User Manual for SAHM package for VisTrails

Colin B. Talbert and Marian K. Talbert

February 9, 2012

Contents

User Manual For for SAHM package for VisTrails

Colin B. Talbert and Marian K. Talbert

1 Introduction

The Software for Assisted Habitat Modeling (SAHM) has been created to both expedite habitat modeling and help maintain a record of the various input data, pre- and post- processing steps and modeling options incorporated in the construction of a species distribution model. The four main advantages to using the combined VisTrail: SAHM package for species distribution modeling are:

1. formalization and tractable recording of the entire modeling process
2. easier collaboration through a common modeling framework
3. a user-friendly graphical interface to manage file input, model runs, and output
4. extensibility to incorporate future and additional modeling routines and tools.

This user manual provides detailed information on each module within the SAHM package, their input, output, common connections, optional arguments, and default settings. This information can also be accessed for individual modules by right clicking on the documentation button for any module in VisTrail or by right clicking on any input or output for a module and selecting view documentation. This user manual is intended to accompany the user guide which provides detailed instructions on how to install the SAHM package within VisTrails and then presents information on the use of the package. A step-by-step tutorial to create cheatgrass habitat suitability maps for Rocky Mountain National Park, USA, is provided in the user guide as well.

2 AggregationMethod

This module is a required class for other modules and scripts within the SAHM package. It is not intended for direct use or incorporation into the VisTrails workflow by the user.

Input Ports

- *value* (optional) NA

Output Ports

- *value_as_string* (optional) ToDo

3 BoostedRegressionTree

BRT uses decision trees to partition the the parameter space into the most homogeneous groups in terms of the response. BRT starts with a single decision tree, then adds a tree that best explains error in the first tree, and so on. Like random forest, BRT models automatically model interactions and nonlinear relationships and are robust to missing observations. Our implementation makes approximately 1,000 trees. It incorporates advanced algorithms for tuning the model settings, simplifying the model using a cross-validation technique, and for detecting important interactions between covariates. If more than 500 presence or absence records are found a random subset will be used for learning rate estimation and model simplification but all data will be used in the final model fitting step. The cross-validation step within BRT should not be confused with that provided by the Model Selection Cross Validation step. The former is used to optimize parameter values when defaults are not provided while the later is used to select models based on between model comparisons of evaluation metrics. All discussion of cross-validation related to setting parameters in the BRT argument documentation refers to the algorithm used for parameter optimization and does not affect the cross validation split selected by Model Selection and Cross Validation.

Several options are available for fitting BRTs when run using VisTrails special attention is required before moving away from the defaults because selection of certain parameters will disallow selection of others. Optional parameters are described briefly here but a more in depth description can be found in Elith and Leathwich 2008.

Input Ports

- *mdsFile* (mandatory) The the input data set consisting of locational data for each sample point, the values of each predictor variable at those points This input file is almost always generated by the upstream steps.

  Common connections

  - The mdsFile can be produced by any of MDSBuilder, ModelEvaluationSplit, ModelSelectionCrossValidation, ModelSelectionSplit, or CorariateCorrelationAndSelection.

- *makeBinMap* (optional) Indicate whether to discretize the continues probability map into presence absence. See the ThresholdOptimizationMethod for how this is done. If time is a concern and many models are to be fit and assessed maps can be produced after model selection for only the best models using the Select and Test the Final Model tool. Options are available for producing Probability, Binary and MESS maps there as well.

  Default value = False (Unchecked)

  Options

  - True (Checked)
  - False (Unchecked)

- *makeProbabilityMap* (optional) Indicate whether a map of predicted values is to be produced for the model fit.

  Default value = False (Unchecked)

  Options

  - True (Checked)
  - False (Unchecked)

- *makeMESMap* (optional) Indicate whether to produce a multivariate environmental similarity surface (MESS) and a map of which factor is limiting at each point see Elith et. al. 2010 for more details. If time is a concern and many models are to be fit and assessed maps can be produced after model selection for only the best models using the Select and Test the Final Model tool. Options are available for producing Probability, Binary and MESS maps there as well.

  Default value = False (Unchecked)

  Options

  - True (Checked)
  - False (Unchecked)

- *ThresholdOptimizationMethod* (optional) Determines how the threshold is optimized in order to discretize continuous predictions into binary. These are used for evaluation metrics calculated based on the confusion matrix as well as for the binary map. The value calculated for the train portion of the data will be applied to the test portion and if cross validation was specified, the value is calculated separately for each fold using the threshold from the training data and applying it to the test data for the hold out fold.

  Default value = 2

  Options

  - 1: Threshold=0.5

  - 2: Sens=Spec sensitivity=specificity
  - 3: MaxSens+Spec maximizes (sensitivity+specificity)/2
  - 4: MaxKappa maximizes Kappa
  - 5: MaxPCC maximizes PCC (percent correctly classified)
  - 6: PredPrev=Obs predicted prevalence=observed prevalence
  - 7: ObsPrev threshold=observed prevalence
  - 8: MeanProb mean predicted probability
  - 9: MinROCdist minimizes distance between ROC plot and (0,1)

- *Seed* (optional) The random number seed used by BRT. If one desires to reproduce results from a previous BRT fit, one must enter the random number seed that is reported in the textual output from that model fit. The seed used is always reported in the textual output.

  Default value = Randomly Generated

  Options

  - Any integer between -2147483647 and 2147483647

- *TreeComplexity* (optional) Sets the level of interactions fitted in the model. A tree complexity of 1 fits no interactions, 2 will fit up to but not necessarily all two way interactions and so on.

  Default value = If not set, tree complexity will be selected based on the number of observations and what produces the best model.

  Options

  - any positive integer (generally no greater than 3)

- *BagFraction* (optional) Controls the proportion of the data that is used to fit the model at each step. Using a bag fraction of 1 will give a fully deterministic model but this is generally not preferable as stochasticity generally improves model performance (Elith and Leathwick 2008).

  Default value = .75

  Options

  - Any positive number greater than 0 and less than or equal to 1

- *NumberOfFolds* (optional) If cross-validation is used for model simplification, this sets the number of folds used for cross-validation.

  Default value = 3

  Options

  - A positive integer (generally between 2 and 10)

- *Alpha* (optional) Controls when the algorithm stops in the model simplification step. The change in deviance is calculated between the previous and current iteration in model simplification and if the average change in deviance per observation is less than the standard error of the original deviance multiplied by alpha then the simplification step is accepted as long as we have not reached the maximum number of drops allowed.

  Default value = 1

  Options

  - Any positive floating point value is valid

- *PrevalenceStratify* (optional) This specifies whether cross validation samples should be stratified to match the overall prevalence. This is currently only valid for presence absence data and is only used in model simplification.

  Default value = True (Checked)

  Options

  - True (Checked)
  - False (Unchecked)

- *ToleranceMethod* (optional) Method used in determining when to stop model simplification.

  Default value = "auto"

  Options

  - Either “auto” or “fixed”

- *Tolerance* (optional) Can be set to control the stopping rule in model simplification. If ToleranceMethod is set to auto this value will be multiplied by the mean total deviance of the null model. Change in deviance is compared to the tolerance to determine when to stop model simplification.

  Default value = .001

  Options

  - Any positive floating point value is valid

- *LearningRate* (optional) Controls the amount each tree contributes to the model. A small learning rate restricts individual tree contributions to the overall model.

  Default value = If not specified, learning rate will be determined based on the number of trees and the tree complexity.

  Options

  - Any positive number greater than 0 and less than 1

- *MaximumTrees* (optional) The absolute upper limit on the total number of tress to fit. Setting this below 5000 will result in an error.

  Default value = 10,000

  Options

  - Any positive integer greater than 5,000

Output Ports

- *modelWorkspace* The R workspace where all internal details regarding the fitted model are stored. This is used by the Select and Test the Final Model module.

  Common connections

  - 'modelWorkspace' port of SAHMModelOutputViewerCell for viewing the aspatial model output.
  - 'modelWorkspace' port of SAHMSpatialOutpuViewerCell for viewing the spatial model output in a mini GIS.

- *BinaryMap* If specified using MakeBinaryMap=True then a surface of binary predictions is produced by discretizing the probability map based on the selected threshold. This map indicates whether one could expect each site to be occupied or unoccupied based on the model.

- *ProbabilityMap* If specified using MakeProbabilityMap=True then a surface of predicted values is produced based on the tiffs in the input .mds file and the fitted model. These can but do not always indicate the probability of finding the species at a given site. For example if model calibration is poor then these will not agree well with the true probabilities though discrimination between presence and absences might still be good.

- *ResidualsMap* Model residual plots show the spatial relationship between the model deviance residuals. Most models assume residuals will be independent thus spatial pattern in the deviance residuals can be indicative of a problem with the model fit and inference based on the fit. It can for example indicate that important predictors were not included in the model and can be compared with the spatial pattern of predictors that were not included in the model.

- *MessMap* If specified by selecting makeMESMap=True the the MESS and MoD surfaces will be produced. The MESS surface is the multivariate environment similarity surface and shows how well each point fits into the univariate ranges of the points for which the model was fit. Negative values in this map indicate that the point is out of the range of the training data.

- *MoDMap* If specified by selecting makeMESMap=TRUE the the MESS and MoD surfaces will be produced. The MoD map is related to the MESS map and indicates which variable was furthest from the range over which the model was fit for each spatial location. See Elith et. al. 2010 for details on how the MESS map calculations are performed.

- *modelEvalPlot* For binary data this will be a Receiver operating characteristic curve. Which shows the relationship between sensitivity and specificity as the threshold for discretizing continuous predictions into presence absence is varied. The threshold selected using the specified ThresholdOptimizationMethod is shown. If a model selection test training split was specified the ROC curve for this will be shown in red and if a cross-validation split was specified ROC curves for each cross-validation fold will be overlaied with box plots summarizing cross-validation results. For count data this display will show several standard plots for assessment of model residuals.

- *ResponseCurves* Model response curves show the relationship between each predictor included in the model, while holding all other predictors constant at their means, and the fitted values. MARS response curves are shown on a logit scale thus the response axis will not necessarily be bounded on the 0 to 1 interval. BRT response curves will show response surfaces for any interaction terms included in the final model along with the percent relative influence.

- *Text_Output* This file contains a summary of the model fit. The information contained here includes the number of presence observations (counts equal to or greater than 1 for count models), the number of absence points, the number of covariates that were considered by the model selection algorithm. Note all of these can differ from the numbers in the original .mds due to incomplete records being deleted, and predictors with only one unique value being removed. The random number seed is recorded if applicable which allows completely reproducible results as well as a summary of the model fit. Evaluation Statistics are reported for the data used to fit the model as well as for the test or cross-validation split if applicable. References for how to interpret most of these are ubiquitous in the literature but it is worth mentioning that interpretation of the calibration statistics is described by Pearce and Ferrier 2000 as well as Miller and Hui 1991. Most metrics reported here can also be found in related graphical displays.

- *modelCalibrationPlot* The calibration plot shows the predicted probability of occurrence plotted against the actual proportions of occurrence for each of 5 bins along the probability axis. A logistic regression model is fit to the logits of the predicted probabilities of occurrence and is shown on the plot. These plots are used to determine how reliably a model will predict if a site is occupied or unoccupied (Pearce and Ferrier 2000)

References

Bivand, R.S., Pebesma, E.J., and Gomez-Rubio, V. (2008). Applied Spatial Data Analysis with R. Springer New York, NY.

Dormann, C.F., McPherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., et al. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:60928.

Elith, J., Kearney, M., Phillips, S. (2010). The art of modeling range-shifting species. Methods Ecol Evol 1:330342

Elith, J., Leathwick, J.R. and Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77, 802813.

Miller, M.E., Hui, S.L., Tierney, W.M. (1991). Validation techniques for logistic regression models. Statistics in Medicine 10: 1213-26

Pearce, J., and S. Ferrier. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133:225245.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

4 CovariateCorrelationAndSelection

The CovariateCorrelationAndSelection view provides a breakpoint in the modeling workflow for the user to assess how well each variable explains the distribution of the sampled data points and to remove any variables that may exhibit high correlation with others.

The display shows the n variables that have the highest total number of correlations above a threshold with other predictors using the maximum of the Pearson, Spearman and Kendall coefficient. The column heading over each variable displays the number of other variables with which the environmental predictor is correlated using the user supplied threshold which defaults to .7. Radio buttons are available to limit the display and correlation calculations to any combination of presence, absence, or background points. The first column in the plot shows the relationship between the response and each predictor. Row labels indicate the maximum of the Spearman and Pearson correlation coefficient and a locally weighted smooth has been added to help distinguish the nature of the relationship.

The remaining plots make up a square with histograms for each variable displayed on the diagonal. Their respective graphical display and correlation with other variables can be found by locating the row/column intersection between each (above and below the diagonal). The scatter plot along with a locally weight smooth is shown below the diagonal. Presence records are represented by red points, absence by green, and background are yellow. Above the diagonal is the correlation coefficient between the two predictors. If Spearman or Kendall correlation coefficient is larger than the Pearson correlation coefficient then an s or k will show up in the bottom right corner of this box.

A user is provided with the opportunity to select a new set of the environmental predictor variables and Update the Covariate Correlation screen to investigate the relationships among the new variables selected. Variables with a high degree of correlation with other variables should generally be unchecked in their respective radio buttons, and will be excluded from subsequent analysis steps in the model workflow.

Multiple iterations can be run at this screen, allowing the user to investigate the relationships among the environmental predictor variables and choose the most appropriate set to be used in the subsequent modeling. When the desired set of variables has been chosen, the OK button is selected and processing will resume in the VisTrails workflow.

Input Ports

- *inputMDS* (mandatory) The file to select from. If this file contains unselected layers (0 in the second header line) these will initially appear deselected in the GUI.

  Common connections

  - The inputMDS can come from any module that outputs an MDS file. These are: MDSBuilder, ModelEvaluationSplit, ModelSelectionSplit, and ModelSelectionCrossValidation.

- *selectionName* (optional) This serves two purposes. First to uniquely identify a given selection. This unique name is used to determine if a selection has been previously made, to apply for example. And secondly to provide something that can be changed to trigger VisTrails to rerun this module even if nothing upstream has changed.

  Options

  - Any text can be used.

- *ShowGUI* (optional) This Boolean indicates whether to stop execution and display the GUI for user interaction. In some cases such as exploration you might want to make a selection in a previous run and then change this to false so that the selection will apply to subsequent runs without interrupting execution.

  Default value = True

  Options

  - True - The GUI will be shown.
  - False - The GUI will not be shown, execution will not be interrupted, but the previous selection made with the specified selectionName will be applied to the current MDS file.

- *numPlots* (optional) The number of variables to display at a time in the plot frame.

  Default value = 8

  Options

  - An integer greater than 1 and generally no greater than 12

- *minCor* (optional) The minimum correlation used to summarize the number of other variables each variable is highly correlated with.

  Default value = 0.7

  Options

  - A decimal number between 0 and 1.

- *corsWithHighest* (optional) If one desires to view only other parameters that have a correlation above the specified threshold with the parameter than has the highest number of total correlations with other parameters then this should be set to true. Otherwise, by default, the parameters that are selected for display will be the set of parameters that have the highest number of correlations with other parameters above the given threshold.

  Default value = False

  Options

  - True (Checked)
  - False (Unchecked)

Output Ports

- *outputMDS* This is the output MDS file with the user supplied selection applied.

  Common connections

  - The output from the CovariateCorrelationAndSelection will generally connect to one of the model modules (BoostedRegressionTree, GLM, MARS, RandomForest, or Maxent)
  - If using Maxent the output from CovariateCorrelationAndSelection might also connect to the RasterFormatConverter.

5 FieldData

The FieldData module allows a user to add presence/absence points or count data recorded across a landscape for the phenomenon being modeled (e.g., plant sightings, evidence of animal presence, etc.). The input data for this module must be in the form of a .csv file that follows one of two formats:

Format 1: A .csv file with the following column headings, in order: “X,” “Y,” and “responseBinary”. In this case, the “X” field should be populated with the horizontal (longitudinal) positional data for a sample point. The “Y” field should be populated with the vertical (latitudinal) data for a sample point. These values must be in the same coordinate system/units as the template layer used in the workflow. The column “responseBinary” should be populated with either a ‘0’ (indicating absence at the point) or a ‘1’ (indicating presence at the point).

Format 2: A .csv file with the following column headings, in order: “X,” “Y,” and “responseCount”. In this case, the “X” field should be populated with the horizontal (longitudinal) positional data for a sample point. The “Y” field should be populated with the vertical (latitudinal) data for a sample point. These values must be in the same coordinate system/units as the template layer used in the workflow. The column “responseCount” should be populated with either a ‘-9999’ (indicating that the point is a background point) or a numerical value (either ‘0’ or a positive integer) indicating the number of incidences of the phenomenon recorded at that point.

Output Ports

- *value* (mandatory) This is the actual file object that is being passed to other modules in the workflow.

  Common connections

  - The ‘fieldData_file’ input port of the FieldDataQuery Module if the field data needs subsetting or aggregation.
  - The ‘fieldData’ input port of the FieldDataAggregateAndWeight

  - Module if the field data needs to be aggregated or weighted to match the spatial resolution of the template layer.
  - The 'fieldData' input port of the MDS builder Module if the field data needs no further pre- processing prior to modeling.

- *value_as_string* (optional) This is a VisTrails port that is not used in general SAHM workflows.

  Common connections

  - This does not commonly connect to other SAHM modules.

6 FieldDataAggregateAndWeight

In many instances data collected in the field can be at a different projection and spatial resolution than we are modeling at. For example we might have observations collected every five meters along a 200 m. transect when we are modeling with covariates with 1000 m. cells. When running species distribution models (SDMs) such as those contained in SAHM, spatial issues need to be addressed in order to avoid introduction of pseudo-replication. For instance, considering multiple field data observations which are all spatially located in the same modeled pixel will generate replicate values or redundant information. When running a model, this redundancy causes pseudo-replication and can negatively influence model development. The FieldDataAggregateAndWeight tool helps aggregate field data locations so only one field data observation is represented per pixel or multiple points are down-weighted proportionately. Additionally the FieldDataAggregationAndWeight module allows the user to change the datum / projection system of the FieldData x, y locations to match that used in the template.

Currently only GLM, MARS, and Boosted Regression Trees accept weights. Any Weights column will be ignored by Random Forest.

Input Ports

- *templateLayer* (mandatory) Raster file used to determine cell size and extent. Note - The projection and coordinate system used in the template file must match that given in the FieldData‘s X and Y columns.

  Common connections

  – This input port generally will connect to the ’value‘ port of a TemplateLayer Module.

- *fieldData* (mandatory) The file containing field data. Must be in X, Y, ResponseBinary/ResponseCount format

  Common connections

  – The’value‘ port of a FieldData module

  – The ’fieldData‘ port of the FieldDataQuery module

- *PointAggregationOrWeightMethod* (mandatory) The method used to either weight or aggregate field data points.

  Default value = Collapse In Pixel

  Options

  – Collapse In Pixel : All field data points falling within a single pixel will be collapse into a single point at the center of that pixel. If using Presence(Absence) data the point will be given a value of 1 if any are presense, otherwise 0 if any are absence, otherwise -9999 if all are background. If using count data the point value will be the average of all points in a pixel.

  – Weight Per Pixel : All field data points are retained but a weight column is added and points in pixels with multiple points are given a weight of 1 over the number of points in that pixel. For example all the points in a pixel with 4 points would be given a weight of 1/4.

- *FD_EPSG_projection* (optional) This optional parameter is a means of specifying the datum and projection information that the field data X and Y locations are in. If this parameter is supplied the point locations will be transformed to the datum projection of the template layer. Otherwise it will be assumed that the points and template are in the same projection and datum. The value to enter in this port must be a valid EPSG code, see below.

EPSG codes are numbers representing all commonly used geographic and projected coordinate systems. These are generally 4 to 6 digit numbers. For example 4326 represents geographic WGS84 data, 4260 represents geographic NAD83, 26912 represents NAD83 / UTM zone 12N, etc. See the options below for a list of ways to lookup EPSG codes.

Default value = None, This assumes that the points are in the same coordinate system as your template layer.

Options

- Within your GDAL installation directory there is a gdal-data directory with two csv files which list all of the supported EPSG numbers and the name and information of the coordinate system. Use gcs.csv for geographic systems and pcs.csv for projected systems.
- EPSG codes can also be looked up at: http://spatialreference.org/ref/
- If you have a .prj file or well known text (WKT) for your coordinate system you can find the EPSG using: http://prj2epsg.org/search
- If you have ESRI ArcGIS and a layer in the coordinate system you can find the EPSG code in the items metadata. Select the item in ArcCatalog, open the description tag, scroll down to the Spatial Reference section, the EPSG will be the number in the 'WELL-KNOWN-IDENDIFIER' tag.

Output Ports

- *fieldData* (mandatory) This is a CSV file in a X,Y,Response,(Weight) format.

  Common connections

– The input port ’fieldData‘ of the MDSBuilder module.

7 FieldDataQuery

Often raw field data come to us in a format that contains more information than we need to include in any single model. This can take the form of additional columns that contain extraneous information, additional columns that contain occurrence data for additional species, or rows that from a time period, collection method, or species that we are not interested in modeling. The Field Data Query module contains functionality to subset and reformat this output into the format used by the SAHM package. Columns can be specified with either a positional argument (1, 2, 3, etc) if you want to select the first, second, third etc column. Note these numbers start from 1. Alternatively you can select a column based on name by entering the text of the column name found in the header.

When selecting rows there are two types of queries that can be specified: Simple - Select a Query_Column and enter a value in the Query port. If the value for a row in the selected column equals the value entered in the Query that row will be kept. For example you might have a ’year‘ column and you would want to select all 2009 entries. NOTE: DO NOT ENCLOSE THE QUERY TEXT IN QUOTES IF YOU ARE TRYING TO MATCH A STRING!

Complex - Optionally you can construct complex queries using Python syntax. To do this enclose the column name in square brackets as part of a line of Python code. Since the columns used are specified in the query string there is no need to use the Query_column port and it will be ignored. For example to select years greater than 2005 you would use: [year] ¿ 2005 To include string equality make sure you enclose the entire bracketed field name in quotes as well. For example “[Observer]” == “Colin” Complex queries involving multiple columns are possible as well, for example “[Observer]” != “Colin” and [year] ¿ 2005.

Input Ports

- *fieldData_file* (mandatory) The file containing Field data. The acceptable formats vary but it must have a column with X, Y, and response values. Additional columns are permissible.

  Common connections

  – This port can be connected to a FieldData module or the FieldData file can be specified directly in the module information pane.

- *x_column* (optional) The column that contains the 'X' coordinates. These values must be in the same coordinates, projection, and units as those defined in the template layer.

  Columns can be specified with either a positional argument (1, 2, 3, etc) if you want to select the first, second, third etc column. Note these numbers start from 1. Alternatively you can select a column based on name by entering the text of the column name found in the header.

  Default value = 1, which is to say the first column in the input field data file.

  Common connections

  – NA

- *y_column* (optional) The column that contains the 'Y' coordinates. These values must be in the same coordinates, projection, and units as those defined in the template layer.

  Columns can be specified with either a positional argument (1, 2, 3, etc) if you want to select the first, second, third etc column. Note these numbers start from 1. Alternatively you can select a column based on name by entering the text of the column name found in the header.

  Default value = 2, which is to say the second column in the input field data file.

- *Response_column* (optional) The column that contains the response of interest.

  Columns can be specified with either a positional argument (1, 2, 3, etc) if you want to select the first, second, third etc column. Note these numbers start from 1. Alternatively you can select a column based on name by entering the text of the column name found in the header.

Default value = 3, which is to say the third column in the input field data file.

- *ResponseType* (optional) The type of response recorded in the response column

  Default value = 'Presence(Absence)'

  Options

  - 'Presence(Absence)' = 1 for Presence, optionally also 0 for Absence and -9999 for background points.
  - 'Count' = 0, 1, 2, 3 etc. observed count data. Optionally also -9999 for background points

- *Response_Presence_value* (optional) The value in the response column that will be taken to indicate a presence.

  Default value = By default any value in the list '1', 'True', 'T', 'Present', 'Presence' will be assigned a value of 1 (presence) in the output. Note: not case sensitive.

  Options

  - And number or string can be entered, quotes are not required.

- *Response_Absence_value* (optional) The value in the response column that will be taken to indicate an absence.

  Default value = By default any value in the list '0', 'False', 'F', 'Absent', 'Absence' will be assigned a value of 0 (absence) in the output. Note: not case sensitive.

  Options

  - And number or string can be entered, quotes are not required.

- *Query_column* (optional) The column which contains values you would like to use to with the simple equality query option. The values in this column will be checked against the value entered in the query port.

- *Query* (optional) If using the simple equality query functionality simply enter the value you would like to filter on here. NOTE: DO NOT ENCLOSE THE QUERY TEXT IN QUOTES IF YOU ARE TRYING TO MATCH A STRING! Also do not include any additional spaces.

  If using the complex Python syntax query a valid Python equality statement with the values from each individual row indicated with square bracketed field (header row) names.

  Options

  - Simple - Select a Query_Column and enter a value in the Query port. If the value for a row in the selected column equals the value entered in the Query that row will be kept. For example you might have a 'year' column selected in the Query_column port and enter 2009 here to to select all 2009 entries. NOTE: DO NOT ENCLOSE THE QUERY TEXT IN QUOTES IF YOU ARE TRYING TO MATCH A STRING!
  - Complex - Optionally you can construct complex queries using Python syntax. To do this enclose the column name in square brackets as part of a line of Python code. Since the columns used are specified in the query string there is no need to use the Query_column port and it will be ignored. For example to select years greater than 2005 you would use: [year] ¿ 2005 To include string equality make sure you enclose the entire bracketed field name in quotes as well. For example “[Observer]” == “Colin” Complex queries involving multiple columns are possible as well, for example “[Observer]” != “Colin” and [year] ¿ 2005.

Output Ports

- *fieldData* (mandatory) The file containing field data. This output file will be in: X, Y, ResponseBinary/ResponseCount format.

  Common connections

  - FieldDataAggregateAndWeight FieldData - For collapsing or weighting points relative to the template pixels.

  - MDS_builder - fieldData - for modeling without using the FieldDataAggregateAndWeight functionality.

8 GLM

This is basically linear regression adapted to binary presence-absence or count data. We used a bidirectional stepwise procedure to select covariates to be used in the model. That is, we began with a null model and calculated the AIC (Akaike Information Criterion) score for each covariate which could be added to the model. AIC is a measure of how well the model fits the data with a penalty based on the number of covariates in the model. In the first step, we add the covariate with the best AIC score. In the next step we calculate AIC scores for all two- covariate models and again add the covariate that most improves the AIC, and so on. At each step, we also look at the change in AIC from dropping each covariate currently in the model. The stepwise procedure ends when no additions or removals result in an improvement in AIC.

Input Ports

- *mdsFile* (mandatory) The the input data set consisting of locational data for each sample point, the values of each predictor variable at those points This input file is almost always generated by the upstream steps.

  Common connections

  - The mdsFile can be produced by any of MDSBuilder, ModelEvaluationSplit, ModelSelectionCrossValidation, ModelSelectionSplit, or CovariateCorrelationAndSelection.

- *makeBinMap* (optional) Indicate whether to discretize the continues probability map into presence absence. See the ThresholdOptimizationMethod for how this is done. If time is a concern and many models are to be fit and assessed maps can be produced after model selection for only the best models using the Select and Test the Final Model tool. Options are available for producing Probability, Binary and MESS maps there as well.

Default value = False (Unchecked)

Options

  - True (Checked)
  - False (Unchecked)

- *makeProbabilityMap* (optional) Indicate whether a map of predicted values is to be produced for the model fit.

  Default value = False (Unchecked)

  Options

  - True (Checked)
  - False (Unchecked)

- *makeMESMap* (optional) Indicate whether to produce a multivariate environmental similarity surface (MESS) and a map of which factor is limiting at each point see Elith et. al. 2010 for more details. If time is a concern and many models are to be fit and assessed maps can be produced after model selection for only the best models using the Select and Test the Final Model tool. Options are available for producing Probability, Binary and MESS maps there as well.

  Default value = False (Unchecked)

  Options

  - True (Checked)
  - False (Unchecked)

- *ThresholdOptimizationMethod* (optional) Determines how the threshold is optimized in order to discretize continuous predictions into binary. These are used for evaluation metrics calculated based on the confusion matrix as well as for the binary map. The value calculated for the train portion of the data will be applied to the test portion and if cross validation was specified, the value is calculated separately for each fold using the threshold from the training data and applying it to the test data for the hold out fold.

Default value = 2

Options

  - 1: Threshold=0.5
  - 2: Sens=Spec sensitivity=specificity
  - 3: MaxSens+Spec maximizes (sensitivity+specificity)/2
  - 4: MaxKappa maximizes Kappa
  - 5: MaxPCC maximizes PCC (percent correctly classified)
  - 6: PredPrev=Obs predicted prevalence=observed prevalence
  - 7: ObsPrev threshold=observed prevalence
  - 8: MeanProb mean predicted probability
  - 9: MinROCdist minimizes distance between ROC plot and (0,1)

- *SimplificationMethod* (optional) This alters the decision rule governing how the model is pruned in the stepwise model selection step.

  Default value = AIC

  Options

  - AIC or BIC

Output Ports

- *modelWorkspace* The R workspace where all internal details regarding the fitted model are stored. This is used by the Select and Test the Final Model module.

  Common connections

  - 'modelWorkspace‘ port of SAHMModelOutputViewerCell for viewing the aspatial model output.
  - 'modelWorkspace‘ port of SAHMSpatialOutpuViewerCell for viewing the spatial model output in a mini GIS.

- *BinaryMap* If specified using MakeBinaryMap=True then a surface of binary predictions is produced by discretizing the probability map based on the selected threshold. This map indicates whether one could expect each site to be occupied or unoccupied based on the model.

- *ProbabilityMap* If specified using MakeProbabilityMap=True then a surface of predicted values is produced based on the tiffs in the input .mds file and the fitted model. These can but do not always indicate the probability of finding the species at a given site. For example if model calibration is poor then these will not agree well with the true probabilities though discrimination between presence and absences might still be good.

- *ResidualsMap* Model residual plots show the spatial relationship between the model deviance residuals. Most models assume residuals will be independent thus spatial pattern in the deviance residuals can be indicative of a problem with the model fit and inference based on the fit. It can for example indicate that important predictors were not included in the model and can be compared with the spatial pattern of predictors that were not included in the model.

- *MessMap* If specified by selecting makeMESMap=True the the MESS and MoD surfaces will be produced. The MESS surface is the multivariate environment similarity surface and shows how well each point fits into the univariate ranges of the points for which the model was fit. Negative values in this map indicate that the point is out of the range of the training data.

- *MoDMap* If specified by selecting makeMESMap=TRUE the the MESS and MoD surfaces will be produced. The MoD map is related to the MESS map and indicates which variable was furthest from the range over which the model was fit for each spatial location. See Elith et. al. 2010 for details on how the MESS map calculations are performed.

- *modelEvalPlot* For binary data this will be a Receiver operating characteristic curve. Which shows the relationship between sensitivity and specificity as the threshold for discretizing continuous predictions into presence absence is varied. The threshold selected using the specified ThresholdOptimizationMethod is shown. If a model selection test training split was specified the ROC curve for this will be shown in red and if a cross-validation split was specified ROC curves for each cross-validation fold will be overlaied with box plots summarizing cross-validation results. For count data this display will show several standard plots for assessment of model residuals.

- *ResponseCurves* Model response curves show the relationship between each predictor included in the model, while holding all other predictors constant at their means, and the fitted values. MARS response curves are shown on a logit scale thus the response axis will not necessarily be bounded on the 0 to 1 interval. BRT response curves will show response surfaces for any interaction terms included in the final model along with the percent relative influence.

- *Text_Output* This file contains a summary of the model fit. The information contained here includes the number of presence observations (counts equal to or greater than 1 for count models), the number of absence points, the number of covariates that were considered by the model selection algorithm. Note all of these can differ from the numbers in the original .mds due to incomplete records being deleted, and predictors with only one unique value being removed. The random number seed is recorded if applicable which allows completely reproducible results as well as a summary of the model fit. Evaluation Statistics are reported for the data used to fit the model as well as for the test or cross-validation split if applicable. References for how to interpret most of these are ubiquitous in the literature but it is worth mentioning that interpretation of the calibration statistics is described by Pearce and Ferrier 2000 as well as Miller and Hui 1991. Most metrics reported here can also be found in related graphical displays.

- *modelCalibrationPlot* The calibration plot shows the predicted probability of occurrence plotted against the actual proportions of occurrence for each of 5 bins along the probability axis. A logistic regression model is fit to the logits of the predicted probabilities of occurrence and is shown on the plot. These plots are used to determine how reliably a model will predict if a site is occupied or unoccupied (Pearce and Ferrier 2000)

References

Bivand, R.S., Pebesma, E.J., and Gomez-Rubio, V. (2008). Applied Spatial Data Analysis with R. Springer New York, NY.

Dormann, C.F., McPherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., et al. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:60928.

Elith, J., Kearney, M., Phillips, S. (2010). The art of modeling range-shifting species. Methods Ecol Evol 1:330342

Miller, M.E., Hui, S.L., Tierney, W.M. (1991). Validation techniques for logistic regression models. Statistics in Medicine 10: 1213-26

Pearce, J., and S. Ferrier. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133:225245.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

9 MARS

MARS is a non-parametric technique that builds flexible models by fitting piecewise logistic regressions. In effect, it is similar to GLM except that rather than fitting a straight line response to each predictor, piecewise functions of each predictor are fit, which allows MARS to better accommodate nonlinear response to predictors and also reduces the risk that outlying observations might have high leverage. The model is deliberately over-fit and then pruned back. The original code was developed from that provided in the supporting material of Leathwick and Elith 2006 which contains more details on how model fitting occurs.

Input Ports

- *mdsFile* (mandatory) The the input data set consisting of locational data for each sample point, the values of each predictor variable at those points This input file is almost always generated by the upstream steps.

  Common connections

  - The mdsFile can be produced by any of MDSBuilder, ModelEvaluationSplit, ModelSelectionCrossValidation, ModelSelectionSplit, or CovariateCorrelationAndSelection.

- *makeBinMap* (optional) Indicate whether to discretize the continues probability map into presence absence. See the ThresholdOptimizationMethod for how this is done. If time is a concern and many models are to be fit and assessed maps can be produced after model selection for only the best models using the Select and Test the Final Model tool. Options are available for producing Probability, Binary and MESS maps there as well.

  Default value = False (Unchecked)

  Options

  - True (Checked)
  - False (Unchecked)

- *makeProbabilityMap* (optional) Indicate whether a map of predicted values is to be produced for the model fit.

  Default value = False (Unchecked)

  Options

  - True (Checked)
  - False (Unchecked)

- *makeMESMap* (optional) Indicate whether to produce a multivariate environmental similarity surface (MESS) and a map of which factor is limiting at each point see Elith et. al. 2010 for more details. If time is a concern and many models are to be fit and assessed maps can be produced after model selection for only the best models using the Select and Test the Final Model tool. Options are available for producing Probability, Binary and MESS maps there as well.

Default value = False (Unchecked)

Options

- True (Checked)
- False (Unchecked)

- *ThresholdOptimizationMethod* (optional) Determines how the threshold is optimized in order to discretize continuous predictions into binary. These are used for evaluation metrics calculated based on the confusion matrix as well as for the binary map. The value calculated for the train portion of the data will be applied to the test portion and if cross validation was specified, the value is calculated separately for each fold using the threshold from the training data and applying it to the test data for the hold out fold.

  Default value = 2

  Options

  - 1: Threshold=0.5
  - 2: Sens=Spec sensitivity=specificity
  - 3: MaxSens+Spec maximizes (sensitivity+specificity)/2
  - 4: MaxKappa maximizes Kappa
  - 5: MaxPCC maximizes PCC (percent correctly classified)
  - 6: PredPrev=Obs predicted prevalence=observed prevalence
  - 7: ObsPrev threshold=observed prevalence
  - 8: MeanProb mean predicted probability
  - 9: MinROCdist minimizes distance between ROC plot and (0,1)

- *MarsDegree* (optional) The level of interaction allowed: 1=no interactions (default) terms are allowed in the model 2=1st order interactions 3=2nd order interactions and so on.

  Default value = 1

  Options

  - A positive integer generally no greater than 3 or possibly 4

- *MarsPenalty* (optional) The cost per degree of freedom charge in fitting the mars model (from the mda library).

  Default value = 2

  Options

  - A positive float

Output Ports

- *modelWorkspace* The R workspace where all internal details regarding the fitted model are stored. This is used by the Select and Test the Final Model module.

  Common connections

  - 'modelWorkspace' port of SAHMModelOutputViewerCell for viewing the aspatial model output.
  - 'modelWorkspace' port of SAHMSpatialOutpuViewerCell for viewing the spatial model output in a mini GIS.

- *BinaryMap* If specified using MakeBinaryMap=True then a surface of binary predictions is produced by discretizing the probability map based on the selected threshold. This map indicates whether one could expect each site to be occupied or unoccupied based on the model.

- *ProbabilityMap* If specified using MakeProbabilityMap=True then a surface of predicted values is produced based on the tiffs in the input .mds file and the fitted model. These can but do not always indicate the probability of finding the species at a given site. For example if model calibration is poor then these will not agree well with the true probabilities though discrimination between presence and absences might still be good.

- *ResidualsMap* Model residual plots show the spatial relationship between the model deviance residuals. Most models assume residuals will be independent thus spatial pattern in the deviance residuals can be indicative of a problem with the model fit and inference based on the fit. It can for example indicate that important predictors were not included in the model and can be compared with the spatial pattern of predictors that were not included in the model.

- *MessMap* If specified by selecting makeMESMap=True the the MESS and MoD surfaces will be produced. The MESS surface is the multivariate environment similarity surface and shows how well each point fits into the univariate ranges of the points for which the model was fit. Negative values in this map indicate that the point is out of the range of the training data.

- *MoDMap* If specified by selecting makeMESMap=TRUE the the MESS and MoD surfaces will be produced. The MoD map is related to the MESS map and indicates which variable was furthest from the range over which the model was fit for each spatial location. See Elith et. al. 2010 for details on how the MESS map calculations are performed.

- *modelEvalPlot* For binary data this will be a Receiver operating characteristic curve. Which shows the relationship between sensitivity and specificity as the threshold for discretizing continuous predictions into presence absence is varied. The threshold selected using the specified ThresholdOptimizationMethod is shown. If a model selection test training split was specified the ROC curve for this will be shown in red and if a cross-validation split was specified ROC curves for each cross-validation fold will be overlaied with box plots summarizing cross-validation results. For count data this display will show several standard plots for assessment of model residuals.

- *ResponseCurves* Model response curves show the relationship between each predictor included in the model, while holding all other predictors constant at their means, and the fitted values. MARS response curves are shown on a logit scale thus the response axis will not necessarily be bounded on the 0 to 1 interval. BRT response curves will show response surfaces for any interaction terms included in the final model along with the percent relative influence.

- *Text_Output* This file contains a summary of the model fit. The information contained here includes the number of presence observations (counts equal to or greater than 1 for count models), the number of absence points, the number of covariates that were considered by the model selection algorithm. Note all of these can differ from the numbers in the original .mds due to incomplete records being deleted, and predictors with only one unique value being removed. The random number seed is recorded if applicable which allows completely reproducible results as well as a summary of the model fit. Evaluation Statistics are reported for the data used to fit the model as well as for the test or cross-validation split if applicable. References for how to interpret most of these are ubiquitous in the literature but it is worth mentioning that interpretation of the calibration statistics is described by Pearce and Ferrier 2000 as well as Miller and Hui 1991. Most metrics reported here can also be found in related graphical displays.

- *modelCalibrationPlot* The calibration plot shows the predicted probability of occurrence plotted against the actual proportions of occurrence for each of 5 bins along the probability axis. A logistic regression model is fit to the logits of the predicted probabilities of occurrence and is shown on the plot. These plots are used to determine how reliably a model will predict if a site is occupied or unoccupied (Pearce and Ferrier 2000)

References

Bivand, R.S., Pebesma, E.J., and Gomez-Rubio, V. (2008). Applied Spatial Data Analysis with R. Springer New York, NY.

Dormann, C.F., McPherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., et al. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:60928.

Elith, J., Kearney, M., Phillips, S. (2010). The art of modeling range-shifting species. Methods Ecol Evol 1:330342

Hastie, T. and Tibshirani., R. mda: Mixture and flexible discriminant analysis. Ported to R by Leisch, F., Hornik, K. and Ripley B. D. (2011). R package version 0.4-2.

Leathwick J.R., Elith, J., Hastie, T. (2006). Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling 199: 188-96

Miller, M.E., Hui, S.L., Tierney, W.M. (1991). Validation techniques for logistic regression models. Statistics in Medicine 10: 1213-26

Pearce, J., and S. Ferrier. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133:225245.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

10 MAXENT

11 MDSBuilder

The Merged Data Set (MDS) Builder module is a utility that extracts the values of each predictor layer to the point locations included in the field data set. The module produces a .csv file that contains the x and y locations of the sample points and a column indicating whether each point represents a presence recording, an absence recording, a presence count, or a background point. Following these first three columns, each environmental predictor layer is appended as a column with row entries representing the value present in the raster layer at each field sample point. There are a total of three header rows in the output .csv of the MDSBuilder. The first row contains the columns "x," "y," "ResponseBinary" or "ResponseCount," and the names of each of the raster predictor files that were passed to the MDS Builder. The second row contains a binary value indicating whether the column should be included when the model is finally applied; these values are later modified during the Covariate Correlation and Selection process that takes place downstream in the workflow. The final header row contains the full path on the file system to each of the raster predictor files.

The output from this module is in the format expected by most of the pre-modeling data manipulation modules, all of the model modules, as well as the RasterFormatConverter module. As such it can reasonably be connected to numerous other modules depending on the type of modeling being conducted. A typical workflow would linearly connect MDSBuilder -¿ ModelEvaluationSplit -¿ ModelSelectionSplit or ModelSelectionCrossValidation -¿ CovariateCorrelationAndSelection -¿ any or all of the models (BoostedRegressionTree, GLM, MARS, RandomForest, Maxent). If using Maxent the output from CovariateCorrelationAndSelection would also go into the RasterFormatConverter module which would connect to the projectionlayers of the maxent module.

Input Ports

- *RastersWithPARCInfoCSV* (mandatory) This is a csv file which contains information about all of the predictors used. The user will not generally need to create or edit this file as it is an output of the PARC module.

  The following columns are in a RastersWithPARCInfoCSV: PARCOutputFile - The raster file produced by PARC Categorical - 0=not categorical data, 1=categorical data Resampling - One of NearestNeighbor, bilinear, cubic, cubicspline, or lanczos Aggregation - One of min, max, mean, or majority OriginalFile - The location and name of the input file used by PARC

  Common connections

  - This port will generally connect with the output from the PARC module.

- *fieldData* (mandatory) The field data input corresponds to a .csv file containing presence/absence points or count data recorded across a landscape for the phenomenon being modeled (e.g., plant sightings, evidence of animal presence, etc.).

Common connections

  - The output port of the FieldData Module
  - The output port of the FieldDataQuery Module
  - The output port of the FieldDataAggregationAndWeight Module

- *backgroundPointCount* (optional) This is an optional value that specifies how many randomly placed background points to add to the output. These points will be randomly placed at pixel centroids within the template extent with no more than one point assigned to any one pixel. In typical SAHM workflows these points are only used by the Maxent modeling package. These points will be added to the output .csv file with a value of "-9999" denoting them as background points.

  Default value = 0, which is to say that no background point are added to the output.

- *backgroundProbSurf* (optional) Background Probability Surface: This is an optional parameter that applies only to workflows that employ the Maxent modeling package. In some analysis, it may be appropriate to spatially limit background points to a particular subset of the study area (e.g., islands within a study area polygon, particular regions within a study area polygon, or a region determined by the known bias present in the field data). Specifying a background probability surface raster allows a user to control where random points will be scattered within the extent of the study area. The raster layer specified by a user should have the same projection and extent as the template layer and contain values ranging from 0 to 100. These values represent the probability that a randomly generated point will be retained should it fall within a particular cell. That is, randomly generated points will not be generated in any part of the probability grid with a value of "0" while all points falling in an area with a value of "100" will be retained. A point falling in an area with a value of "50" will be kept as a background point 50% of the time.

  Default value = 0 (No background points are added to the output.)

  Options

  - Any positive integer.

- *Seed* (optional) The seed is used to be able recreate a specific output. The seed used in each run will be noted on the console output and saved in the output log in the session folder.

  Default value = If no seed is specified a random seed between -1*(($2^3$2)/2) and (($2^3$2)/2) will be generated and used.

  Options

  - Any integer between -1*(($2^3$2)/2) and (($2^3$2)/2)

Output Ports

- *mdsFile* (optional) This is the CSV flat file containing the location data and values extracted from each of the covariates.

  Common connections

  - The input port ’InputMDS‘ in the ModelEvaluationSplit module.
  - The input port ’InputMDS‘ in the ModelSelectionSplit module.
  - The input port ’InputMDS‘ in the ModelSelectionCrossValidation module.
  - The input port ’InputMDS‘ in the CovariateCorrelationAndSelection module.
  - The input port ’InputMDS‘ in the RasterFormatConverter module.
  - The input port ’InputMDS‘ in the Maxent module.
  - The input port ’InputMDS‘ in the BoostedRegressionTree module.
  - The input port ’InputMDS‘ in the GLM module.
  - The input port ’InputMDS‘ in the MARS module.
  - The input port ’InputMDS‘ in the RandomForest module.

12 MergedDataSet

This module is a required class for other modules and scripts within the SAHM package. It is not intended for direct use or incorporation into the VisTrails workflow by the user.

Input Ports

- *mdsFile* (optional) NA

Output Ports

- *value* (optional) ToDo

13 Model

This module is a required class for other modules and scripts within the SAHM package. It is not intended for direct use or incorporation into the VisTrails workflow by the user.

Input Ports

- *mdsFile* (optional) NA
- *makeBinMap* (optional) NA
- *makeProbabilityMap* (optional) NA
- *makeMESMap* (optional) NA
- *ThresholdOptimizationMethod* (optional) NA

Output Ports

- *modelWorkspace* (optional) ToDo
- *BinaryMap* (optional) ToDo
- *ProbabilityMap* (optional) ToDo
- *ResidualsMap* (optional) ToDo
- *MessMap* (optional) ToDo
- *MoDMap* (optional) ToDo
- *modelEvalPlot* (optional) ToDo
- *ResponseCurves* (optional) ToDo
- *Text_Output* (optional) ToDo

14 ModelEvaluationSplit

The ModelEvaluationSplit module provides the opportunity to reserve a specified portion of the data for producing and reporting evaluation metrics on an independent test set following model exploration and selection. The ModelEvaluationSplit must be applied before the CovariateCorrelationAndSelection module. The nearly identical ModelSelectionSplit reserves a portion of the data from the model fitting process but reports the evaluation metrics on all models not just the those selected as the final models to be reported in the analysis. This module can be placed either directly before or directly after the CovariateCorrelationAndSelection. If both a ModelEvaluationSplit and a ModelSelectionSplit are specified then the training portion of the ModelEvalutationSplit will be further partitioned by the ModelSelectionSplit thus the ModelEvalutationSplit should come first in the workflow. Both of these algorithms stratify the splits by the response. That is, the ratio of presence to absence points should be nearly equal in the testing and training split. If a ModelSelectionSplit is included evaluation metrics applied to the reserved data will be reported in the textual output, model evaluation plots including AUC plots as well as the across model plots and the csv. Both of these modules ignore background points and treat all observations with values greater than 0 as presence for the purpose of stratification by response.

Input Ports

- *inputMDS* (mandatory) This is the input data set consisting of location data for each sample point, the values of each predictor variable at those points. This input is usually provided by the upstream steps that precede the Test Training Split module. Any value entered here (e.g., specifying another existing MDS on the file system) will override the input specified by a model connection in the visual display.

  Common connections

  - This port can connect to the output port on MDSBuilder.
  - While it could technically also connect to the output from ModelSelectionCrossValidation, CovariateCorrelationAndSelection, ModelSelectionSplit, or another ModelEvaluationSplit this would not make sense for SAHM workflows. The results of connecting to one of these modules has not been tested and could cause errors or subtle but significant problems with subsequent modeling.

- *trainingProportion* (optional) This is the proportion of the sample points that will be used to train the model, relative to the total number of points. Entered values should be greater than 0 but less than 1. For example, a value of '0.9‘ will result in 90% of the sample points being used to train the model, with 10% of the sample being held out to test the model's performance. Choosing an appropriate training proportion can depend on various factors, such as the total number of sample points available. Selecting an appropriate value for the training proportion is a complex issue that depends on many factors including the total number of observations, the complexity of the models that will be fit, and the signal to noise ratio in the data (Hastie et. al. 2009).

- *RatioPresAbs* (optional) This optional field is populated with a number corresponding to the desired ratio of presence to absence points to be used in the analysis. If not populated then all occurrence records (not background points) will be portioned into either the test or training split with no reduction in the total number of points. If populated, this entry should be a number greater than zero. (A value of '1' will result in an equal number of both presence and absence points being used, a value of '2' indicates that twice as many presence points will be used, a value of 0.5' indicates that twice as many absence points will be used, etc.). All field data points with a value equal to or greater than 1 are interpreted as presence points. Although the original field data is unmodified, this option will reduce the sample size as the merged dataset containing sample points will have points deleted from it to achieve the specified ratio as such it should be used with caution. A warning will be generated if more than 50% of either the presence or absence points will be deleted based on the ratio specified by the user. Background points are ignored by this module (they are read in and written out, but not assigned to either the test or training split).

- *Seed* (optional) The random number seed used to split the data. If one desires to reproduce results from a previous split fit, one must enter the random number seed that is reported in the console output from that split step.

  Default value = Randomly selected

  Options

  - Any integer between -2147483647 and 2147483647

Output Ports

- *outputMDS* This is an MDS file to be further used in the downstream workflow. It is nearly identical to the input MDS file but with an added column under the header "EvalSplit" for each non- background observation will be labeled either "test", "train", or "NA" indicating whether the observation will be withheld for producing final model evaluation metrics, training the model, or excluded from the analysis respectively.

  Common connections

  - This port can connect to either the CovariateCorrelationAndSelection, ModelSelectionSplit, ModelSelectionCrossValidation, or directly to any of the R model modules (BoostedRegressionTree, GLM, MARS, or RandomForest).
  - While it could technically also connect to the input port of ModelEvaluationSplit this would not make sense for SAHM workflows. The results of connecting to one of these modules has not been tested and could cause errors or subtle but significant problems with subsequent modeling.

References

Hastie T, Tibshirani R, Friedman JH. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag. 744 pp. 2nd ed.

15 ModelOutputType

This module is a required class for other modules and scripts within the SAHM package. It is not intended for direct use or incorporation into the VisTrails workflow by the user.

Input Ports

- *value* (optional) NA

Output Ports

- *value_as_string* (optional) ToDo

16 ModelSelectionCrossValidation

The ModelSelectionCrossValidation module provides another tool for model selection by splitting the field data observations into cross validation folds. This should not be used with the ModelSelectionSplit but can be used with the ModelEvaluationSplit in which case only the training portion of the ModelEvalutationSplit is partitioned into folds. If specified then the individual models will fit a model using all of the data and report this as the training results. Following the model fitting step sub-models with be fit to each set of n-1 folds and then evaluation metrics calculated on the remaining fold. These will show up as ranges in the AUC plot, means and standard deviations are reported in textual output and box plots in across model comparison plots. Evaluation metrics for each individual fold are reported in the across model comparison csv. The cross validation method incorporated here was originally written for evaluation of MARS models by Leathwick et. al. 2006. The current implementation does not attempt any sort of model averaging but rather is only used for calculation of evaluation metrics. The ModelSelectionCrossValidation module makes better use of data then the ModelSelectionSplit as it uses all of the data to fit the final model but can be substantially more time consuming.

Under most circumstances the cross validation evaluation metrics reported by this module do not indicate how the the model might perform if applied to an independent set of data but rather are to be used only for model selection purposes. The first issue is that when cross validation is applied any feature selection based on the relationship between the response and the predictors must be carried out on each cross validation training set. The CovariateCorrelationAndSelection module includes an exploration of the relationship between the predictors and the response and thus would need to be carried out for each for each cross validation training set. The second issue is that it is invalid to use an evaluation metric for model selection and then report that metric for only the best performing model without acknowledgment to the total number of models that were considered and the range of the evaluation metrics. This module ignores background points.

Input Ports

- *inputMDS* (mandatory) This is the input data set consisting of locational data for each sample point, the values of each predictor variable at those points. This input is usually provided by the upstream steps that precede the Test Training Split module. Any value entered here (e.g., specifying another existing MDS on the file system) will override the input specified by a model connection in the visual display.

  Common connections

  – This port can connect to the output port on MDSBuilder, ModelEvaluationSplit

  – While it could technically also connect to the output from another ModelSelectionCrossValidation, CovariateCorrelationAndSelection, or ModelSelectionSplit this would not make sense for SAHM workflows. The results of connecting to one of these modules has not been tested and could cause errors or subtle but significant problems with subsequent modeling.

- *nFolds* (optional) The number of folds into which the data should be partitioned. A trade-off exists in selecting the number of folds to use for cross validation. When nFolds is close to the total number of observations the prediction error is nearly unbiased as the cross validation sample size is nearly equal to the total sample size but because the training sets are nearly identical in this case variance of the prediction error can be quite high (Hastie et. al 2009).

  Default value = 10

  Options

  – Any integer less than the number of data points is valid but this is generally either set to 3 or 10

- *Stratify* (optional) Indicate whether cross-validation folds should be stratified by the response

  Default value = True (Checked)

  Options

  - True (Checked)
  - False (Unchecked)

- *Seed* (optional) The random number seed used to split the data. If one desires to reproduce results from a previous split fit, one must enter the random number seed that is reported in the console output from that split step.

  Default value = Randomly selected

  Options

  - Any integer between -2147483647 and 2147483647

Output Ports

- *outputMDS* This is an MDS file to be further used in the downstream workflow. It is nearly identical to the input MDS file but with an added column under the header "Split". for each non- background observation will be labeled either a number from 1 to the number of folds selected or "NA" indicating which fold each observation has been partitioned into or whether it will not be included in the Model Selection step likely because it is being withheld for model evaluation.

  Common connections

  - This port can connect to either the CovariateCorrelationAndSelection or directly to any of the R model modules (BoostedRegressionTree, GLM, MARS, or RandomForest).
  - While it could technically also connect to the input port of ModelEvaluationSplit, ModelSelectionCrossValidation or ModelSelectionSplit this would not make sense for SAHM workflows. The results of connecting to one of these modules has not been tested and could cause errors or subtle but significant problems with subsequent modeling.

References

Hastie T, Tibshirani R, Friedman JH. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag. 744 pp. 2nd ed.

17 ModelSelectionSplit

The ModelSelectionSplit reserves a portion of the data from the model fitting process but reports the evaluation metrics on all models not just the those selected as the final models to be reported in the analysis (in contrast to the ModelEvaluationSplit). This module should be placed directly the CovariateCorrelationAndSelection. If both a ModelEvaluationSplit and a ModelSelectionSplit are specified then the training portion of the Model-EvalutationSplit will be further partitioned by the ModelSelectionSplit thus the ModelEvalutationSplit should come first in the workflow. Both of these algorithms stratify the splits by the response. That is, the ratio of presence to absence points should be nearly equal in the testing and training split. If a ModelSelectionSplit is included evaluation metrics applied to the reserved data will be reported in the textual output, model evaluation plots including AUC plots as well as the across model plots and the csv. Both of these modules ignore background points and treat all observations with values greater than 0 as presence for the purpose of stratification by response.

It is not valid to select models based on their performance on the reserved portion of the data and then report these metrics only for the top performing models claiming that we would expect similar performance on an independent dataset see Hastie 2009 for this discussion. If one desires metrics for how the models might be expected to perform on an independent dataset then the ModelEvaluationSplit must be used.

Input Ports

- *inputMDS* (mandatory) This is the input data set consisting of location data for each sample point, the values of each predictor variable at those points. This input is usually provided by the upstream steps that precede the Test Training Split module. Any value entered here (e.g., specifying another existing MDS on the file system) will override the input specified by a model connection in the visual display.

Common connections

- This port can connect to the output port on MDSBuilder, ModelEvaluationSplit
- While it could technically also connect to the output from ModelSelectionCrossValidation, CovariateCorrelationAndSelection, or even another ModelSelectionSplit this would not make sense for SAHM workflows. The results of connecting to one of these modules has not been tested and could cause errors or subtle but significant problems with subsequent modeling.

- *trainingProportion* (optional) This is the proportion of the sample points that will be used to train the model, relative to the total number of points. Entered values should be greater than 0 but less than 1. For example, a value of '0.9' will result in 90% of the sample points being used to train the model, with 10% of the sample being held out to test the model's performance. Choosing an appropriate training proportion can depend on various factors, such as the total number of sample points available. Selecting an appropriate value for the training proportion is a complex issue that depends on many factors including the total number of observations, the complexity of the models that will be fit, and the signal to noise ratio in the data (Hastie et. al. 2009).

- *RatioPresAbs* (optional) This optional field is populated with a number corresponding to the desired ratio of presence to absence points to be used in the analysis. If not populated then all occurrence records (not background points) will be portioned into either the test or training split with no reduction in the total number of points. If populated, this entry should be a number greater than zero. (A value of '1' will result in an equal number of both presence and absence points being used, a value of '2' indicates that twice as many presence points will be used, a value of 0.5' indicates that twice as many absence points will be used, etc.). All field data points with a value equal to or greater than 1 are interpreted as presence points. Although the original field data is unmodified, this option will reduce the sample size as the merged dataset containing sample points will have points deleted from it to achieve the specified ratio as such it should be used with caution. A warning will be generated if more than 50% of either the presence or absence points will be deleted based on the ratio specified by the user. Background points are ignored by this module (they are read in and written out, but not assigned to either the test or training split).

- *Seed* (optional) The random number seed used by split the data. If one desires to reproduce results from a previous split fit, one must enter the random number seed that is reported in the console output from that split step.

  Default value = Randomly selected

  Options

  - Any integer between -2147483647 and 2147483647

Output Ports

- *outputMDS* This is an MDS file to be further used in the downstream workflow. It is nearly identical to the input MDS file but with an added column under the header "Split" for each non- background observation will be labeled either "test", "train", or "NA" indicating whether the observation will be used for producing evaluation metrics, training the model, or excluded from the analysis respectively. "NA" generally indicates that the observation was reserved for evaluating the final models but can also occur if a desired ratio of presence to absence was set using the RatioPresAbs.

  Common connections

  - This port can connect to either the CovariateCorrelationAndSelection or directly to any of the R model modules (BoostedRegressionTree, GLM, MARS, or RandomForest).
  - While it could technically also connect to the input port of ModelEvaluationSplit, ModelSelectionCrossValidation or even another ModelSelectionSplit this would not make sense for SAHM workflows. The results of connecting to one of these modules has not

been tested and could cause errors or subtle but significant problems with subsequent modeling.

References

Hastie T, Tibshirani R, Friedman JH. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer-Verlag. 744 pp. 2nd ed.

18 PARC

The Projection, Aggregation, Resampling, and Clipping (PARC) module is a powerful utility that automates the preparation steps required for using raster layers in most geospatial modeling packages. In order to successfully consider multiple environmental predictors in raster format, each layer must have coincident cells (pixels) of the same size, have the same coordinate system (and projection, if applicable), and the same geographic extent. The PARC module ensures that all of these conditions are met for the input layers by transforming and or reprojecting each raster to match the coordinate system of the template layer. This process usually involves aggregation (necessary when an input raster layer must be up-scaled to match the template layer– e.g., generalizing a 10 m input layer to a 100 m output layer), and or resampling (necessary for interpolating new cell values when transforming the raster layer to the coordinate space or cell size of the template layer). Lastly, each raster predictor layer is clipped to match the extent of the template layer.

The settings used during these processing steps follow a particular set of decision rules designed to preserve the integrity of data as much as possible. However, it is important for a user to understand how these processing steps may modify the data inputs. For additional information about the PARC module, please see the extended help and documentation for the SAHM package.

Input Ports

- *predictor* (optional) A single raster with resampling, aggregation, and categorical options.

  Common connections

  - value' port of a Predictor module Note - Multiple single Predictor modules can be connected to this single input port.

- *PredictorList* (optional) This is an in memory data construct that contains a list of predictors each with resampling, aggregation, and categorical options.

  Common connections

  - value‘ port of any of the 'Individual Predictors selector‘ modules Note - Multiple single Predictors selectors modules can be connected to this single input port.

- *RastersWithPARCInfoCSV* (optional) This is a CSV containing a list of files to include in the PARC operation.

  The format of this list conforms to the 'PredictorListFile‘ specs: Column 1: The full file path to the input raster layer including the drive. Column 2: A binary value indicating whether the input layer is categorical or not. A value of “0” indicates that an input raster is non-categorical data (continuous), while a value of “1” indicates that an input raster is categorical data. Column 3: The resampling method employed to interpolate new cell values when transforming the raster layer to the coordinate space or cell size of the template layer, if necessary. The resampling type should be specified using one of the following values: “nearestneighbor,” “bilinear,” “cubic,” or “lanczos.” Column 4: The aggregation method to be used in the event that the raster layer must be up-scaled to match the template layer (e.g., generalizing a 10 m input layer to a 100 m output layer). Care should be taken to ensure that the aggregation method that best preserves the integrity of the data is used. The aggregation should be specified using one of the following values: “Min,” “Mean,” “Max,” “Majority,” or “None.”

In formatting the list of predictor files, the titles assigned to each of the columns are unimportant as the module retrieves the information based on the order of the values in the .csv file (the ordering of the information and the permissible values in the file however, are strictly enforced). The module also anticipates a header row and will ignore the first row in the .csv file.

Common connections

  - 'value' port of PredictorListFile module

- *templateLayer* (mandatory) The template layer raster file used to define the Extent, Cell size, Projection, raster snap, and coordinate system of the outputs.

  Common connections

  - 'value' port of TemplateLayer module

- *ignoreNonOverlap* (optional) Option of using the intersection of all covariates and template or enforcing the template extent.

  Options

  - True (checked) = Use intersection of all covariates extents. Area of template extent will be reduce such all covariate layers extents can be completely covered by the new extent.
  - False (Unchecked) = The template extent will be used for all outputs and an error will be raised if any of the covariates are not completely covered by the template.

- *multipleCores* (optional) Option of running processing on multiple threads/cores.

  Options

  - True (checked) = Individual layers will be run consecutively on separate threads.
  - False (Unchecked) = All processing will occur on the same thread as the main program.

Output Ports

- *RastersWithPARCInfoCSV* (mandatory) The VisTrails output from the PARC module is a interim CSV file that contains information about each of the files processed. This is used by the MDS builder to determine which files to extract values from and which layers are categorical.

  Common connections

  – Input port 'RastersWithPARCInfoCSV' of MDSBuilder module

19 PointAggregationMethod

This module is a required class for other modules and scripts within the SAHM package. It is not intended for direct use or incorporation into the VisTrails workflow by the user.

Input Ports

- *value* (optional) NA

Output Ports

- *value_as_string* (optional) ToDo

20 Predictor

The Predictor module allows a user to select a single raster layer for consideration in the modeled analysis. Besides selecting the file the user also specifies the parameters to use for resampling, aggregation, and whether the data is categorical.

Input Ports

- *categorical* (optional) This parameter allows a user to indicate the type of data represented. The distinction between continuous and categorical data will maintained through a workflow by appending the word '_categorical' to categorical layer names in the resulting MDS file. It is also import to select the nearest neighbor resampling option for categorical layers.

  Default value = False (Unchecked)

  Options

  - True (Checked) - The data contained in the raster layer is categorical (e.g., landcover categories).
  - False(Unchecked) - The data contained in the raster is continuous (e.g., a DEM layer).

- *ResampleMethod* (mandatory) The resample method employed to interpolate new cell values when transforming the raster layer to the coordinate space or cell size of the template layer.

  Options

  - near: nearest neighbor resampling Fastest algorithm, worst interpolation quality, but best choice for categorical data.
  - bilinear: bilinear resampling, good choice for continuous data.
  - cubic: cubic resampling.
  - cubicspline: cubic spline resampling.
  - lanczos: Lanczos windowed sinc resampling.
  - see: http://www.gdal.org/gdalwarp.html for context

- *AggregationMethod* (mandatory) The aggregation method to be used in the event that the raster layer must be up-scaled to match the template layer (e.g., generalizing a 10 m input layer to a 100 m output layer). Care should be taken to ensure that the aggregation method that best preserves the integrity of the data is used. See the PARC module documentation for more information on how resampling and aggregation are performed.

Options

  - Mean: Average value of all constituent pixels used.
  - Max: Maximum value of all constituent pixels used.
  - Min: Minimum value of all constituent pixels used.
  - Majority: The value occurring most frequently in constituent pixels used.
  - None: No Aggregation used.

- *file* (mandatory) The location of the raster file. A user can navigate to the location on their file system. When a user is selecting an ESRI grid raster, the user should navigate to the 'hdr.adf' file contained within the grid folder

Output Ports

- *value* (mandatory)

  Common connections

  - The output from this port only connects to the PARC input port 'predictor'.

- *value_as_string* (optional) This is a VisTrails port that is not used in general SAHM workflows.

  Common connections

  - Does not generally connect to other SAHM modules.

21 PredictorList

This module is a required class for other modules and scripts within the SAHM package. It is not intended for direct use or incorporation into the VisTrails workflow by the user.

Input Ports

- *value* (optional) NA
- *addPredictor* (optional) NA

Output Ports

- *value* (optional) ToDo
- *value_as_string* (optional) ToDo

22 PredictorListFile

The PredictorListFile module allows a user to load a .csv file containing a list of rasters for consideration in the modeled analysis. The .csv file should contain a header row and four columns containing the following information, in order, for each raster input.

Column 1: The full file path to the input raster layer.

Column 2: A binary value indicating whether the input layer is categorical or not. A value of "0" indicates that an input raster is non-categorical data (continuous), while a value of "1" indicates that an input raster is categorical data.

Column 3: The resampling method employed to interpolate new cell values when transforming the raster layer to the coordinate space or cell size of the template layer, if necessary. The resampling type should be specified using one of the following values: "nearestneighbor," "bilinear," "cubic," or "lanczos."

Column 4: The aggregation method to be used in the event that the raster layer must be up- scaled to match the template layer (e.g., generalizing a 10 m input layer to a 100 m output layer). Care should be taken to ensure that the aggregation method that best preserves the integrity of the data is used. The aggregation should be specified using one of the following values: “Min,” “Mean,” “Max,” “Majority,” or “None.”

In formatting the list of predictor files, the titles assigned to each of the columns are unimportant as the module retrieves the information based on the order of the values in the .csv file (the ordering of the information and the permissible values in the file however, are strictly enforced). The module also anticipates a header row and will ignore the first row in the .csv file.

Input Ports

- *csvFileList* (optional) This is the CSV file on the file system. While not strictly mandatory this port will almost always have an input.
- *predictor* (optional) Allows a user to add individual Predictor modules to a PredictorListFile

  Common connections

  - The output port ’value‘ of a Predictor module.

Output Ports

- *RastersWithPARCInfoCSV* (mandatory) This port generally connects to the input port ’RastersWithPARCInfoCSV‘ on the PARC module.

23 RandomForest

Input Ports

- *mdsFile* (mandatory) The the input data set consisting of locational data for each sample point, the values of each predictor variable at those points This input file is almost always generated by the upstream steps.

Common connections

  - The mdsFile can be produced by any of MDSBuilder, ModelEvaluationSplit, ModelSelectionCrossValidation, ModelSelectionSplit, or CovariateCorrelationAndSelection.

- *makeBinMap* (optional) Indicate whether to discretize the continues probability map into presence absence. See the ThresholdOptimizationMethod for how this is done. If time is a concern and many models are to be fit and assessed maps can be produced after model selection for only the best models using the Select and Test the Final Model tool. Options are available for producing Probability, Binary and MESS maps there as well.

  Default value = False (Unchecked)

  Options

  - True (Checked)
  - False (Unchecked)

- *makeProbabilityMap* (optional) Indicate whether a map of predicted values is to be produced for the model fit.

  Default value = False (Unchecked)

  Options

  - True (Checked)
  - False (Unchecked)

- *makeMESMap* (optional) Indicate whether to produce a multivariate environmental similarity surface (MESS) and a map of which factor is limiting at each point see Elith et. al. 2010 for more details. If time is a concern and many models are to be fit and assessed maps can be produced after model selection for only the best models using the Select and Test the Final Model tool. Options are available for producing Probability, Binary and MESS maps there as well.

  Default value = False (Unchecked)

  Options

  - True (Checked)
  - False (Unchecked)

- *Seed* (optional) The random number seed used by BRT. If one desires to reproduce results from a previous randomForest fit, one must enter the random number seed that is reported in the textual output from that model fit. The seed used is always reported in the textual output.

  Default value = Randomly Generated

  Options

  - Any integer between -2147483647 and 2147483647

- *ThresholdOptimizationMethod* (optional) Determines how the threshold is optimized in order to discretize continuous predictions into binary. These are used for evaluation metrics calculated based on the confusion matrix as well as for the binary map. The value calculated for the train portion of the data will be applied to the test portion and if cross validation was specified, the value is calculated separately for each fold using the threshold from the training data and applying it to the test data for the hold out fold.

  Default value = 2

  Options

  - 1: Threshold=0.5
  - 2: Sens=Spec sensitivity=specificity
  - 3: MaxSens+Spec maximizes (sensitivity+specificity)/2
  - 4: MaxKappa maximizes Kappa
  - 5: MaxPCC maximizes PCC (percent correctly classified)
  - 6: PredPrev=Obs predicted prevalence=observed prevalence
  - 7: ObsPrev threshold=observed prevalence
  - 8: MeanProb mean predicted probability
  - 9: MinROCdist minimizes distance between ROC plot and (0,1)

- *mTry* (optional) By default this is optimized using the tuneRF function so that OOB error is minimized. See the CRAN website for more details.

  Default value = this is optimized using the tuneRF function so that out of bag error is minimized.

  Options

  – A number between 1 and the total number of valid parameters used in model fitting

- *nTrees* (optional) See the randomForest documentation on the CRAN website for details http://cran.r-project.org/web/packages/randomForest/index.html.

  Default value = randomForest function default

- *nodesize* (optional) See the randomForest documentation on the CRAN website for details http://cran.r-project.org/web/packages/randomForest/index.html.

  Default value = randomForest function default

  Options

  – See randomForest documentation for valid input

- *replace* (optional) See the randomForest documentation on the CRAN website for details http://cran.r-project.org/web/packages/randomForest/index.html.

  Default value = randomForest function default

  Options

  – See randomForest documentation for valid input

- *maxnodes* (optional) See the randomForest documentation on the CRAN website for details http://cran.r-project.org/web/packages/randomForest/index.html.

  Default value = randomForest function default

  Options

  – See randomForest documentation for valid input

- *importance* (optional) See the randomForest documentation on the CRAN website for details http://cran.r-project.org/web/packages/randomForest/index.html.

  Default value = randomForest function default

  Options

  - See randomForest documentation for valid input

- *localImp* (optional) See the randomForest documentation on the CRAN website for details http://cran.r-project.org/web/packages/randomForest/index.html.

  Default value = randomForest function default

  Options

  - See randomForest documentation for valid input

- *proximity* (optional) See the randomForest documentation on the CRAN website for details http://cran.r-project.org/web/packages/randomForest/index.html.

  Default value = randomForest function default

  Options

  - See randomForest documentation for valid input

- *oobProx* (optional) See the randomForest documentation on the CRAN website for details http://cran.r-project.org/web/packages/randomForest/index.html.

  Default value = randomForest function default

  Options

  - See randomForest documentation for valid input

- *normVotes* (optional) See the randomForest documentation on the CRAN website for details http://cran.r-project.org/web/packages/randomForest/index.html.

  Default value = randomForest function default

  Options

  - See randomForest documentation for valid input

- *doTrace* (optional) See the randomForest documentation on the CRAN website for details http://cran.r-project.org/web/packages/randomForest/index.html.

  Default value = randomForest function default

  Options

  - See randomForest documentation for valid input

- *keepForest* (optional) See the randomForest documentation on the CRAN website for details http://cran.r-project.org/web/packages/randomForest/index.html.

  Default value = randomForest function default

  Options

  - See randomForest documentation for valid input

Output Ports

- *modelWorkspace* The R workspace where all internal details regarding the fitted model are stored. This is used by the Select and Test the Final Model module.

  Common connections

  - 'modelWorkspace' port of SAHMModelOutputViewerCell for viewing the aspatial model output.
  - 'modelWorkspace' port of SAHMSpatialOutpuViewerCell for viewing the spatial model output in a mini GIS.

- *BinaryMap* If specified using MakeBinaryMap=True then a surface of binary predictions is produced by discretizing the probability map based on the selected threshold. This map indicates whether one could expect each site to be occupied or unoccupied based on the model.

- *ProbabilityMap* If specified using MakeProbabilityMap=True then a surface of predicted values is produced based on the tiffs in the input .mds file and the fitted model. These can but do not always indicate the probability of finding the species at a given site. For example if model calibration is poor then these will not agree well with the true probabilities though discrimination between presence and absences might still be good.

- *ResidualsMap* Model residual plots show the spatial relationship between the model deviance residuals. Most models assume residuals will be independent thus spatial pattern in the deviance residuals can be indicative of a problem with the model fit and inference based on the fit. It can for example indicate that important predictors were not included in the model and can be compared with the spatial pattern of predictors that were not included in the model.

- *MessMap* If specified by selecting makeMESMap=True the the MESS and MoD surfaces will be produced. The MESS surface is the multivariate environment similarity surface and shows how well each point fits into the univariate ranges of the points for which the model was fit. Negative values in this map indicate that the point is out of the range of the training data.

- *MoDMap* If specified by selecting makeMESMap=TRUE the the MESS and MoD surfaces will be produced. The MoD map is related to the MESS map and indicates which variable was furthest from the range over which the model was fit for each spatial location. See Elith et. al. 2010 for details on how the MESS map calculations are performed.

- *modelEvalPlot* For binary data this will be a Receiver operating characteristic curve. Which shows the relationship between sensitivity and specificity as the threshold for discretizing continuous predictions into presence absence is varied. The threshold selected using the specified ThresholdOptimizationMethod is shown. If a model selection test training split was specified the ROC curve for this will be shown in red and if a cross-validation split was specified ROC curves for each cross-validation fold will be overlaied with box plots summarizing cross-validation results. For count data this display will show several standard plots for assessment of model residuals.

- *ResponseCurves* Model response curves show the relationship between each predictor included in the model, while holding all other predictors constant at their means, and the fitted values. MARS response curves are shown on a logit scale thus the response axis will not necessarily be bounded on the 0 to 1 interval. BRT response curves will show response surfaces for any interaction terms included in the final model along with the percent relative influence.

- *Text_Output* This file contains a summary of the model fit. The information contained here includes the number of presence observations (counts equal to or greater than 1 for count models), the number of absence points, the number of covariates that were considered by the model selection algorithm. Note all of these can differ from the numbers in the original .mds due to incomplete records being deleted, and predictors with only one unique value being removed. The random number seed is recorded if applicable which allows completely reproducible results as well as a summary of the model fit. Evaluation Statistics are reported for the data used to fit the model as well as for the test or cross-validation split if applicable. References for how to interpret most of these are ubiquitous in the literature but it is worth mentioning that interpretation of the calibration statistics is described by Pearce and Ferrier 2000 as well as Miller and Hui 1991. Most metrics reported here can also be found in related graphical displays.

- *modelCalibrationPlot* The calibration plot shows the predicted probability of occurrence plotted against the actual proportions of occurrence for each of 5 bins along the probability axis. A logistic regression model is fit to the logits of the predicted probabilities of occurrence and is shown on the plot. These plots are used to determine how reliably a model will predict if a site is occupied or unoccupied (Pearce and Ferrier 2000)

References

Bivand, R.S., Pebesma, E.J., and Gomez-Rubio, V. (2008). Applied Spatial Data Analysis with R. Springer New York, NY.

Dormann, C.F., McPherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., et al. (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecography 30:60928.

Elith, J., Kearney, M., Phillips, S. (2010). The art of modeling range-shifting species. Methods Ecol Evol 1:330342

Liaw, A. and Wiener M. (2002). Classification and Regression by randomForest. R News 2(3), 18–22.

Miller, M.E., Hui, S.L., Tierney, W.M. (1991). Validation techniques for logistic regression models. Statistics in Medicine 10: 1213-26

Pearce, J., and S. Ferrier. (2000). Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133:225245.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

24 RasterFormatConverter

The RasterFormatConverter module allows a user to easily convert between raster file types for a group of rasters. This group can be specified as either all the rasters in a single directory or the rasters specified in a single MDS file (see below). All outputs will be sent to a folder named “ConvertedRasters” (followed by an underscore and a number corresponding to the run sequence of the module) within the user’s current VisTrail‘s session folder. Typically this module will be used within a workflow to convert the geotiff format used by the rest of the modules to the ascii format needed by Maxent. But the following file formats are accepted for both inputs and outputs: Arc/Info ASCII Grid, ESRI BIL, ERDAS Imagine, and JPEG and others. See the compiled by default options at http://www.gdal.org/formats_list.html for a complete list of the accepted file types.

Input Ports

- *inputMDS* (optional) Any merged dataset (MDS) format csv can be used as input to this module. All of the rasters pointed to in the third line of the file will be converted to the output format.

  Common connections

  - This can be connected to the output 'mdsFile' port on any of the following modules: MDSBuilder, ModelEvaluationSplit, ModelSelectionSplit, ModelCrossValidationSplit, or CovariateCorrelationAndSelection.

- *inputDir* (optional) An directory can be used to specify which files to process. All of the rasters (of any format) will be converted to the output format specified.

  Common connections

  - This does not generally connect to any other SAHM modules.

- *format* (optional) The format to convert all the input grids into. For Maxent this will be ASCII.

  Default value = Geotif

  Options

  - Geotif
  - Arc/Info Grid
  - ASCII Grid
  - ESRI Bil
  - ERDAS Imagine
  - JPEG
  - BMP
  - Additional uncommon file types are supported by GDAL. For the complete list see the 'compiled by default' options at http://www.gdal.org/formats_list.ht for a complete list of the accepted file types.

- *multipleCores* (optional) Option of running processing on multiple threads/cores.

  Options

  - True (checked) = Individual layers will be run consecutively on separate threads.
  - False (Unchecked) = All processing will occur on the same thread as the main program.

Output Ports

- *outputDir* (optional) The directory where all output files will be saved to.

  Default value = This directory name is created by the module and it will be located in the session folder.

  Common connections

  - This port will connect to the maxent input port 'projectionlayers'.

25 ResampleMethod

This module is a required class for other modules and scripts within the SAHM package. It is not intended for direct use or incorporation into the VisTrails workflow by the user.

Input Ports

- *value* (optional) NA

Output Ports

- *value_as_string* (optional) ToDo

26 ResponseType

This module is a required class for other modules and scripts within the SAHM package. It is not intended for direct use or incorporation into the VisTrails workflow by the user.

Input Ports

- *value* (optional) NA

Output Ports

- *value_as_string* (optional) ToDo

27 SAHMModelOutputViewerCell

The SAHM Spatial Output Viewer Cell provides a convenient means for viewing the numerous spatial outputs produced by individual model runs as well as the input presence and absence points and background points if applicable. The spatial viewer displays the outputs in an interactive Matplotlib chart which functions much like a full GIS.

Attached to each cell is a toolbar that allows changing of the displayed layer and the overlaid points

Input Ports

- *row* (optional) Entering a value here forces the output for this cell to appear on the row specified on the output spreadsheet. Row counts start from 1.

Default value = VisTrails will autoselect the next empty cell, but you do not have control over which outputs appear where.

- *column* (optional) Entering a value here forces the output for this cell to appear on the column specified on the output spreadsheet. Row counts start from 1.
- *model_workspace* (optional) This is the model workspace output by any of the R models. This widget finds all of the various outputs for display relative to the location of this file.

  Common connections

  - Connects to the output port 'modelWorkspace' on any of the R models (MARS, BRT, RandomForest, or GLM)
- *InitialModelOutputDisplay* (optional) The display tab to show initially.

  Default value = AUC Curves

  Options

  - Text
  - Response Curves
  - AUC Curves
  - Calibration
  - Confusion
  - Residuals

28 SAHMSpatialOutputViewerCell

SAHMModelOutputViewerCell is a VisTrails Module that displays the various output from a SAHM Model run in a single cell

Input Ports

- *row* (optional) Entering a value here forces the output for this cell to appear on the row specified on the output spreadsheet. Row counts start from 1.

  Default value = VisTrails will autoselect the next empty cell, but you do not have control over which outputs appear where.

- *column* (optional) Entering a value here forces the output for this cell to appear on the column specified on the output spreadsheet. Row counts start from 1.

- *model_workspace* (optional) This is the model workspace output by any of the R models. This widget finds all of the various outputs for display relative to the location of this file.

  Common connections

  – Connects to the output port 'modelWorkspace' on any of the R models (MARS, BRT, RandomForest, or GLM)

- *max_cells_dimension* (optional) For very large GIS datasets the Matplotlib based viewer used to display GIS data runs into memory issues. To get around this limitation we downsample the full resolution and only display the resampled image. This number represents the number of resampled pixels that will be displayed in the largest dimension. For example if your template layer is 20,000 by 10,000 pixels and the max_cells_dimension is set at 5000 your output will be displayed at 5,000 by 2,500 resolution. Zooming in the map does cause the resampled cells to be displayed at their original resolution as is the case in most GIS programs. If the resampled resolution is not sufficient for your needs you will need to display the final outputs in an external GIS program.

  Default value = 5000

29 TemplateLayer

The second fundamental input in an analysis is the template layer. It is used to define the extent and resolution that will be used in all subsequent analysis. The TemplateLayer is a raster data layer with a defined coordinate system, a known cell size, and an extent that defines the study area. The data type and values in this raster are not important. All additional raster layers used in the analysis will be resampled and reprojected as needed to match the template, snapped to the template, and clipped to have an extent that matches the template. Users should ensure that additional covariates considered in the analysis have complete coverage of the template layer used.

Output Ports

- *value* (mandatory) This is the actual file object that is being passed to other modules in the workflow.

  Common connections

  - The 'TemplateLayer' input port of the FieldDataAggregationAndWeight Module.
  - The 'TemplateLayer' input port of the PARC Module.

- *value_as_string* (optional) This is a VisTrails port that is not used in general SAHM workflows.

  Common connections

  - This does not commonly connect to other SAHM modules.