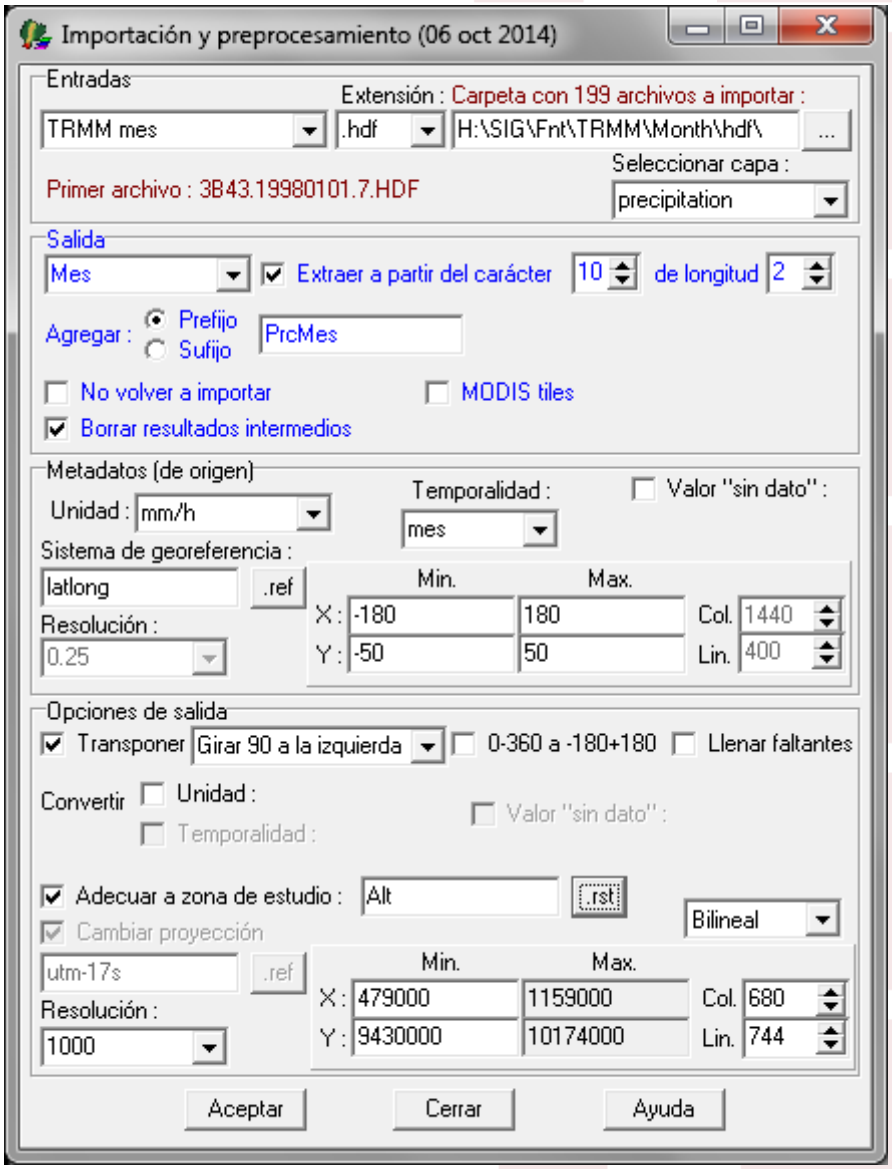
Time series analysis

Method	Purpose	Apply to deseasoned serie
Series Trend Analysis	Interannual trend	Should
STA (Seasonal Trend Analysis)	Annual seasonal progression	**No**
PCA (Principal Components Analysis) / EOF	Irregular but recurrent patterns in space/time	Should
EOT (Empirical Orthogonal Teleconnections)	Related patterns of variation between widely separated areas of the globe	Should Time-consuming
CCA (Canonical Correlation Analysis)	Requires 2 series	
Fourier PCA Spectral Analysis	Cyclical components	Should not
Linear Modeling	Requires 2 series	Should

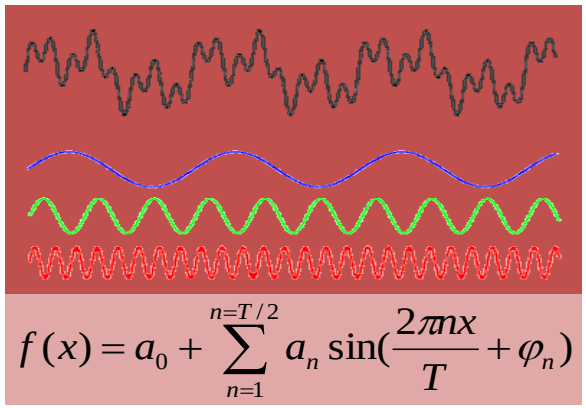
Spatio-temporal cube preparation

Module developped by
INSPI-EpiSIG
(running under TerrSet)

Seleccionar también **Año** a partir de **6** con longitud **4**



Seasonal Trend Analysis



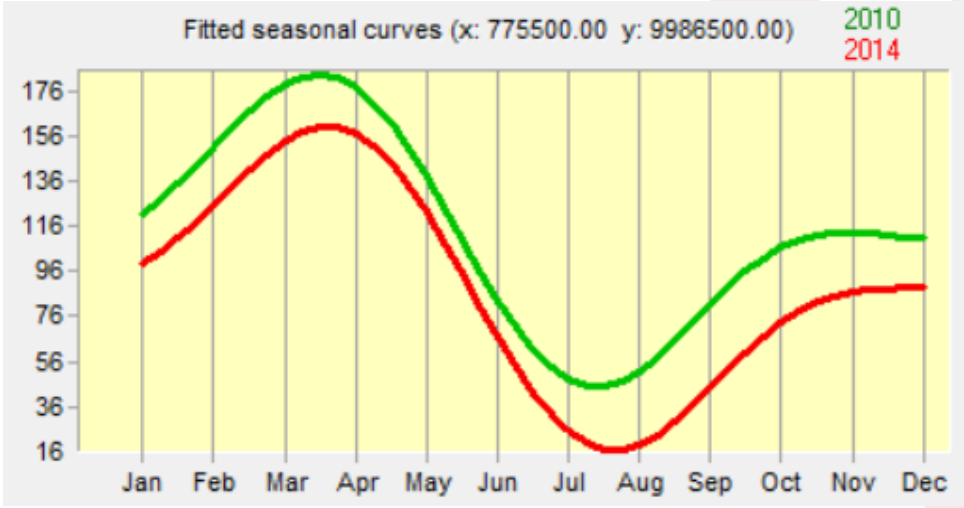
$$f(x) = a_0 + \sum_{n=1}^{n=T/2} a_n \sin\left(\frac{2\pi n x}{T} + \varphi_n\right)$$

Two stage analytical process

1. Harmonic Analysis of each year in the series to determine the best fit mean value (Amplitude 0), annual cycle (Amplitude 1 and Phase 1) and semi-annual cycle (Amplitude 2 and Phase 2)
2. Theil-Sen Median Slope operator to determine trends in these five parameters

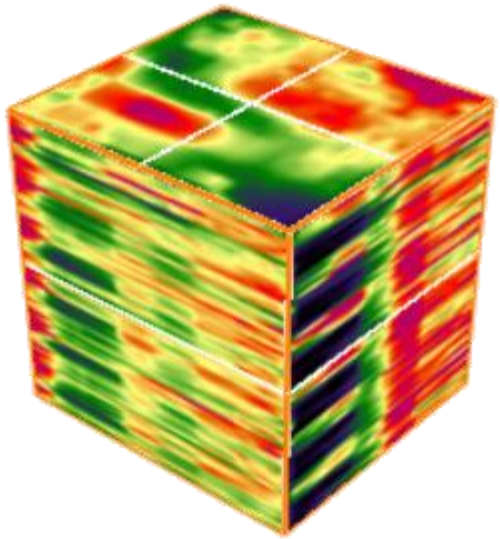
Advantages

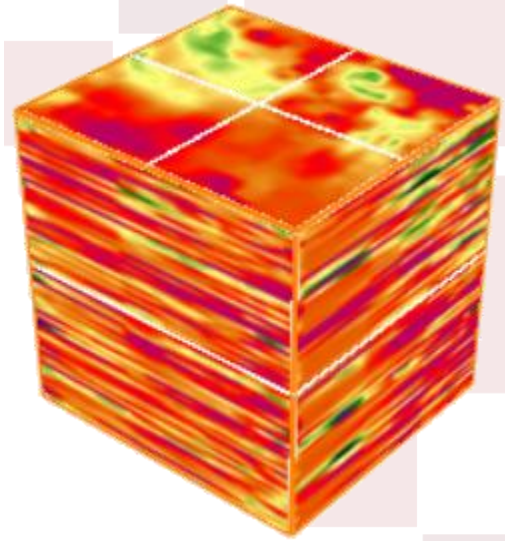
1. No need to identify seasonal/phenological events – models each year as a totality
2. Removes short-term variability with less than a 6-month frequency
3. Removes interannual up to 30% of the length of the series.

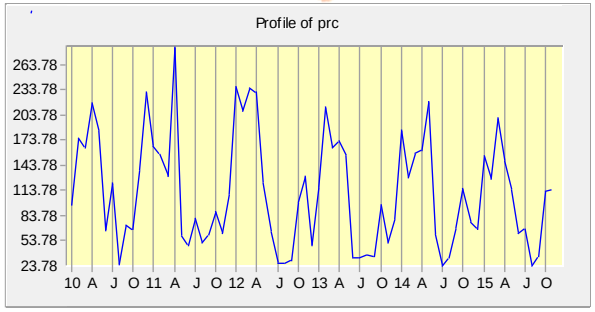


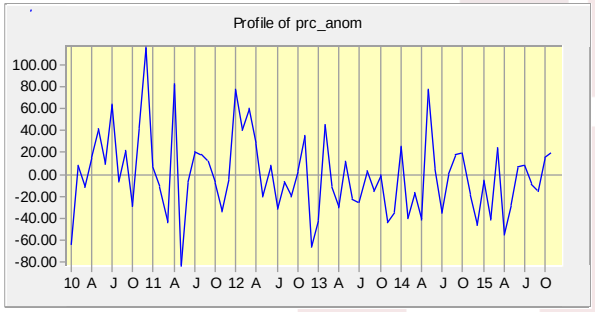
Deseason

Removal of seasonality in a time series : calculate the deviation from the mean; standardized anomalies includes the division by the standard deviation.

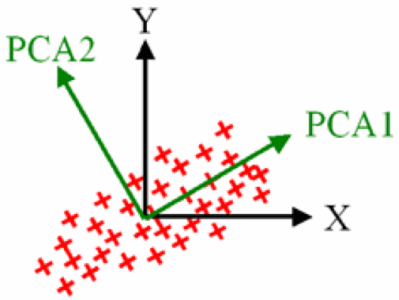








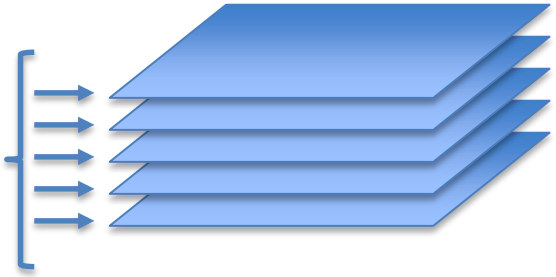
Principal Component Analysis



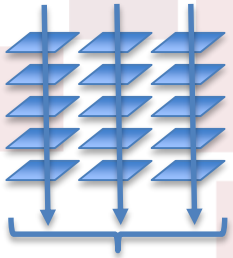
orthogonal transformation of n-dimensional image data that produces a new set of images (components) that are uncorrelated with one another and ordered with respect to the amount of variation (information) they represent from the original image set

T-mode

Variables = time layers



S-mode



Variables = pixels

- With T-mode analysis, the image is the component and the graph expresses the loading – the correlation between the component image pattern and each image in the series.
- With S-mode (Empirical Orthogonal Function analysis), the graph is the component and the image contains the loadings.

MaxEnt methodology

Let X be any geographic region of interest.

X:

- is a set of discrete cells (a finite set of cells)
- $x_1, \dots, x_m \in X$, is a set of point.

It represents the locations where the species has been observed y recorded.

$$x_1, \dots, x_m \in X$$

selected independently of X according to an unknown probability distribution π and our objective is to estimate π.

(π coincides with the biological concept of potential distribution or fundamental niche)

MaxEnt methodology

Now the problem is one of "density estimation"

The set of points $x_1, \dots, x_m$ has been chosen independently of an unknown distribution π and the purpose is to build a distribution $\hat{\pi}$ that approximates π.

For the construction of $\hat{\pi}$

Let $f_1, \dots, f_n$ be a set of functions where

$$f_j : X \to \mathbb{R}$$

f: represents the vector of the n functions or environmental variables

$$\forall f, \qquad f : X \to \mathbb{R}$$

$\pi[f]$ is defined by way of

$$\tilde{\pi}(x) = \frac{|\{1 \le i \le m : x_i = x\}|}{m}, \text{ empirical distribution}$$

$$\forall f, \tilde{\pi}[f] \text{ is empirical average of } f$$

which will be as near as possible to its true value $\pi[f]$), thus an approximation of π must be find under which

$$\hat{\pi}[f] \approx \tilde{\pi}[f]$$

MaxEnt methodology

There are several distributions which satisfy these constraints.

The one with maximum entropy is chosen, it is the function which is closest to the uniform distribution.

Observe that the entropy of any distribution p in X is defined by:

$$H(p) = -\sum_{x \in X} p(x)\ln(p(x))$$

So the problem is:

Estimate π by $\hat{\pi}$ distribution of maximum entropy subject to the condition that

$$\forall f_j, \hat{\pi}[f_j] \approx \tilde{\pi}[f_j]$$

To solve this problem, a family of distributions functions named "Distributions Gibbs" and defined as follows is used:

$$q_\lambda(x) = \frac{e^{\lambda . f(x)}}{Z_\lambda}$$

MaxEnt methodology

Where,

$$Z_\lambda = \sum_{x \in X} e^{\lambda . f(x)}$$

is a normalization constant and $\lambda \in \mathbb{R}^n$.

It can be demonstrated that the distribution of maximum entropy (MaxEnt) is the same as the Gibbs distribution maximum likelihood, this is the distribution q_λ that minimizes

$$RE(\tilde{\pi} \| q_\lambda)$$

Where

$$RE(p \| q) = \sum_{x \in X} p(x) \ln \left(\frac{p(x)}{q(x)} \right)$$

denotes the relative entropy or Kullback-Leibler divergence.

Reference : S. J. Phillips, R. P. Anderson, and R. E. Schapire, 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190:231-259.

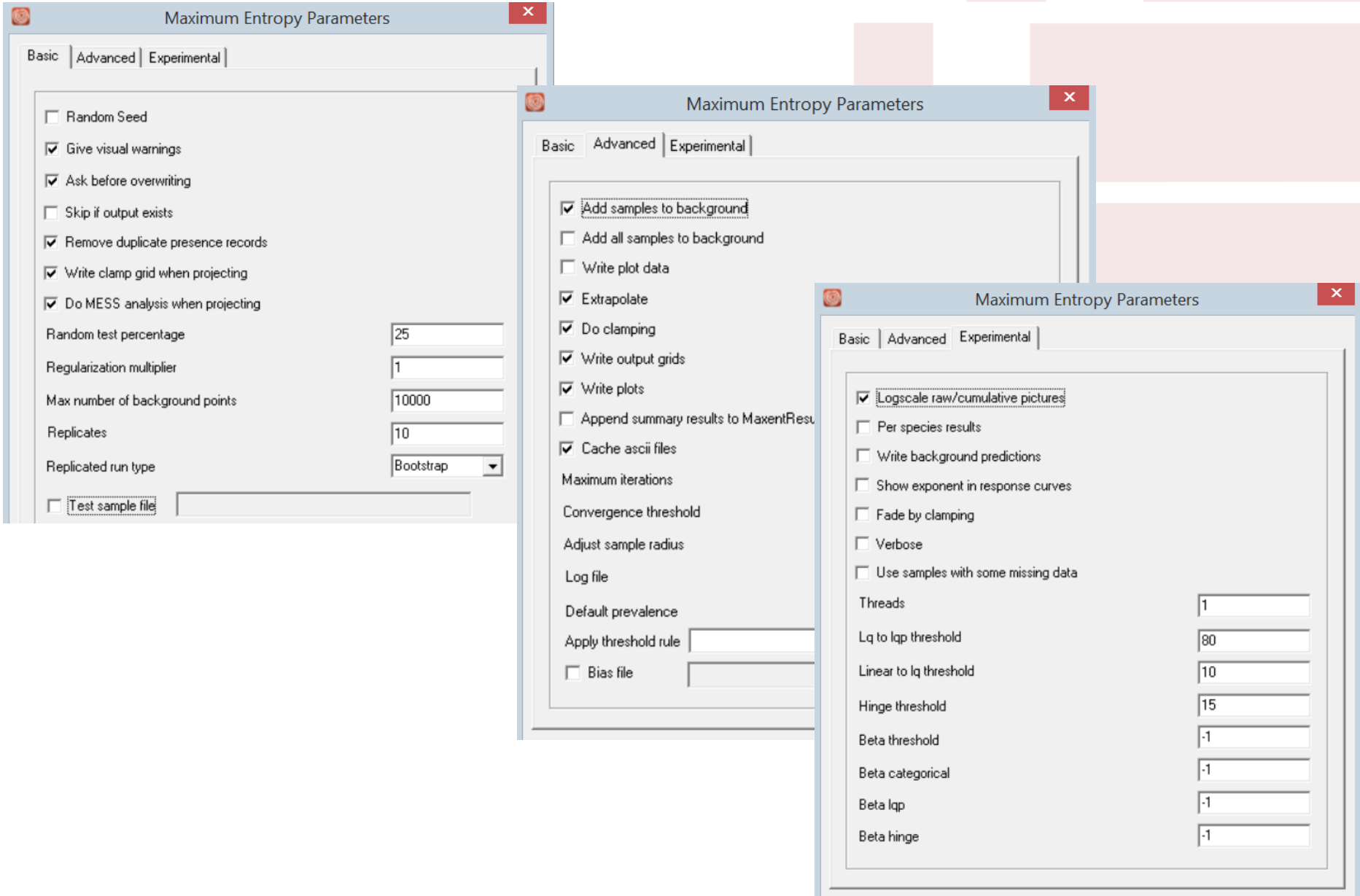
Parameters settings

- 75% of the presence points for training, 25% for validation
- 10 replicates (Bootstrapping, for sensibility analysis)
- To improve the AUC, it was observed that it is necessary to have a number of *background* points twice more than presence points

Notes :

- explicative rasters must have Flag value (indicate -9999 if none) and Flag description (Background);
- it is highly suggested to copy all the ecplicative rasters in the same folder and create a raster group;
- MaxEnt requires Java program installed.

Setting parameters



Maxent settings

Basic

Give visual warnings **No**

Random test percentage **25**

Replicates **10**

Replicated run type **Bootstrap**

Advanced

Write output grids **No**

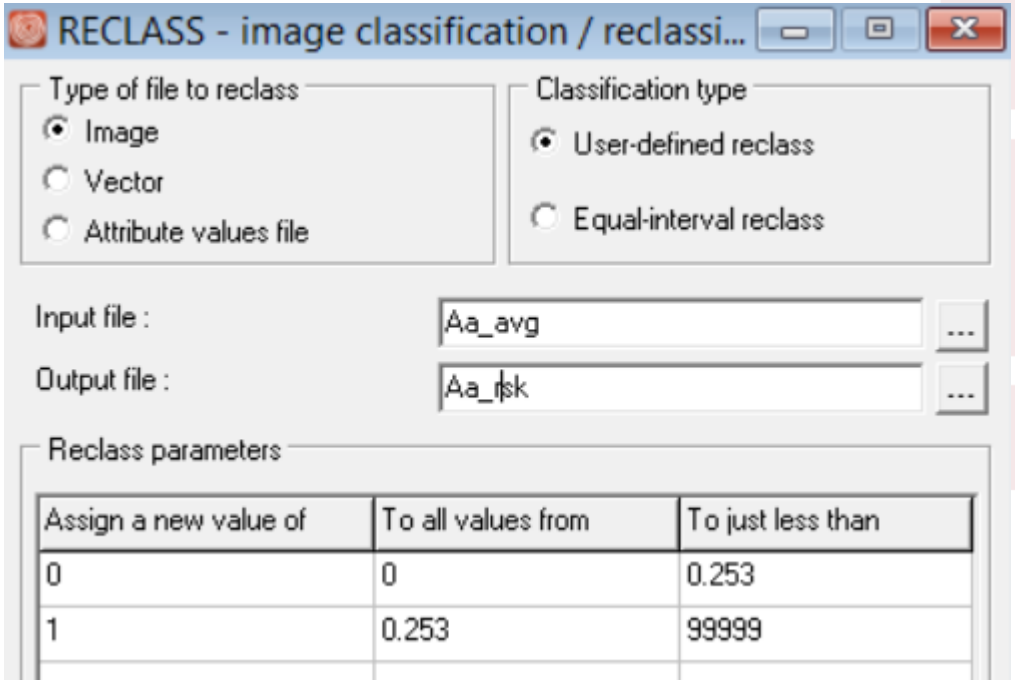
Apply threshold rule **Equal training sensitivity and specificity**

Maxent memory usage

1024mb (2048mb can give error)

Threshold for reclassification

Logistic threshold	Description
0.253	Equal training sensitivity and specificity



Practical questions

- How to summarize time series images for MaxEnt?
- How to set general parameters of MaxEnt optimally?
- Period to consider for time series?
- Presence points very close : useful or noisy?

Precisión	Decimal de grado	Puntos de presencia
100 km	0	27
10 km	1	162
1 km	2	486
100 m	3	1305
10 m	4	1943
1 m	5	2002

For more information

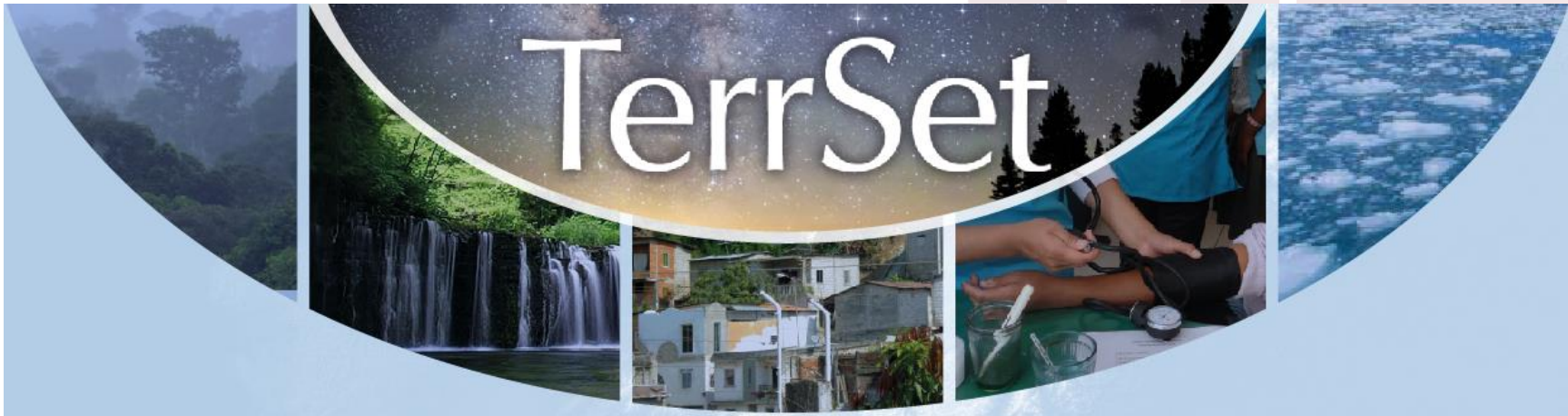
- E-mails

  episig.inspi@gmail.com, episig@inspi.gob.ec

- Web site of EpiSIG

  http://www.investigacionsalud.gob.ec/webs/episig/

  – Request form online



El Instituto Nacional de Investigación en Salud Pública INSPI "Dr. Leopoldo Izquieta Pérez", a través de su Plataforma integrada de epidemiología, geomática, bioinformática y bioestadística EpiSIG, invita al:

CURSO TALLER DE GEOMÁTICA "APLICACIONES INTEGRADAS DE SIG UTILIZANDO TERRSET"

3-7 octubre 2016 / 8h00 – 17h30 / Quito, Ecuador

Cupo limitado a 45 personas

Se hará una previa selección basada en la formación o empleo actual y motivación de los candidatos.

Taller sin costo

Para más información ingresa a: http://www.investigacionsalud.gob.ec/webs/episig/taller/#taller3

Instituto Nacional de Investigación en Salud Pública INSPI
Dr. Leopoldo Izquieta Pérez

EPISIG

CLARK LABS