Herramientas para el Análisis ROC de modelos espaciales

Instrucciones de instalación y ejemplos de aplicación

Tools for ROC analysis of spatial models

Installation instructions and application examples

J.F. Mas

Centro de Investigaciones en Geografía Ambiental

Universidad Nacional Autónoma de México (UNAM)

jfmas@ciga.unam.mx

http://www.ciga.unam.mx/

Septiembre 2013

Índice

Herramientas para el Análisis ROC

Instrucciones de instalación y ejemplos de aplicación

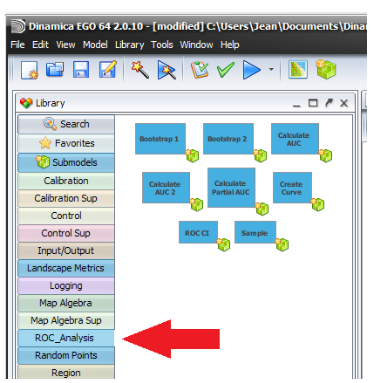
Resumen: El análisis ROC es comúnmente utilizado para evaluar el desempeño de algoritmos de clasificación. Estos submodelos desarrollados en el ámbito del paquete de modelación DINAMICA EGO (www.csr.ufmg.br/dinamica) permiten llevar a cabo un análisis ROC utilizando datos espaciales (mapas digitales en formato de celda) y pueden ser fácilmente utilizados para evaluar modelos de cambio de uso/cubierta del suelo o de distribución de especies, entre otros. Con ellos, es posible crear una curva ROC, calcular el AUC, el AUC parcial, los AUCs alto y bajo, el intervalo de confianza del AUC, la densidad de ocurrencia del evento en categorías de probabilidad y llevar a cabo pruebas estadísticas sobre la diferencia entre los AUCs de dos modelos.

Agradecimientos: Estas herramientas se desarrollaron en el ámbito de los proyectos *Elaboración y aplicación de modelos prospectivos de cambio de cobertura/uso del suelo* (PAPIIT clave RR113511) y *¿Puede la modelación espacial ayudarnos a entender los procesos de cambio de cobertura/uso del suelo y de degradación ambiental ?* (SEP-CONACYT clave 178816). Se dan agradecimientos especiales al equipo de desarrollo de DINAMICA EGO en la Universidad Federal de Minas Gerais, Brasil por su apoyo durante el desarrollo de los modelos.

Instalación

Obtener DINAMICA de www.csr.ufmg.br/dinamica/, instalarlo.

Descomprimir la carpeta ROC_tools, copiar el contenido de la carpeta "submodels" en la carpeta de submodelos de DINAMICA que se encuentra en la carpeta documentos del usuario (por ejemplo C:\Users\nombre_usuario\Documents\Dinamica EGO\Submodels). Los submodelos se encontrarán en la librería "ROC_Analysis".



Librería "ROC Analysis"

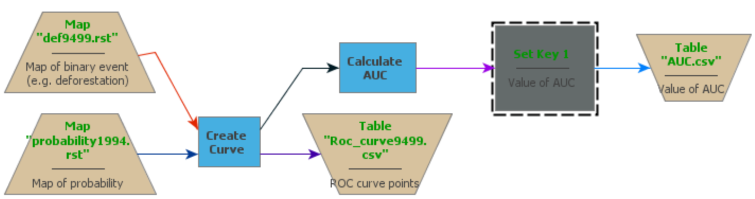
Ejemplos de aplicación

Estos ejemplos ilustran la aplicación de los modelos a la evaluación de modelos de cambios de cubierta/uso del suelo (ejemplo 1, Mato Grosso) y de distribución potencial de especies (ejemplo 2). Los modelos y los insumos están disponibles en las carpetas example 1 and example 2. Los modelos están descritos en Mas et al. (2013).

Ejemplo 1 : Deforestación en Mato-Grosso

Un mapa de probabilidad de deforestación (mapa *probability1994*) es evaluado por un análisis ROC basado en la deforestación observada entre 1994 y 1999 (mapa *def9499*).

El modelo *calculate_AUC* permite crear una tabla para construir la curve ROC y calcular el AUC.



El modelo *Compare_with_random_model* realiza una prueba para determinar si el AUC obtenida por el modelo es significativamente diferente del AUC obtenida por un modelo aleatorio (Monte Carlo).

Example 2. Modelación de Nicho

Mapas de probabilidad de distribución potencial de Bradypus variegatus fueron elaborados utilizando dos métodos: MaxEnt (mapa *MaxEnt_p*) y los pesos de evidencia (WOE, mapa *WoE_p*). Se lleva a cabo un análisis ROC con base en información no utilizada en la elaboración de los mapas de probabilidad (mapa *occurrences*).

Los Modelos cumulative*_distribution_function_MaxEnt* y *_WoE* permiten elaborar las gráficas de la función de distribución cumulativa.

El modelo *calculate_pAUC* calcula el AUC parcial y producen tablas y mapas de densidad de ocurrencia por intervalos de probabilidad.

El modelo *calculate_AUC_lower_upper* calcula los valores del AUC bajo, trapezoidal y alto.

El modelo *calculate_AUC_CI* permite calcular el intervalo de confianza del valor de AUC del modelo de los pesos de evidencia. Utiliza el submodelo *Bootstrap 1* para producir 2000 recombinaciones del mapa de probabilidad y calcula los límites del intervalo con los enfoque normalizado (Student) y de niveles porcentuales.

El modelo *compare_2_models* realiza una prueba estadística para evaluar si la diferencia de AUC de los mapas obtenidos por MaxEnt y por los pesos de evidencia es estadísticamente significativa.

El modelo *reduce_size* produce mapa de probabilidad y de ocurrencia aleatoriamente remuestrados (10% de no-evento y 100% de evento). Estos mapas más "ligeros" se usan en los procesos iterativos (bootstrapping).

Nota: Los modelos compare_2_models y *calculate_AUC_CI* que se basan en un gran número de iteraciones (bootstrapping) corren más rápido (menos lento!) en el Console Launcher.

Tools for ROC analysis of spatial models

Installation instructions and application examples

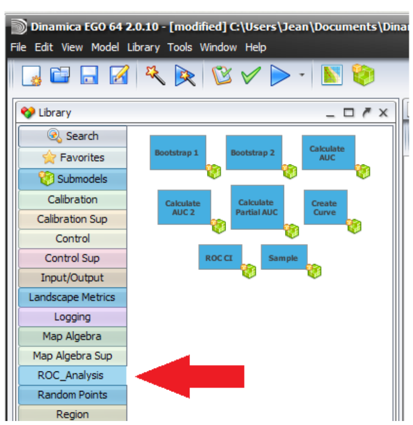
Summary: ROC analysis is commonly used to evaluate the performance of classification algorithms. These sub-models, developed within the modeling package DINAMICA (www.csr.ufmg.br/dinamica) enable the user to perform ROC analysis using spatial data (digital maps in raster format) and can be easily used to assess change use / land cover change and species distribution models, among others. The tools include the elaboration of the ROC curve, the computing of the area under the curve (AUC), partial AUC, lower and upper AUCs, the confidence interval of AUC (or pAUC), the density of event in probability bins and tests to evaluate the difference between the AUCs of two models.

Acknowledgements: These tools are developed in the scope of the projects *Elaboración y aplicación de modelos prospectivos de cambio de cobertura/uso del suelo* (PAPIIT # RR113511) and *¿Puede la modelación espacial ayudarnos a entender los procesos de cambio de cobertura/uso del suelo y de degradación ambiental ?* (SEP-CONACYT # 178816). Special thanks are given to the development team of DINAMICA EGO at the Federal University of Minas Gerais, Brazil for their support during the development of the models.

Installation

Obtain DINAMICA from **www.csr.ufmg.br/dinamica/***,* install it.

Unzip the file ROC_tools, copy the files of the folder "submodels" into the submodels folder of DINAMICA that can be found in the User's Documents Folder (for example C:\Users\user_name\Documents\Dinamica EGO\Submodels). In DINAMICA, submodels will be available at the ROC_Analysis library.



ROC Analysis Library

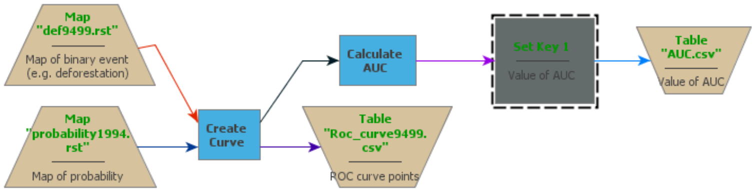
Examples of application

These examples illustrate the application of the models to evaluate land use / land cover change models (Example 1, Mato Grosso) and potential distribution of species (Example 2). The models and inputs are available in folders example 1 and example 2. The models are described in Mas et al. (2013).

Example 1 : Deforestation in Mato-Grosso

A map of the probability of deforestation (map *probability1994*) is evaluated through a ROC analysis based on observed deforestation between 1994 and 1999 (map *def9499*).

Model *calculate_AUC* enables user to create a table in order to elaborate the ROC curve and calculate the AUC.



The model *Compare_with_random_model* performs a statistical test to determine whether the AUC obtained by the model is significantly different from the AUC obtained by a random model (Monte Carlo).

Example 2. Niche modelling

Maps of potential distribution of Bradypus variegatus were elaborated using two methods: MaxEnt (map *MaxEnt_p*) and the Weights of Evidence (WOE, map WoE_p). ROC analysis is carried out using a data not used in the elaboration of the probability maps (map *occurrences*).

Models *cumulative_distribution_function_MaxEnt* and *cumulative_distribution_function_WoE* enables user to elaborate the graphs of cumulative distribution functions.

Model *calculate_pAUC* computes the partial AUC and produces tables and maps of the density of species occurrence in each bin.

Model *calculate_AUC_lower_upper* calculates the value of AUC lower, AUC (trapezoidal) and AUC upper.

Model *calculate_AUC_CI* enables user to calculate the confidence interval of the WoE model's AUC. It uses the submodel *Bootstrap 1* to produce 2000 replicates of the probability map and calculates the bounds of the confidence interval using both Normal (Student) and percentile approaches.

Model *compare_2_models* carries out a statistical test to evaluate whether the difference of the AUC of maps obtained the MaxEnt and WoE methods is significant.

Model *reduce_size* produces a map of probability and occurrence randomly resampled (10% of no-event and 100% of event cells). These "lighter" maps are used in iterative processing (bootstrapping).

Note: Models *compare_2_models* and *calculate_AUC_CI* which are based on a large number of iterations (bootstrapping) run more rapidly (less slowly !) using the Console Launcher.

Brief description on ROC Analysis submodels

ROC create curve:

Inputs:

Map of occurrence: Map of occurrence of an event such as map of observation of a species (1: cell corresponding to an occurrence point, 0 otherwise) or map of land cover change (change: 1, no change 0). The map can have no data values, no data value is declared when loading the input map (functor *Load map*).

Map of probability: Map of probability of an event (change potential map in LUCC model, probability of presence of a species.

Thresholding value: Threshold to bin the probability image (default 10%), smaller the value, more precise is the ROC curve and the AUC estimates but processing time increases.

Optional inputs:

Map of frequency: map indicating the number of time an event was observed at a cell, this input will be used only if the port "use a frequency map" is set to Yes, but anyway a map has to be connected to this port. In case of no using a frequency map, connect the occurrence map to this port.

Calculate Density of event in each interval bin: if set to yes a map and a table of event density in each bin are produced.

Peterson modification: if set to Yes, creates and uses a ROC curve that presents the % of area set to "presence" as the ROC curve X axis.

Slicing method: allows to choose among three methods to slice the probability map into bins: 1) equal probability, 2) equal area or user defined thresholds.

Outputs:

False positive values table (X axis of ROC curve).

True positive values table (Y axis of ROC curve).

ROC curve: table with False and true positive rates, allows to elaborate the ROC curve in a spreadsheet program.

Cumulative program of probability table: allows to construct the cumulative distribution function graphs.

Sliced map: sliced map of probability

Table of event density in each bin: table indicating the proportion of event cell in each bin.

Map of event density in each bin: map indicating the proportion of event cell in each bin.

Note: False and true positive values tables are the inputs of submodels *Calculate AUC* and *Calculate partial AUC.*

Examples of use: *calculate_AUC* in example 1, *calculate_pAUC* in example 2

Calculate AUC: calculates AUC for ROC analysis.

Inputs:

False positive values table (X axis of ROC curve).

True positive values table (Y axis of ROC curve) (both produced by submodel Create curve).

Output: AUC (value).

Examples of use: calculate_AUC in example 1

Calculate partial AUC: calculates partial AUC for ROC analysis.

Inputs:

False positive values table (X axis of ROC curve).

True positive values table (Y axis of ROC curve) (both produced by submodel *Create curve*).

Inferior Threshold: inferior threshold value to define the portion of the curve partial AUC is based on.

Superior Threshold: superior threshold value to define the portion of the curve partial AUC is based on.

Sensitivity Vertical Axis of the curve: if yes the thresholds defined by the user are assigned to the True positive rate (Y axis) of the ROC curve. If no, the thresholds defined by the user are assigned to the False positive rate (X axis) of the ROC curve and corresponding thresholds in the Y axis are found to calculate the partial AUC.

Outputs:

pAUC (value)

Normalized pAUC value using the equation (1) in order to present the same variation as AUC (0.5 for a non-discriminant ROC curve derived from a random probability map, one for a perfect ROC curve).

$$pAUC_{st} = \frac{1}{2}\left(\frac{pAUC - \min}{\max - \min} + 1\right) \quad (1)$$

where $pAUC_{st}$ is the standardized pAUC, min is the pAUC obtained by the random model over the same portion of the ROC curve, and max is the pAUC over the same portion of the perfect ROC curve.

Examples of use: *calculate_pAUC* in example 2

Calculate AUC 2: calculates AUC for ROC analysis along with lower and upper AUCs.

Inputs:

False positive values table (X axis of ROC curve).

True positive values table (Y axis of ROC curve) (both produced by submodel *Create curve*)

Outputs: AUC upper, AUC and AUC lower (values). Lower and upper AUC are based on stair-stepped shape curve that lies below and above the ROC trapezoidal curve respectively (see Pontius and Parmentier, submitted).

Examples of use: *calculate_AUC_lower_upper* in example 2

Bootstrap 1: Produces bootstrap replicates of a probability map.

Inputs:

Map of occurrence: Map of occurrence of an event.

Map of probability: Map of probability of an event.

Interval hist: Threshold value to bin the probability image (default 1%).

Outputs:

"Replicated" probability map

Note: in order to carry out bootstrapping, this submodel has to be inserted in a *Repeat* functor. This submodel enables user to calculate the AUC of bootstrap replicates of the probability map in order to estimate the AUC (or pAUC) confidence interval using submodel ROC CI.

Examples of use: *calculate_AUC_CI* in example 2

Bootstrap 2: Produces bootstrap replicates of two probability maps and an occurrence map. Useful to compare the AUC of two models based on the same occurrence data.

Inputs:

Map of occurrence: Map of occurrence of an event.

Map of probability: Map of probability of an event.

Map of occurrence: Map of occurrence of an event.

Outputs:

Replicate P1: "Replicated" probability map 1.

Replicate P2: "Replicated" probability map 2.

Replicate O: "Replicated" occurrence map.

Examples of use: *compare_2_models* in example 2

ROC CI: Calculates confidence interval of AUC estimate

Inputs:

Table of Bootstrap replicates: table of AUC (or pAUC) values obtained by bootstrapping.

Z scores table: Normal Distribution Table see tables z-score_pt.csv and z-score_tp.csv.

AUC: value of AUC (or pAUC) obtained by original data (that is: without bootstrapping).

Alpha error: value of alpha (type I error, 5% by default).

Studentized Percentile Bootstrap: Yes (default) the confidence bounds estimation is based on a normal distribution assumption and estimates the CI bounds using the standard deviation of the replicated AUCs and critical values of the Student's t distribution. If set to No the bootstrap percentile interval method is applied using the empirical quantiles of the bootstrap replicates to estimate the confidence interval bounds.

Outputs:

Lower bound CI: lower bound of the confidence interval

Upper bound CI: upper bound of the confidence interval

Examples of use: *calculate_AUC_CI* in example 2

Sample: resample the input probability and occurrence map in order to deal with smaller data

Inputs:

input binary map: occurrence map

Proportion of zero. Proportion (value between 0 and 1) of cells with 0 value conserved in the resampled data (0.1 or 10% by default)

Proportion of one. Proportion (value between 0 and 1) of cells with 1 value conserved in the resampled data (1 or 100% by default)

Output

Output resampled binary map. Map with resampled data only. If input maps have no data value cells these are also dropped out from the output map. Spatial structure of input maps is not conserved. in the output map that have only one line. However, this does not affect the results as operations are based on cell-to-cell operations.

Note: Using resampled maps is useful when running iterative procedures (Monte Carlo or Bootstrapping).

Examples of use: *reduce_size* in example 2*.*

Referencias / References

Fawcett, T., 2006. An introduction to ROC analysis, *Pattern Recognition Letters* 27: 861–874.

Mas, J.F., B. Soares Filho, R. G. Pontius Jr, M. Farfan Gutiérrez, H. Rodrigues, 2013, A suite of tools for ROC analysis of spatial models, *ISPRS International Journal of Geo-Information.* 2(3):869-887. http://www.mdpi.com/2220-9964/2/3/869

Peterson, A.T. Monica Papeş and Jorge Soberón, 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modelling, *Ecological Modelling* 213(1): 63-72.

Pontius Jr, Robert Gilmore and Kiran Batchu, 2003. Using the relative operating characteristic to quantify certainty in prediction of location of land cover change in India. *Transactions in GIS* 7(4): 467-484.

Pontius Jr, R.G. and Schneider, L.C., 2001. Land-cover change model validation by an ROC method for the Ipswich watershed, Massachusetts, USA. *Agriculture, Ecosystems and Environment* 85: 239-248.

Pontius Jr, R.G. and B. Parmentier, submitted, Recommendations for using the Relative Operating Characteristic (ROC), submitted to *Landscape Ecology*.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M., 2011, pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12:77. Available at http://www.biomedcentral.com/1471-2105/12/7