

Fort Collins Science Center

User Documentation for the Software for Assisted Habitat Modeling (SAHM) package in VisTrails.

By Colin Talbert, Marian Talbert, Drew Ignizio

Report Series XXXX–XXXX

U.S. Department of the Interior

U.S. Geological Survey

U.S. Department of the Interior

KEN SALAZAR, Secretary

U.S. Geological Survey

Marcia K. McNutt, Director

U.S. Geological Survey, Reston, Virginia: 201x  
Revised and reprinted: 201x

For more information on the USGS—the Federal source for science about the Earth,  
its natural and living resources, natural hazards, and the environment—visit   
<http://www.usgs.gov> or call 1–888–ASK–USGS

For an overview of USGS information products, including maps, imagery, and publications,  
visit <http://www.usgs.gov/pubprod>

To order this and other USGS information products, visit <http://store.usgs.gov>

Suggested citation:  
Author1, F.N., Author2, Firstname, 2001, Title of the publication: Place of publication   
(unless it is a corporate entity), Publisher, number or volume, page numbers; information   
on how to obtain if it’s not from the group above.

Any use of trade, product, or firm names is for descriptive purposes only and does not imply   
endorsement by the U.S. Government.

Although this report is in the public domain, permission must be secured from the individual   
copyright owners to reproduce any copyrighted material contained within this report.

**Contents**

Introduction 1

Installation of SAHM and VisTrails 2

Windows Installation 2

VisTrails Primer 11

Creating and Editing Workflows 11

Executing Workflows 13

Viewing Output in the Spreadsheet 14

Workflow History Tree View 15

Parameter Exploration 17

Using the SAHM Package 18

Basic Operation. 18

The session folder. 19

Data Inputs 20

***Field Data*** 20

***Covariates*** 23

***Predictor*** 23

***PredictorListFile*** 24

***Individual Predictors selector widgets.*** 26

Data Preparation 28

***Field Data Query*** 28

***Field Data Aggregate and Weight*** 30

***Project, Aggregate, Resample, Clip (PARC)*** 32

***How PARC works:*** 34

***Raster Format Converter*** 36

Pre-modeling Data Manipulation 37

***Merged Dataset Builder.*** 38

***ModelSelectionSplit and ModelEvaluationSplit.*** 41

***ModelSelectionCrossValidation*** 43

***CovariateCorrelationAndSelection*** 45

*Modeling* 48

***Input Required by GLM, MARS, Boosted Regression Tree, Random Forest*** 49

***Output Produced by All models*** 51

***Generalized Linear Model (GLM)*** 60

***Multivariate Adaptive Regression Splines (MARS)*** 61

***RandomForest*** 61

***BoostedRegressionTree (BRT)*** 62

***Maxent*** 65

***Select and Test the Final Model*** 65

Viewing Output 66

***Spreadsheet*** 66

***SAHM Model Output Viewer Cell*** 67

***SAHM Spatial Output Viewer Cell*** 67

Glossary 70

**Figures**

**Figure 1.** The main VisTrails Builder Window of the VisTrails 2.0 program. 11

**Figure 2.** The VisTrails Spreadsheet. 14

**Figure 3.** VisTrails History View. 16

**Figure 4.** VisTrails Parameter Exploration View. 18

**Figure 5.** A calibration plot showing good calibration. Note that the logit curve follows the diagonal quite closely. The intercept is not significantly different than zero and the slope is not significantly different than 1. 54

**Tables**

**Table 1.** Module execution color interpretation. 13

User Documentation for the Software for Automated Assisted Habitat Modeling (SAHM) package in VisTrails.

By Colin Talbert, Marian Talbert, Drew Ignizio, Nick Young

# 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 guide 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 appendix A.

# Installation of SAHM and VisTrails

## Windows Installation

The following detailed steps can be used to install and setup the VisTrails and the SAHM package on Windows. The various components are all freely available but require separate downloads and installations. These include VisTrails 2.0, R (developed under version 2.14.0), Maxent (developed under version 3.3.3k), GDAL (developed under version 1.8.0), and the GDAL Python bindings for Python 2.7. While not terribly difficult each component must be installed correctly and configured for the individual computer. The most difficult step in this process is getting the GDAL Python bindings installed into the Python installation that comes with VisTrails. Previous installations of any of the above components or PyQt might cause compatibility problems with VisTrails SAHM. These issues can be difficult to troubleshoot and fix. Particular care should be taken to ensure that the versions of each component are compatible. For example if you install the 64-bit version of VisTrails you should also install the 64-bit version of the GDAL Python bindings.

**NOTE: You must have Administrative Rights on the computer you are trying to install on.**

1. **Download and install VisTrails.**

Ready to install executables are available from <http://www.vistrails.org/index.php/Downloads>. Choose the most current release which is currently 2.0-alpha. The bundle is available for both 32-bit and 64-bit machines (to view the specs of a computer to determine which version you should get, go to Control Panel>System and Security>System). Install this application on your system using the provided installers. You will need administrative rights to complete this installation. Unlike some installations which will prompt a user to login as an administrator during the installation, a user should be logged in with admin privileges before attempting the installation, or else it will fail. Complete instructions and documentation are available from their site. Note that for the purpose of these instructions I installed my VisTrails to C:\VisTrails so subsequent instructions will be relative to this location.

1. **Download and install GDAL.**

32-bit and 64 bit versions are available from http://www.gisinternals.com/sdk/

A user should choose the 32-bit or 64-bit version to match their system architecture. We tested these instructions with the 'release-1600' version.

Specifically we used “Generic Installer for the GDAL Core Components”:

<http://www.gisinternals.com/sdk/Download.aspx?file=release-1600-x64-gdal-1-8-mapserver-6-0\gdal-18-1600-x64-core.msi> for 64-bit architecture and

<http://www.gisinternals.com/sdk/Download.aspx?file=release-1600-gdal-1-8-mapserver-6-0\gdal-18-1600-core.msi> for 32-bit architecture.

When it finishes downloading double click the msi and choose the 'complete installation' option.

The final step for GDAL is to add its location to your Path variable and add a couple of required Environmental (System) Variables to your system. On your machine’s system, navigate to Control Panel > System and Security > Advanced System Settings (on the left) > Environmental Variables



Click on the Environment Variables tab and add a new system variable called 'GDAL\_DATA' that points to the 'gdal-data' folder within the GDAL install location. Note this location will be specific to where you installed GDAL and might be C:\Program Files (x86)\GDAL\gdal-data if you installed the 32-bit version on a 64-bit machine.



Next add a new system variable called 'PROJ\_LIB' that points to the 'projlib' folder in the GDAL install location:



Now click on and edit the 'Path' variable. Append the path to the root GDAL folder to the variable value for the 'Path' variable. Include a semicolon at the end, with no spaces, 

when done modifying the 'Path' System Variable:



1. **Download and install the GDAL bindings for Python.**

These are available from the same location we obtained the GDAL core from, <http://www.gisinternals.com/sdk/> Once again match the 32-bit or 64-bit architecture of your system and be sure to grab the version for Python 2.7.

Specifically we used “Installer for the GDAL python bindings (requires to install the GDAL core)”:

<http://www.gisinternals.com/sdk/Download.aspx?file=release-1600-x64-gdal-1-8-mapserver-6-0\GDAL-1.8.1.win-amd64-py2.7.msi> for 64-bit architecture and

<http://www.gisinternals.com/sdk/Download.aspx?file=release-1600-gdal-1-8-mapserver-6-0\GDAL-1.8.1.win32-py2.7.msi> for 32-bit architecture.

Once it finishes downloading double click to install. For the installation of these bindings you must choose ‘Python from another location’ and then point to the python installation used by VisTrails. In our case I used C:\VisTrails\vistrails\Python27\_64 Note that if you've installed the 32-bit version of VisTrails this will be C:\VisTrails\vistrails\Python27 When providing the location after specifying "Python from another location," a user should be providing the location in which the python.exe program resides. We want to provide the location where the bindings should be installed, rather than pointing to the full path of the actual python.exe file.



The final thing to do is copy the DLL files from the GDAL install location to the VisTrails folder in the VisTrails installation.

-In GDAL install location (C:\Program Files\GDAL), sort all files by type. Select all the DLL files (these might be labeled 'Application extension') and copy them. Paste the files here: C:\VisTrails\vistrails



1. **Download and install the SAHM package.**

This is available from <http://www.fort.usgs.gov/xxxxxxxxxxxxxxxxxx>. Unzip this into the VisTrails packages directory which is at C:\VisTrails\vistrails\packages.

1. **Download and install the external software R**

R is available from <http://www.r-project.org/>. If you have a previously installed version of R you might be OK with it but we have only tested our code with version 2.14.0 for Windows. All required libraries will be programmatically downloaded and installed assuming you have internet connectivity at the time you first run an R module. The following libraries are loaded as needed ade4, foreign, gbm, lattice, mda, PresenceAbsence, randomForest, raster, rgdal, sp, spatstat, survival, tcltk2, and tools.

1. **Download and install the external software Maxent**

The Maxent package is available for download from: <http://www.cs.princeton.edu/~schapire/maxent/>

1. **Configure the SAHM package to point to where the above packages are installed.**

Once all the dependencies are installed we just need to let the SAHM package know where the various applications are installed. To do this open the VisTrails application and select Edit -> Preferences -> Module Packages (tab). Select and enable sahm in the list of Disabled packages. 

Click Configure.



By double clicking in the 'Value' area you can change the path to point to the location where various components were installed. In our case we used:

**gdal\_path** = C:\Program Files

note: this is the folder that contains the GDAL folder

**r\_path**  = C:\Program Files\R\R-2.14.0\bin

note: this is the bin folder within your R installation

**maxent\_path** = C:\Maxent

note: this is the folder that contains the maxent.jar file

**Additional configuration:**

While you are setting the above you could also change the following configuration variables:

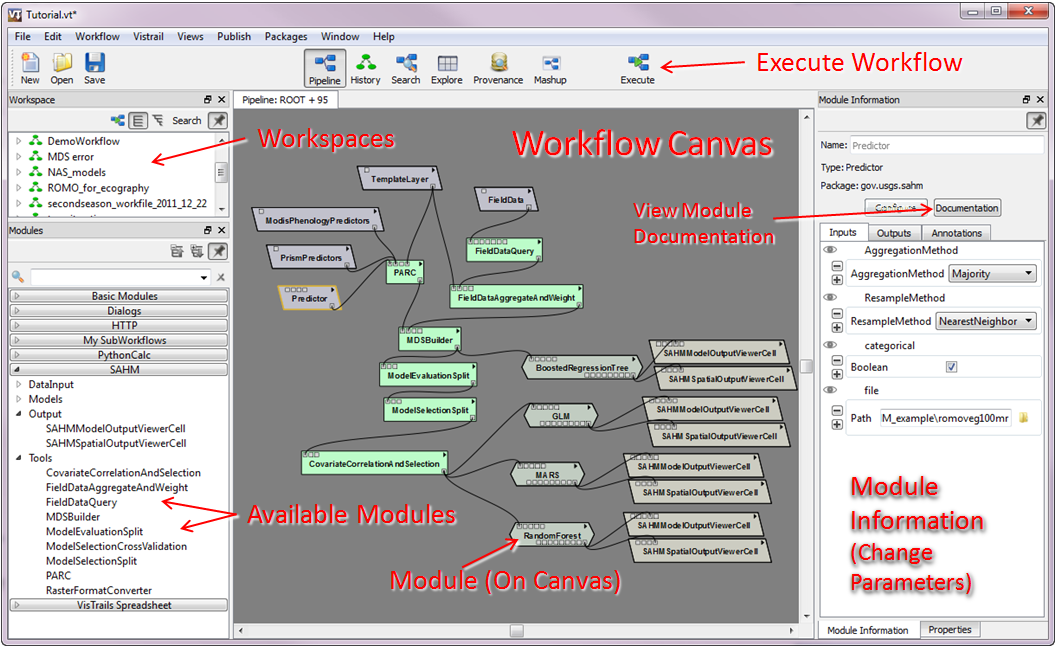
**output\_directory** = The default location you want SAHM to store your session folders.

**verbose** = Whether to print out all output messages to console screen. This

should stay the default value of True.

# VisTrails Primer

VisTrails is covered in detail in their own documentation and much of the content won’t be repeated here. The complete documentation is available at <http://www.vistrails.org/usersguide/VisTrails_Documentation/VisTrails_Documentation.html> This section is meant as a quick introduction to the highlights of VisTrails.



1. The main VisTrails Builder Window of the VisTrails 2.0 program.

## Creating and Editing Workflows

When you first open VisTrails you will see the window above with a blank Workflow Canvas. In the Workflow Canvas window you will view, edit and create workflows. A workflow is composed of individual Modules that each perform a specific function. The workflow as well as a complete history (provenance) of all changes and parameters used can be saved to a VisTrails (.vt) file. The Available Module are show to the left of the Workflow Canvas. These are arranged by the packages they belong to. Some packages are distributed with the VisTrails application others, such as SAHM, are created and distributed by others to enable discipline specific functionality within VisTrails.

Modules are added to a workflow by dragging the item from the Available Modules list onto the Workflow Canvas. You will notice on each module a series of small squares on the top and bottom of its shape. These represent input and output 'ports' or parameters for the module. The inputs and parameters are along the top and outputs are on the bottom. By hovering over a port you will get a pop-up text box that gives the port name and data type. Complete Module documentation is available by either clicking the small triangle in the upper right of the Module and selecting 'View Documentation' or by clicking the 'Documentation' button in the Module Information panel. Additional detailed documentation about each port, including the modules and ports it can be connected to is available by left clicking the port name in the Module Information panel and selecting 'View Documentation'.

Modules are connected into a workflow by clicking and dragging from the output port of one module into the appropriate input port of the next module. There are numerous valid ways of connecting modules depending on the processing required. When starting out it can be difficult to know which modules connect where. It will be helpful at first to open and modify an existing workflow instead of starting from scratch. The module and port documentation as well as the package documentation (this document) can be informative as well. Modules or connections that are no longer needed can be removed by selecting them and clicking 'Delete'.

As workflows become larger you might find it helpful to resize and navigate around the workflow canvas. The scroll wheel and sideways scroll wheel action move around (pan) the workflow canvas. Panning can also be accomplished by holding shift and the left mouse button and moving the mouse. By holding down the right mouse button and moving the mouse up or down you can zoom in and out. A clicking control+r will re-center the canvas.

## Executing Workflows

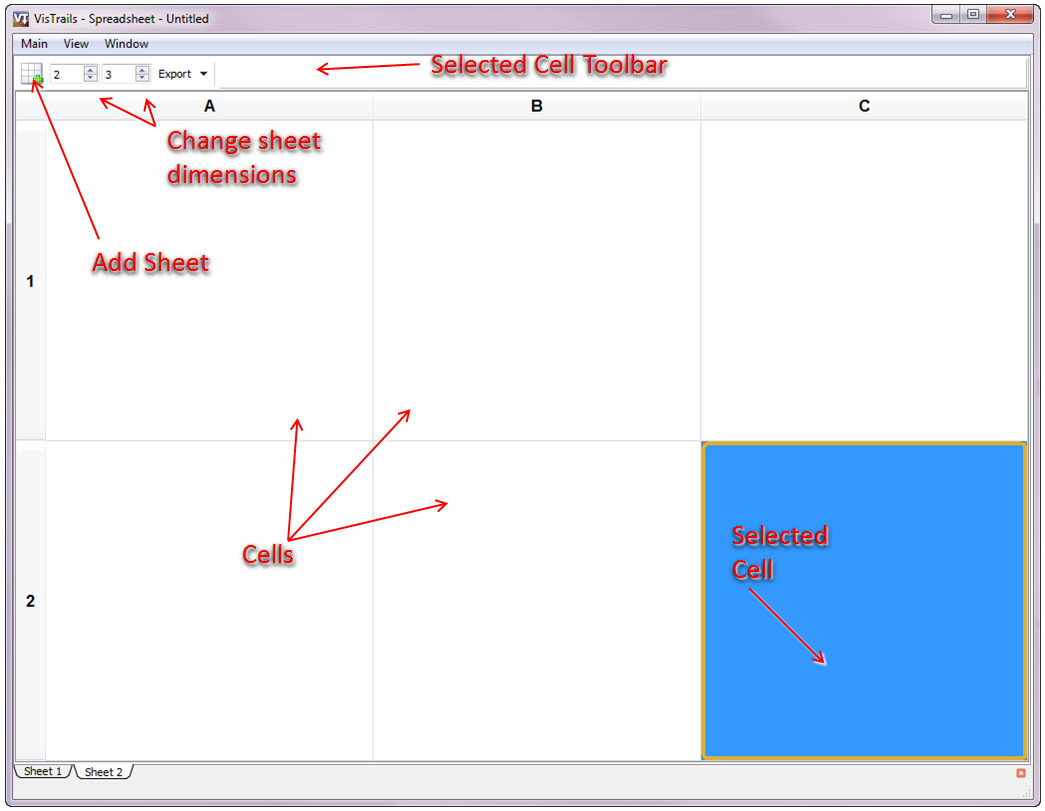
Once a workflow has been set up and configured it can be run by clicking on the 'Execute Workflow' button on the top of the main window. VisTrails will sequentially run through the execution of each module starting with the ones that have no inputs feeding into them and proceeding down the workflow. VisTrails caches the results of each models execution so that subsequent runs will not rerun the modules that completed successfully unless one of their inputs or parameters were changed. All modules downstream of a rerun module will be rerun. To override this property you can click Workflow -> Erase Cache Contents and subsequent runs will run all modules. As execution progresses the color of the modules in a workflow will change according to the table below.

1. Module execution color interpretation.

|  |  |  |  |
| --- | --- | --- | --- |
| Module Color | Meaning | Example | |
| Green | Module was executed |  |
| Red | Module execution failed |  | |
| Yellow | Module is currently being executed |  | |
| Orange | Module was cached |  | |
| Lilac | Module was not executed |  | |

## Viewing Output in the Spreadsheet

VisTrails has a spreadsheet window for viewing the various output produced by a workflow or series of workflows. Certain modules allow outputs or files to added to the spreadsheet. By default these modules will be added to the next available cell but the cell location can also be explicitly specified. Each specific type of spreadsheet cell module can display a particular type of file and will allow for particular types of output manipulation.

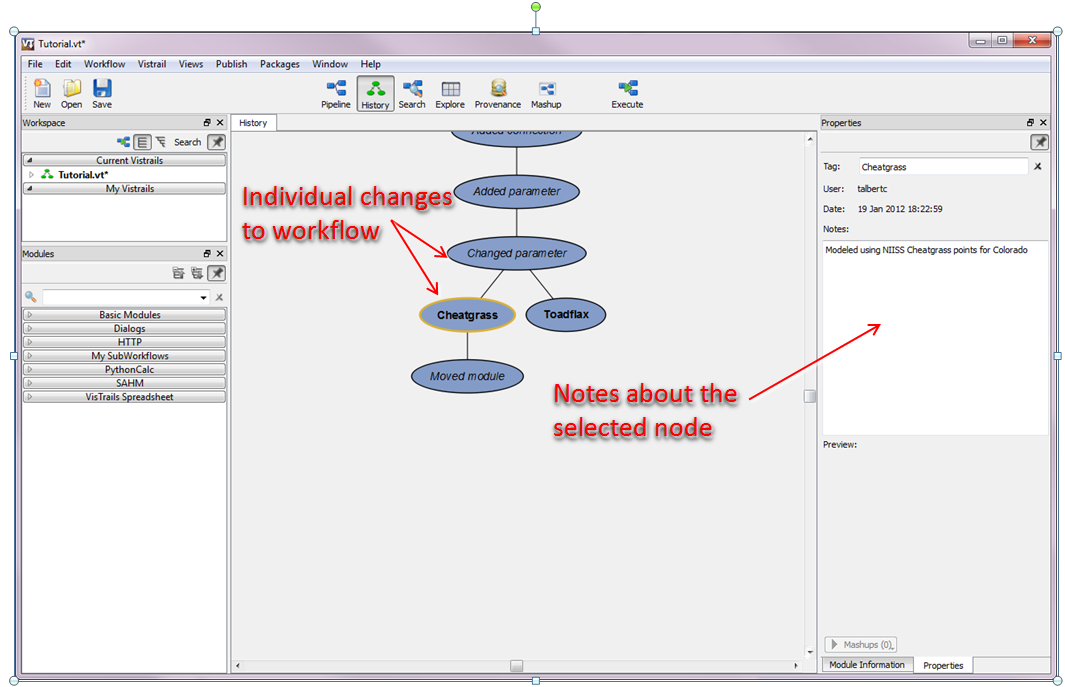


1. The VisTrails Spreadsheet.

The number of rows and columns displayed on a spreadsheet should be changed to match the output from a workflow or series of workflows. Additional sheets can be added for subsequent runs and sheets can be renamed to organize results.

## Workflow History Tree View

One of the strengths of VisTrails is that it tracks a complete history of all steps in the evolution of a workflow. If you would like to review or return to a previous step of your history it can be done using the History Tree View which is accessed with the 'History' button at the top of the main window. Since the History Tree View allows a full tree structure, relabeling of nodes on the history, and additional annotation this is a convenient and powerful means of documenting the process resulting in a set of final results. Additionally it can allow a user to quickly return to a previous workflow state if a particular line of processing proves unfruitful.



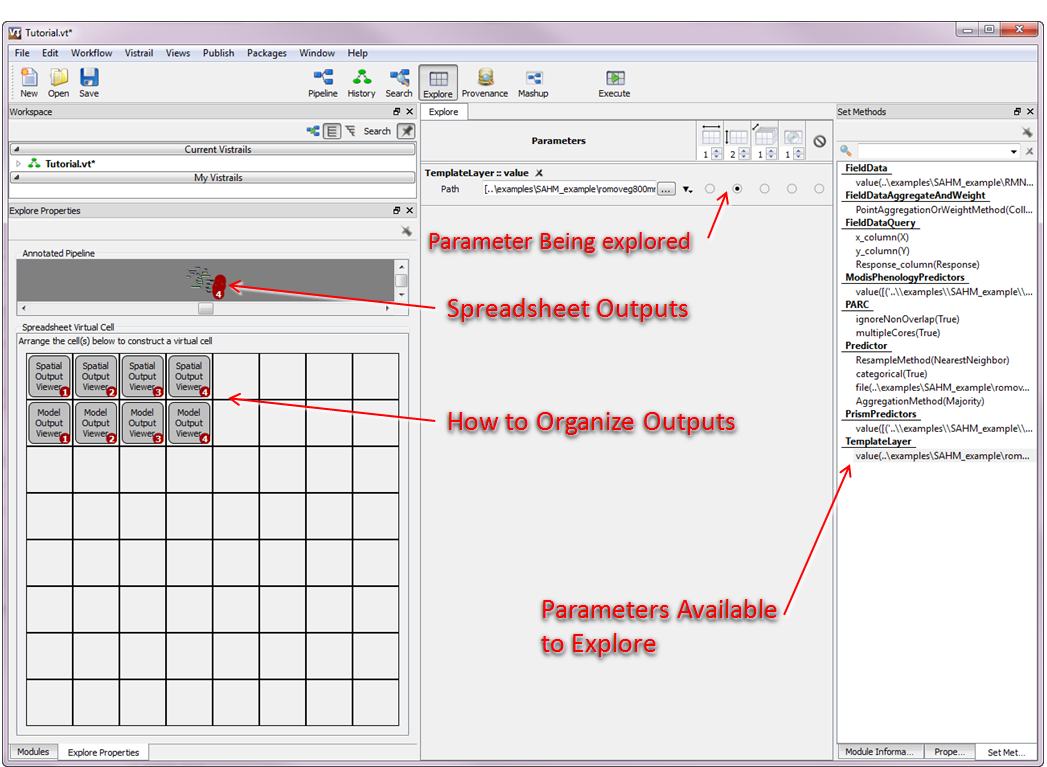
1. VisTrails History View.

The same navigation and panning functionality described in the Creating and Editing workflows section applies to the History View Canvas. To return to a previous workflow state select the node in the appropriate node on the tree and when you change back to the Pipeline View (Workflow Canvas) the modules, parameters, and connections will be as they were at that point in time. To change the label on a node double click on the text in that node and write a more meaningful tag. Additional notes about a node can be entered into the notes field in the properties tab to the right of the main canvas.

In addition to the history tab the Provenance tab provides functionality to view the results of each individual run executed on a workflow. This provides an execution log that complements the information stored in the history view.

## Parameter Exploration

The Parameter Exploration tab includes tools for easily seeing how changes to a parameter or series of parameters affects workflow output. This might be a series of input files or a range of values for a given parameter. To explore a parameter drag the name from the list in the 'Set Methods' tab to the right onto the 'Explore Canvas'. Next Change the values to include how you want to explore the parameter. Clicking the button with three ellipses opens a wizard to assist with this. Next decide how you want each iteration's output to be organized on the output spreadsheet. You can specify rows, columns, additional sheets, or as a movie over time. This is set with the radio buttons next to the parameter being explored on the 'Explore Canvas'. Finally you need to arrange the outputs from each individual iteration using the Spreadsheet Virtual Cell tab to the right of the 'Explore Canvas'.



1. VisTrails Parameter Exploration View.

# Using the SAHM Package

## Basic Operation.

When getting started with the SAHM VisTrails package it is helpful to refer to some previously used workflows to see how modules are commonly connected. To help with this we have provided some template workflows (.vt) that demonstrate the order that components generally go together. These templates are available in the package directory in a folder called templates (ToDo. Design and test these templates). For any module that you are unfamiliar with you can refer to documentation from within VisTrails by clicking on the small triangle in the upper right of the widget and selecting ‘View Documentation’. When connecting modules make sure that the input and output ports are connected to the appropriate location.

The VisTrails application consists of at least two main windows, the GUI and a secondary console window. SAHM displays various messages and warnings to this window. When running SAHM it is generally helpful to periodically check this console window for progress and warning information. If an error is encountered during execution the module that had the error will turn red and an error message will be shown when the module is hovered over. Additional information might be shown in the console window.

## The session folder.

All of the outputs produced during a single session of SAHM in VisTrails are saved to a single directory which we refer to as a ‘session folder’. When you first start up SAHM in VisTrails a new uniquely named folder is created in the directory specified in the SAHM configuration. If you would like to continue using a folder from a previous session of change where outputs are being saved to you can select Packages -> SAHM -> Change Session Folder

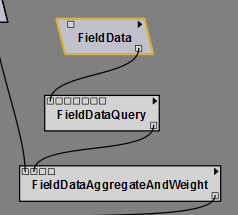


Within the session folder individual outputs are generally uniquely, sequentially named so that subsequent outputs do not overwrite previous outputs. For example the the first output from the Field Data Aggregate and Weight module will be called FDAW\_1.csv. The next time this module is run the output will be FDAW\_2.csv. In the case of modules that produce several related outputs the group will be placed in a folder named with the same unique sequential convention. The exceptions to this pattern are the Covariate Corelation and Selection module which will have the name from its selectionName parameter PARC module which will be send its output to a folder named according to the template layer used.

## Data Inputs

All data input modules are designated with a parallelogram shape and a pale lavender color.



****

### Field Data

The fundamental input to a modeling session is the field data file. At a minimum this file contains rows with the X and Y coordinates and the recorded detection of a species at that point. It is possible to supply a file that contains additional inputs and use the Field Data Query Module to subset/reformat the output (see Field Data Query Module documentation for more information).

The only Method that must be specified is the value (Path) which is the location of the csv file. The name parameter is an artifact of VisTrails and is not used.

The input data for this module must be in the form 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 (e.g., longitudinal, easting, X, etc ) positional data for a sample point. The “Y” field should be populated with the vertical (e.g., latitudinal, northing, Y, etc) 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). Absence points are optional for Maxent.

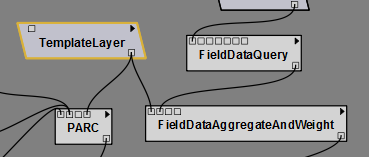
Format 2

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 (e.g., longitudinal) positional data for a sample point. The “Y” field should be populated with the vertical (e.g., 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.

The FieldData port 'path' can be connected to:

1. FieldDataQuery port 'fieldData\_file'
2. FieldDataAggregateAndWeight port 'fieldData'
3. MDSBuilder port 'fieldData'



**Template Layer**

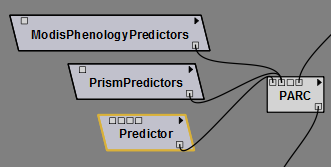
The second fundamental input in an analysis is the template layer. It is used to define the extent and resolution that will be used in all subsequent analyses. The TemplateLayer is a raster data layer with a defined coordinate system, a known cell size, and an extent that defines the study area. The 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.

The TemplateLayer port 'path' can be connected to:

1. PARC port 'templateLayer'
2. FieldDataAggregateAndWeight port 'templateLayer'

### Covariates

The SAHM package provides several ways of specifying which covariates or predictors will be included in subsequent analysis.



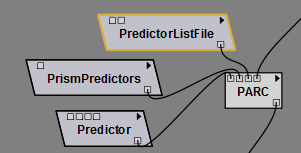
### Predictor

The Predictor module allows a user to select a single raster layer for consideration in the modeled analysis. Four parameters must be specified by the user:

1. Aggregation Method: 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.
2. Resample Method: 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.
3. Categorical (Boolean): Checking this box indicates that the data contained in the raster layer is categorical (e.g., landcover categories). Leaving this box unchecked indicates that the data contained in the raster is continuous (e.g., a DEM layer). This distinction is important in determining an appropriate resampling method.
4. File Path: 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.

The Predictor port 'value' can only be connected to:

1. PARC port 'predictor'



### 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 input raster.

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.

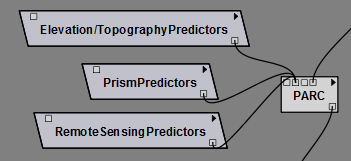
Column 3: The resampling method employed to interpolate new cell values when transforming the raster layer to the coordinate space or cell size of the template layer, if necessary. The resampling type should be specified using one of the following values: "nearestneighbor," "bilinear," "cubic," or "lanczos."

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.

The PredictorListFile output port 'RastersWithPARCInfoCSV' can only be connected to:

1. PARC port ' RastersWithPARCInfoCSV '



### Individual Predictors selector widgets.

The SAHM package allows users to set up dynamic Predictor selection widget which allow users to easily select from commonly used predictors. Examples might include Daymet, PRISM, MODIS, etc. For this to work the layers must exist on the local file system and the user must configure a csv file which contains the locations and parameters for each. This csv is located in the SAHM package and is called layers.csv. The first column (FileName) contains the full path to each individual file. The second column (LayerName) is a short descriptive name for this layer. The third column (Group) specifies which widget this layer will be found in. All of the distinct entries in the Group column will generate individual widgets with similar data. In our case we found it convenient to separate climate layers from phenology layers, from topology layers. The fourth column (Source) provideds a means of grouping layers within an individual layer. Each unique entry within a widgets layers will generate a collapsible group of similar layers. The fifth and final column (Categorical) contains a ‘Y’, ‘N’ Boolean value indicating whether this layer is categorical. If it is categorical Resampling will default to ‘nearest neighbor’ and Aggregation will default to ‘majority’. If it is not categorical Resampling will default to ‘bilinear’ and Aggregation will default to ‘mean’. See the original layers.csv for an example of how to set up this file.

The resulting widgets derived from this file are named the unique Group column name followed by ‘Predictors’. For example you might end up with a ClimatePredictors widget and a PrismPredictors widget. To select individual layers from each widget click the small triangle in the widget’s upper left and select ‘Edit Configuration’ in the dropdown. A box will pop up which contains the layers managed by this widget. Checking the boxes at the left will include that layer in the model. Optionally you can also change the categorical, resampling and aggregation parameters used in subsequent processing.



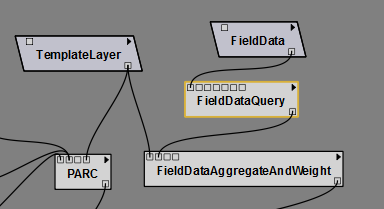
Within the widget convenience functions to select all layers or switch the selection are provided via buttons at the bottom. The Query box can be used to easily add/remove multible layers. To use enter text in the Query box that is unique to the layers that you want to include or remove. The full paths to the layers will be individually searched for the provided text. For example I might want to include all layers that have MonthlyPrecip in the path. Note that the query string is case sensitive.

When the final selection is made click OK to continue.

The IndividualPredictorSelector output port 'PredictorList' can only be connected to:

1. PARC port 'PredictorList'

## Data Preparation

****

#### Field Data Query

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 occurance 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 reformat this output into the format used by the SAHM package.

At a minimum you must specify the columns that contain the X, Y, and response. 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. Finally you must select the response type of either Presence(Absence) or Count. If none is selected it will default to Presence(Absence).

If you would also like to select by rows you can use the Query and Query\_column methods. The Query\_column is the column that contains the data you would like to query on. It will be selected similarly to how columns are selected for X, Y, and response. For the Query column you can either enter a single value or enter an equality statement with x used as a placeholder to represent the values in the query column.

*For example:*

*x < 2005 (would return values less than 2005)*

*x == 2000 or x == 2009 (would return 2000 or 2009)*

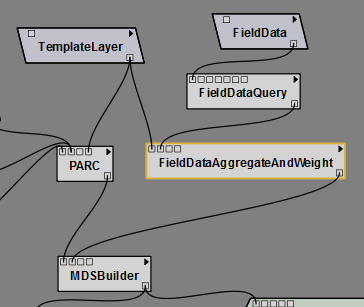
*The syntax is python in case you want to create an involved query.*

The FieldDataQuery input port 'fieldData\_file' can be connected to:

1. FieldData port 'value'

The FieldDataQuery output port 'fieldData' can be connected to:

1. FieldDataAggregateAndWeight port 'fieldData'
2. MDSBuilder port 'fieldData'



### Field Data Aggregate and Weight

In many instances data collected in the field can be redundant, both spatially and temporally. 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. Similarly, field data is often collected temporally through time, where observations of the same location are often recorded annually.

Currently three parameters can be specified by the user:

1. Field Data: The file with the X, Y, Response data. This can be specified here or come from the FieldDataQuery or FieldData modules
2. Template Layer: Raster used to determine cell size and extent. This can be specified here or come from the TemplateLayer module.
3. PointAggregationOrWeightMethod: Currently four options are available allowing several options for how weights are set
4. Collapse In Pixel: All of the field data points that occur within the bounds of a given pixel are collapsed to the centroid of that pixel. The value assigned to this pixel will be presense (1) if any of the collapse points were presense. If using count data the count will be the sum of all the collapsed points.
5. Weight Per Pixel: All of the field data points that occur within the bounds of a given pixel will be a assigned a weight of 1/the number of points in that pixel.
6. Inverse Density: Sets point weights inversely proportional to their spatial proximity to other points using an isotropic Gaussian smoothing kernel to calculate the density using the density.ppp function in the spatstat library in R using the leaveoneout=TRUE option which calculates the density at each point ignoring the point itself so that a completely isolated point will have weight equal to 1. The standard deviation of the kernel is by default a function of the area of the minimum convex polygon containing all points. This can optionally be set and explored by the user and an option is available to set this parameter to half the pixel length. The density calculated for the points is multiplied by 2πσ2 to adjust for the height of the Gaussian. Once the density is calculated, weights are set equal to 1/sqrt(1+density at the point).
7. Total Presence = Total Absence: Sets weights so the sum of the weights for presence points is equal to the sum of the weights for absence point by down-weighting absences. Presence points will have weight 1.

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

The FieldDataAggregateAndWeight input port 'fieldData' can be connected to:

1. FieldData port 'value'
2. FieldDataQuery port 'fieldData'

The FieldDataAggregateAndWeight input port 'TemplateLayer' can be connected to:

1. TemplateLayer port 'value'

The FieldDataAggregateAndWeight output port 'fieldData' can only be connected to:

1. MDSBuilder port 'fieldData'

### 

### Project, Aggregate, Resample, Clip (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.

The PARC module accepts four kinds of inputs:

1. Predictor List: A user should not have to populate this field. This field is populated at execution time with selections made from the pre-loaded .csv file of raster predictor layer inputs (specified during the SAHM install) and connected to the PARC module in the visual display.
2. ListRasterWithPARCInfoCSV: The .csv file list corresponds to the Predictor List File element and allows a user to load a .csv file containing a list of rasters for consideration in the modeled analysis. For additional information, please see the documentation for the Predictor List File element.
3. Predictor: The predictor input allows a user to select a single raster predictor layer to be considered in the analysis. It is recommended that a user add this input as a separate element in the visual display (and then link it to the PARC module) so that the aggregation and resampling settings can be established. The PARC module can accept multiple predictor elements. For additional information, please see the documentation for the Predictor element.
4. Template Layer: The template layer is a raster data layer with a defined coordinate system, a known cell size, and an extent that defines the study area. This raster layer serves as the template for all the other inputs in the analysis. 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 any additional layers considered in the analysis have coverage within the extent of the template layer. The template layer is a required input for the PARC module.

Additional Methods:

1. multipleCores: If checked individual layers will be processed in parallel on each of the available cores on a machine. This can lock up a machine until processing is finished but processing times are much, much faster.
2. ignoreNonOverlap: If checked the requirement that every covariate completely covers the template is ignored. The output extent will be the intersection of the covariates extents. The output will be snapped to the template raster. Use this with caution.

The Parc output port 'RastersWithPARCInfoCSV' can only be connected to:

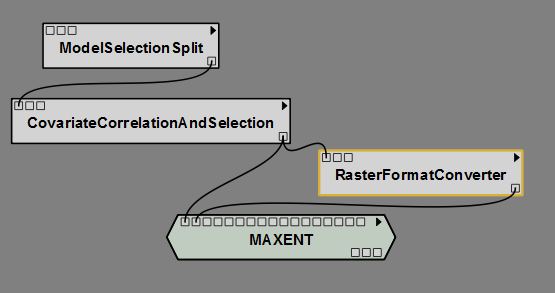
1. MDSBuilder port 'RastersWithPARCInfoCSV '

#### How PARC works:

PARC uses a combinataion of GDAL (Which used Proj.4), numpy, and python code to accomplish its task. The processing steps that PARC uses are fairly complex and geospatial implications should be considered fully. The default transformation chosen by GDAL will be used in the reprojection but in some cases this is not optimal and can introduce significant errors. In particular be careful with switching datums, using unusual projections or dataums, using projections or datums far outside of their intended area of accuracy. When in doubt you might want to perform reprojections in an external GIS package. See <http://www.remotesensing.org/geotiff/proj_list/> for more information.

For each raster that is going to transformed into the projection, extent, and cell size of the template raster the following steps are performed:

1. The ratio of source cell size to template cell size is determined. If the source raster cell dimension is greater that 0.5 of the template raster cell dimension no aggregation will be performed. Otherwise aggregation will be performed if specified. The values used in the comparison are the units in the template projection.
2. The source raster is projected into the template projection and extent. If no aggregation is to be used the output cell size will be identical to the template cell size. If aggregation is to be performed the cell size is calculated to be the value closest to the original source cell size that results in template cells containing even numbers of source cells. For example if aggregating from 30m to 100m the value would be 25m. These new pixels will be aligned with the origin of the template such that a specified number of whole pixels fall within each template pixel. The source pixel type and nodata value will be maintained.
3. If required aggregation is performed on the resulting layer. To aggregate, all of the reprojected source pixels that fall within a given template pixel are considered using the specified algorithm minimum, maximum, mean or majority and the resulting value is written to the output raster.



### Raster Format Converter

The RasterFormatConverter module allows a user to easily convert .tif raster layers into a different raster format for use and display in other software packages. The module accepts as an input either a list of rasters from a Merged Dataset File (MDS) in the workflow or the location of a directory containing multiple raster files. 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.

Three parameters can be specified by the user:

1. Format: The format corresponds to the desired raster output format. The following output file formats are supported: Arc/Info ASCII Grid, ESRI BIL, ERDAS Imagine, and JPEG.

To specify the desired output, users should enter the values shown below.

For an ASCII (.asc) output, enter: "asc"

For an ESRI BIL output, enter: "bil"

For an Erdas Imagine (.img) output, enter: "img"

For a JPEG (.jpg) output, enter: "jpg"

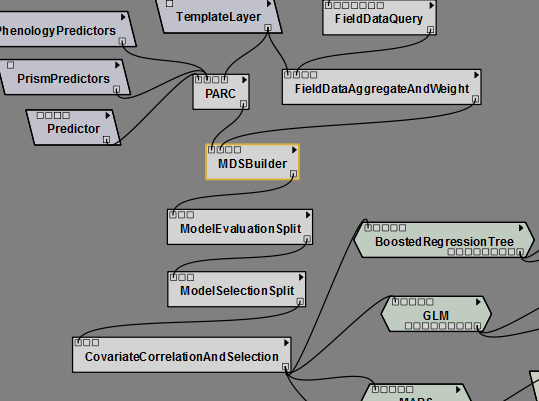
If no value is entered by the user, the module will default to an ASCII (.asc) output format.

1. Input Directory: The input directory allows a user to point to an entire folder as an input to the RasterFormatConverter. The contents of the specified folder will be checked for raster files and all the raster files contained within the directory will be converted to the format specified in the "Format" dialogue box. The module will identify and convert files of the following raster types: .bil, .img, .tif, .jpg, and .asc.
2. Input Merged Dataset (MDS): The input merged dataset allows a user to specify a .csv file created in the VisTrails workflow (containing a list of .tif raster files) as an input to the raster converter. All of the files listed in the MDS will be converted to the raster format specified in the "Format" dialogue box.

## Pre-modeling Data Manipulation

Much of the remaining workflow uses a file format we devised called Merged Dataset or MDS. This format combines the X, Y, response columns of the field data file with the pixel values extracted from the individual covariate rasters at each point. In this regard it is similar to the Samples With Data (SWD) format used by Maxent. The MDS format uses three headerlines to record and transmit additional pieces of information as well. The first header line contains the shortnames for each covariate. Additionally for categorical covariates the work \_categorical is appended to the end of the name. This will indicate to subsequent models how to treat this variable. The second line in an MDS file contains zeros and ones which indicate whether to include or remove this covariate from subsequent models. The third line contains a file path to the location of the raster used to extract values.

In addition to the above items the MDS file stores a few bits of information about its provenance. In cell 1, 2 is the template used for the field data aggregate and weight. In cell 2,2 is the field data file used. Cell 2, 3 contains the name of any selection applied to the MDS. Cell 1,3 contains the raster used for the PARC operation (this should always be identical to cell 1, 2). Cell 2,3 contains the folder that PARC output was saved to.

****

### Merged Dataset Builder.

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 MDSBuilder accepts four inputs from the user:

1. Rasters with PARC Info .csv File: The raster layers listed in this file will have their values extracted to the points in the field data .csv file. This parameter should be supplied by connecting the output of a PARC module within the workflow to the MDSBuilder module. (In order to properly extract the values of each predictor layer to the field data points, all the layers must have matching coordinate systems and cell sizes; outputs from the PARC module will have had the prerequisite processing).
2. Background Point Count: This is an optional value that applies only to workflows that employ the Maxent modeling package. The dialogue box provides the option of specifying a number of background points to be randomly scattered throughout the study area (the extent of the template layer) to capture a more complete sample of the range of values present for each predictor layer. These points will be added to the field data .csv file with a value of "-999" denoting them as background points.
3. Background Probability Surface: This is an optional parameter that applies only to workflows that employ the Maxent modeling package. In some analyses, 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 1. 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 "1" will be retained. A point falling in an area with a value of "0.5" will be kept as a background point 50% of the time.
4. Field Data: 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.). This input file must be in a particular format, and in most cases, a user should populate this field by connecting a FieldData element to the MDSBuilder in the visual display within VisTrails. Please see the documentation for the FieldData module for more details.

### 

### ModelSelectionSplit and 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. Three parameters can be set by the user for each of these modules:

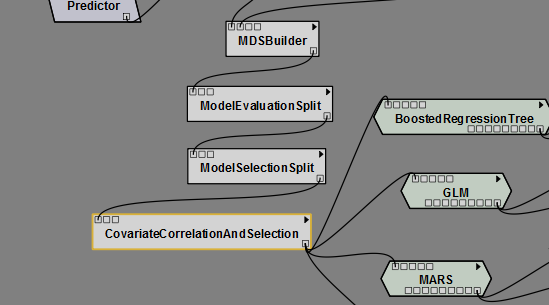
1. Ratio of Presence/Absence Points: This optional field is populated with a number corresponding to the desired ratio of presence to absence points to be used in the analysis. If not populated then all occurrence records (not background points) will be portioned into either the test or training split 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).
2. Input Merged Data Set (MDS): This is the input data set consisting of location data for each sample point, the values of each predictor variable at those points, and if established, a field denoting the weight that will be assigned to each point in modeling. 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.
3. Training Proportion: 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).

### 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 acknowledgement to the total number of models that were considered and the range of the evaluation metrics. This module ignores background points. Three parameters can be set by the user:

1. Stratify: Whether to stratify the folds by the response the default is to true
2. Input Merged Data Set (MDS): This is the input data set consisting of locational data for each sample point, the values of each predictor variable at those points, and if established, a field denoting the weight that will be assigned to each point in modeling. 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.
3. nFolds: The number of folds into which the data should be partitioned. The default is 10. 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).

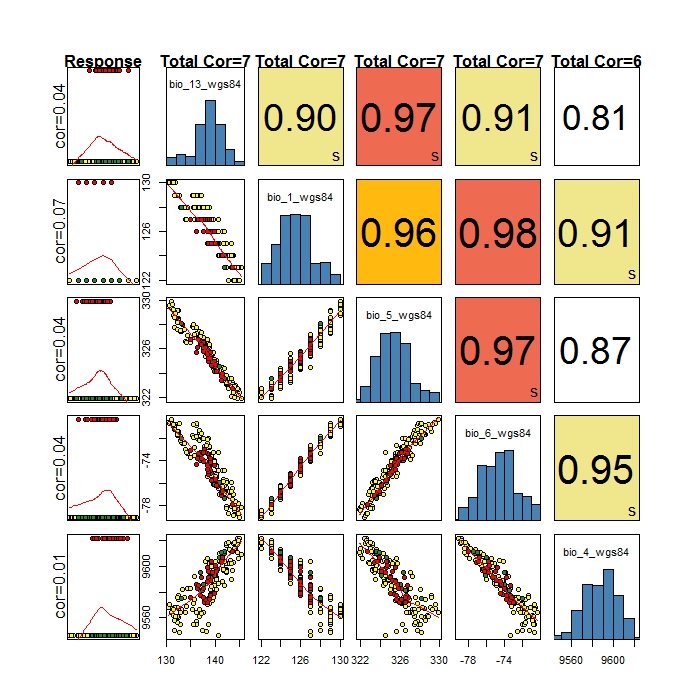
****

### 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 squre 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. Three parameters can/must be specified by the user:

1. ShowGUI : This Boolean indicates whether to stop execution and display the GUI for user interaction. In somecases 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 be apply to subsequent runs without interrupting execution.
2. corsWithHighest: 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.
3. inputMDS : 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.
4. selectionName: 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.

## **Modeling**

Currently five models are available in the SAHM package. Maxent is used for modeling presence only data. Generalized Linear Model (GLM), Multivariate Adaptive Regression Splines (MARS), Boosted Regression Tree and Random Forest are appropriate for fitting models to presence/absence data as well as count data all of these except Random Forest work under the assumption that a count response can be modeled as Poisson. Random Forest is less restrictive. GLM, MARS, Boosted Regression Tree, and Random Forest are all implemented in R and have similar options and input requirements and produce similar output. These will be discussed together while highlighting their similarities and differences. Maxent was written in Java and has quite different requirements and will be discussed separately.

### Input Required by GLM, MARS, Boosted Regression Tree, Random Forest

The following input values are applicable to all four R modules

1. mdsFile: 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 file is almost always generated by the upstream steps. The following information regarding the file contents is fairly technical and can be ignored by most users. Several names are reserved and have special meaning in the model fitting algorithms and not be used as predictors. These names appear in the first row of the csv.

* X: the x location of the observation (generally the first column)
* Y: they y location of the observation (generally the second column)
* Weights: site weights for each observation generally set using the FieldDataAggregateAndWeight module these are always ignored by Random Forest
* EvalSplit: the split used for final model evaluation generally set using the ModelEvaluationSplit module
* Split: the model selection split set using either the ModelSelectionCrossValidation or the ModelSelectionSplit module

All other labels in the first row should be the names of the response and predictors. Any categorical predictors should have names followed by the character string “\_categorical”.

The second row of the CSV includes some information on the steps taken previously in the work flow but should have a value of 0 or 1 for each predictor. 0 indicates that the predictor should not be used in the model fitting algorithm 1 indicates that it should be considered. These values are generally set using the CovariateCorrelationAndSelection but manually changing these values before running the models is also possible.

The third row of the csv contains more information on previous model fitting steps and if maps of the response surface are to be generated it must include a path to the geotiff for each predictor to be included in the model. All other rows should be filled with observations at each point.

1. MakeProbabilityMap: Indicate whether a probability map is to be produced for the model fit.
2. makeBinMap: Indicate whether to discritize the continues probability map into presence absence see the ThresholdOptimizationMethod for how this is done.
3. makeMESMap: 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.
4. ThresholdOptimizationMethod: determines how the threshold is set between to discriztize continuous predictions into binary for evaluation metrics calculated based on the confusion matrix as well as for the binary map. The options, directly from the PresenceAbsences package in R are:

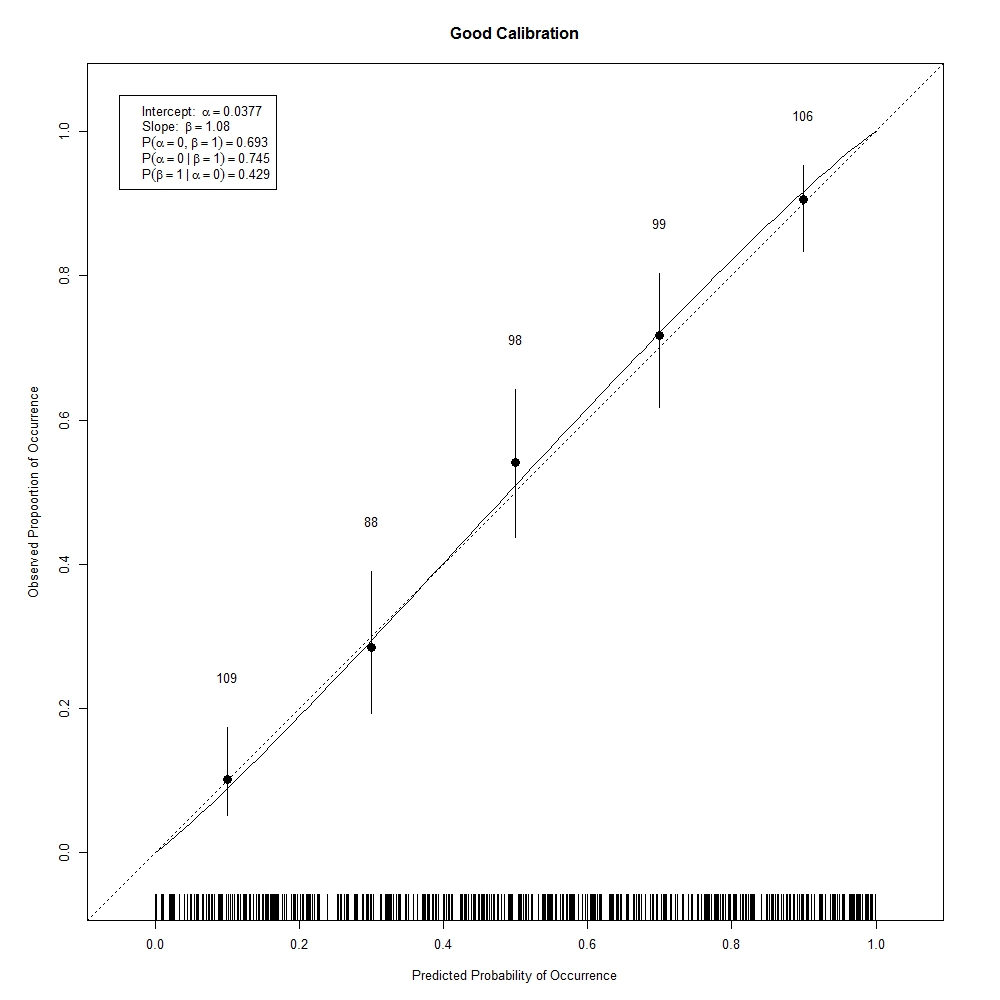
* 1: Threshold=0.5
* 2: Sens=Spec sensitivity=specificity
* 3: MaxSens+Spec maximizes (sensitivity+specificity)/2
* 4: MaxKappa maximizes Kappa
* 5: MaxPCC maximizes PCC (percent correctly classified)
* 6: PredPrev=Obs predicted prevalence=observed prevalence
* 7: ObsPrev threshold=observed prevalence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

This is set to 2 by default so it uses the threshold that results in the same sensitivity and specificity as nearly as possible. 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 each combination of folds.

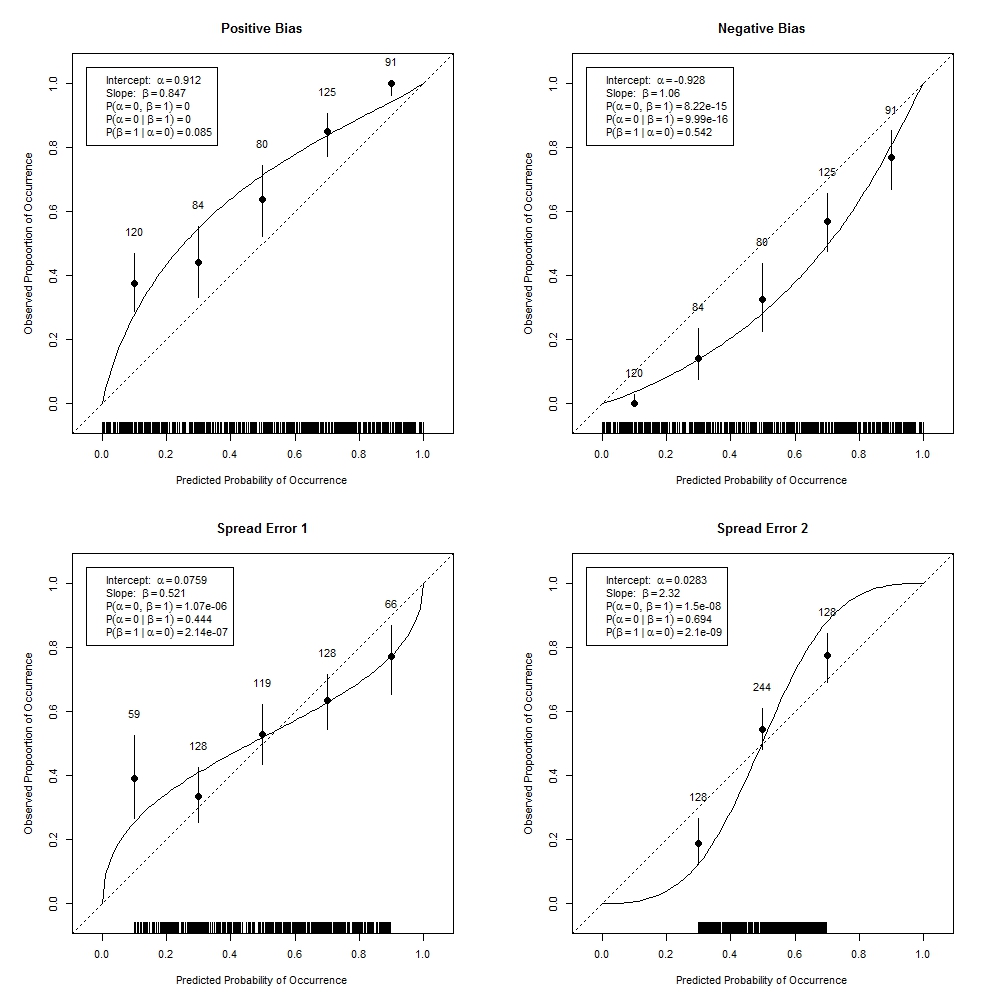
### Output Produced by All models

Several files should be produced by a successful model run while an unsuccessful model run will produce a log file which will indicate the issue that caused execution to halt and possibly some other files depending on where in the workflow execution was halted. When a model is successful the following files are produce in a folder in the vistrails workspace:

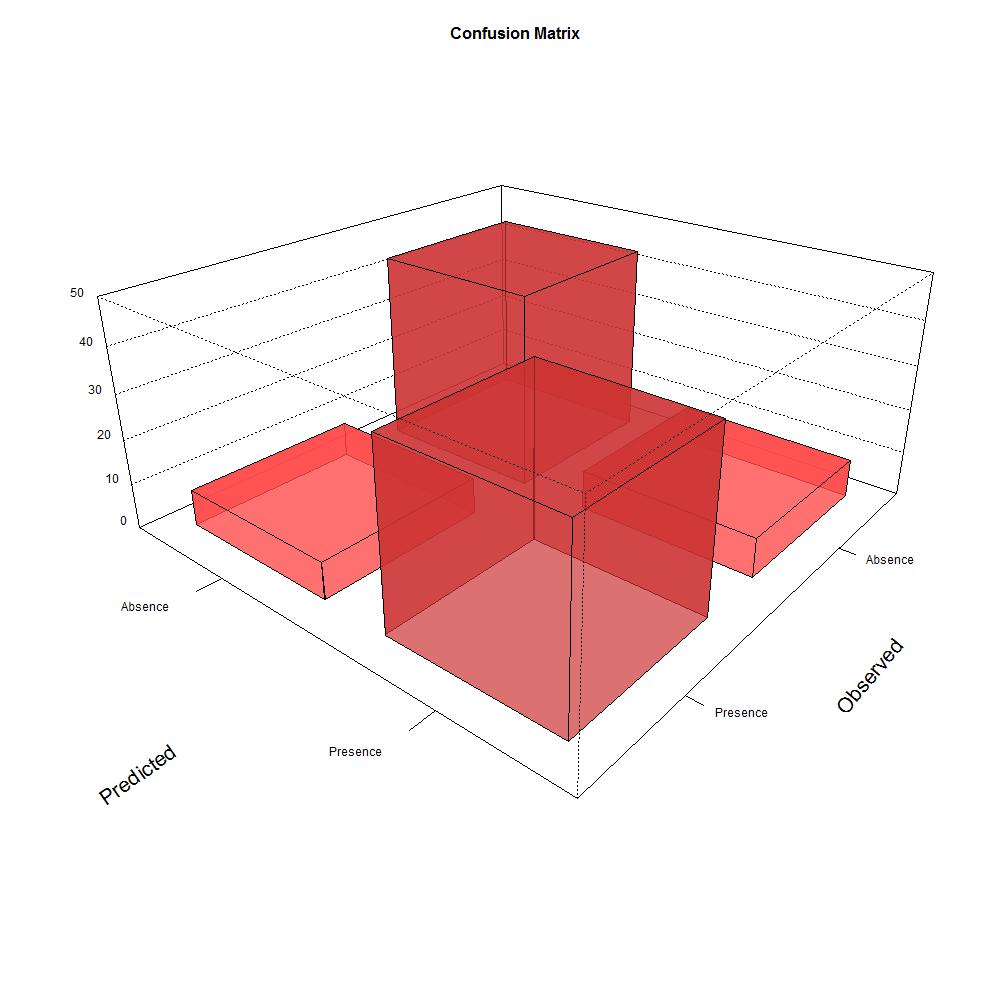
1. “model”\_output.txt : This file contains a summary of the model fit. The information contained here includes the number of presence observations (counts equal to or greater than 1 for count models), the number of absence points, the number of covariates that were considered by the model selection 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.
2. “model”\_modelEvalPlot.jpg : For binary data this will be a Receiver operating characteristic curve. Which shows the relationship between sensitivity and specificity as 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 a blue region will show the standard deviation for the cross validation folds. If the model fits well both sensitivity and specificity should be well above the diagonal line. If there is a strong disparity between the curves for the training data and either the testing split or cross validation standard deviation curves this can be indicative of model overfitting. These plots and the evaluation metrics based on the confusion matrix describe the models ability to discriminate between presence and absence points. The AUC value, or area under the ROC curve, is the probability that the model will rank a randomly chosen presence observation higher than a randomly chosen absence observation. For count data this display will show several standard plots for assessment of model residuals.
3. “model”\_CalibrationPlot: Calibration plots and statistics describe the goodness-of-fit between the predicted values and the actual observations. These are especially usefull for identifying problems with overfitting or underfitting when separate data is used for model fitting and model evaluation. These plots and statistics can be used to determine how reliably the model will predict if a site is occupied or unoccupied (Pearce and Ferrier 2000). 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. This line is a logistic curve because we are not using the logit transform of the predicted and observed. The intercept and slope of this line should be 0 and 1 respectively. Test statistics are reported and significant values indicate that the predicted values have a different mean or spread than the observed data. The five plots below show several different patterns that can be identified using calibration plots and statistics. Currently calibration plots are only produced for presence absence models.

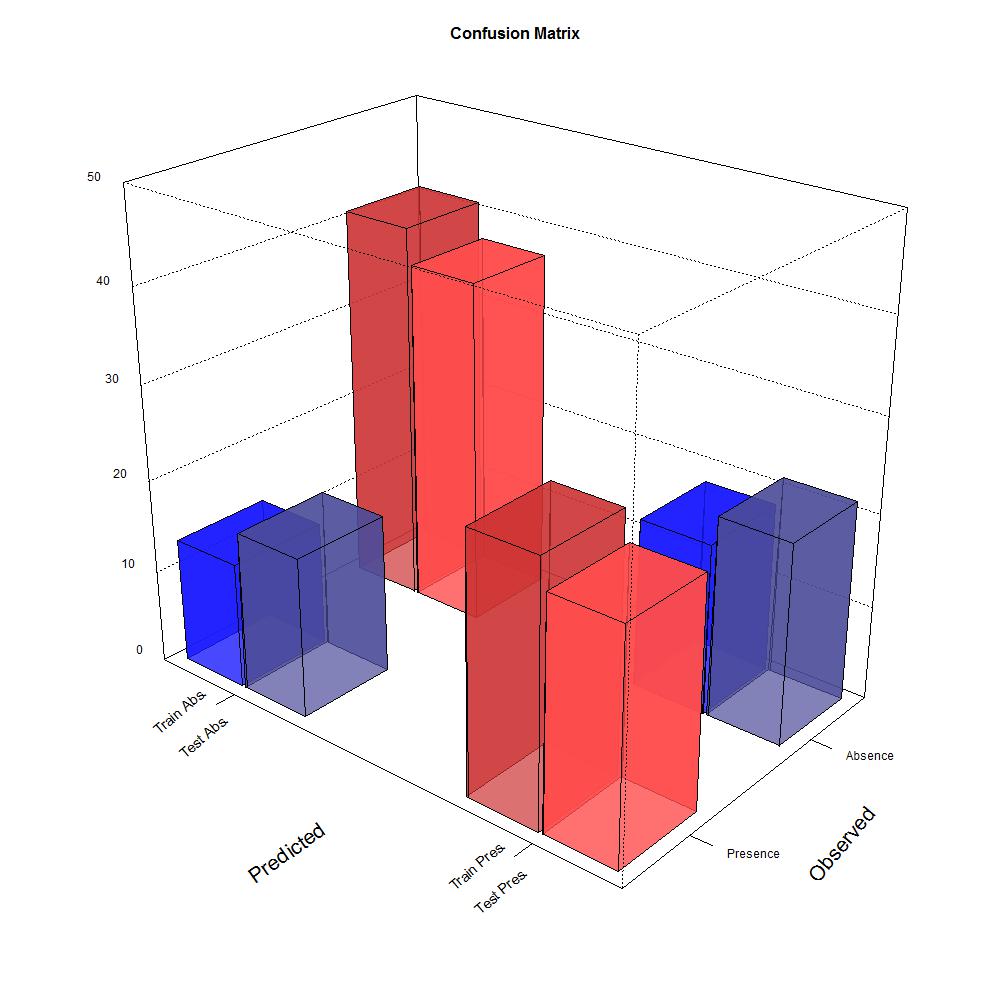


1. A calibration plot showing good calibration. Note that the logit curve follows the diagonal quite closely. The intercept is not significantly different than zero and the slope is not significantly different than 1.



1. These plots show poorly calibrated models. Positive bias occurs when the predicted values are higher than the observed probabilities of occurrence. Negative bias occurs when the predicted probababilities of occurrence are lower than the observed probabilities. Spread error 1 indicates that predicted probabilities of occurrence are higher than than they should be .

1. “model”.confusion.matrix.jpg: The confusion matrix shows the percent of predicted and observed values in each of the presence and absence classes. For predicted values this is based on the threshold used to discritize the predicted values. If a test\training split or cross validation was specified then the percentages for the training split and for the test or total for each evaluation set fold will be shown in the same plot. Several evaluation metrics are based on the discritization of the continuous predictions and could be seen as accompanying this plot. These include the percent correctly classified, sensitivity, specificity, Cohen’s kappa, and the true skill statistic. The calibration plot and related statisticics are only reported for presence absence models. 



1. Confusion matricies for a data with no evalutation split and with a test\training split. A good model fit will have relatively high percentages in the red bars indicating that the predicted presence and predicted absence of observations agree well with the observed values of presence and absence and thus the model has good discrimination. Large discrepancies between the training and evaluation data in this plot could indicate model overfitting.
2. “model”\_response\_curves.pdf: 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.
3. “model”.resid.plot: 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. Whether or not a significant spatial pattern exists in model residuals can at times be difficult to assess visually. We hope to add correlograms of Moran’s I soon. Unfortunately statistical tests based on the Moran’s I statistic for residuals of binary response models lack statistical justification and thus cannot be used to test for a significant spatial pattern (Bivand 2008). See Dormann 2007 for more discussion on evaluation of model residuals and spatial models that are appropriate for species distribution modeling. Residual plots can also be used to determine if certain observations contribute disproportionately to the deviance of the fitted model. For a binary response model deviance residuals with absolute values greater than 2 can be indicative of a problem.
4. “model”\_prob\_map.tif: 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.
5. “model”\_bin\_map.tif: If specified using MakeBinaryMap=TRUE then a surface of binary observations 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.
6. “model”\_MESS\_map and “model\_Mod\_Map: 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. The MoD map is related and indicates which variable was furthest from the range over which the model was. The MESS map takes the minimum value of a statistic calculate for each predictor and thus cannot diagnose hidden extrapolation as one might do using a hat matrix. This surface is only calculated for variables that are selected in the model selection step within each model fitting algorithm so that variables that do not significantly affect the occurrence of the organism over the range of the training data will not be included in the MESS map even though these predictors might be significant outside the range in which the model was fit. Random Forest never drops predictors so if one wishes to compare the MESS and Mod map before and after insignificant predictors were dropped, one can compare the MESS map of a Random Forest fit to that produced from the other model fit algorithms as long as they were fit using the same dataset. See Elith et. al. 2010 for details on how the MESS map calculations are performed.
7. Evaluation metrics appended to AppendOutput .csv and .jpg: An appended output csv is produced to track several evaluation metrics across model runs and if at least two models have been run then a .jpg will accompany this. The name of the csv indicates the type of response as well as the type of model selection split that was specified and separate .csv’s and .jpg’s will be produced for each combination so that a folder might contain any combination of:

* AcrossModelNoSplitBinom
* AcrossModelNoSplitCount
* AcrossModelTestTrainBinom
* AcrossModelTestTrainCount
* AcrossModelCrossValBinom
* AcrossModelCrossValCount

This plot will be used to select and evaluate the final models and is discussed in more detail in the Select and Test the Final Model section.

### Generalized Linear Model (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.

One options is available for fitting GLMs:

* Simplification Method: This can be set either to AIC or BIC and alters the decision rule governing how the model is pruned in the stepwise model selection step.

### Multivariate Adaptive Regression Splines (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. To optional parameters can be set:

* MarsDegree: 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.
* MarsPenalty: the cost per degree of freedom charge in fitting the mars model the default is 2. (from the mda library)

### RandomForest

Random forest is a machine learning ensemble classifier. Numerous (in our case, 1,000) decision trees are computed using random subsets of the covariates. Each tree gets one vote, and whichever class gets the best vote “wins”. The relative importance of each covariate is assessed by the change in a fit statistic, on average, for trees that include it. Random forest models automatically model interactions and nonlinear relationships. According to Breiman, cross-validation or separation of data into test\training splits is not required with Random Forest because internally random forest uses a bootstrap sample to fit individual trees observations not used to fit a given tree are called out-of-bag. Predictions can then be calculated for each observation using only trees for which the observation was out-of-bag (OOB). All evaluation metrics for the training data are based on OOB predictions and thus should be similar to the results from applying evaluation metrics to independent test data. The OOB confusion matrix does not agree with the training data confusion matrix because the threshold for the former is based on wining OOB votes while the later has a threshold based on the threshold optimization method selected by the user. Several parameters can be set for random Forest these are discussed in detail and their defaults are listed in the randomForest function documentation in the randomForest reference manual (<http://cran.r-project.org/web/packages/randomForest/index.html>). Here I will only mention a few slight differences from the standard parameters and defaults.

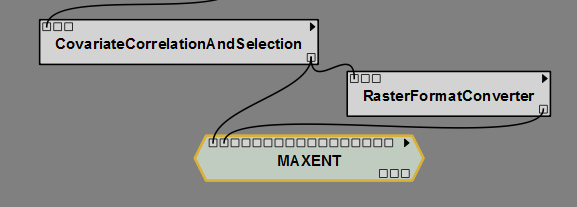
* mtry: by default this is optimized using the tuneRF function rather so that OOB error is minimized.
* Seed: The random number seed used by randomForest. If one desires to reproduce results from a previous BRT fit, one must enter the random number seed that is reported in the textual output from that model fit. A seed is randomly selected by default.

### BoostedRegressionTree (BRT)

BRT is also based on numerous decision trees. 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, and 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 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. This is a summary of the descriptions found in Elith and Leathwich 2008 and more detail can be found in this reference.

* Alpha: 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. The default value is 1.
* BagFraction: 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). The default value is .75.
* LearningRate: controls the amount each tree contributes to the model. A small learning rate restricts individual tree contributions to the overall model. If not specified, learning rate will be determined based on the number of trees and the tree complexity.
* MaximumTrees: The absolute upper limit on the total number of tress to fit. Setting this below 5000 will result in an error. The default is 10,000.
* NumberOfFolds: if cross-validation is used for model simplification, this sets the number of folds used for cross-validation. The default is 3.
* TreeComplexity: 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. If not set, tree complexity will be selected based on the number of observations and what produces the best model.
* PrevalenceStratify: can be set to TRUE or FALSE. 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. Set to TRUE by default.
* Seed: 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. A seed is randomly selected by default.
* Tolerance: 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. The default is .001.
* ToleranceMethod: Method used in determining when to stop model simplification. This can be set to “auto” or “fixed”. The default is “auto”.

****

### Maxent

Catherine to write.

### Select and Test the Final Model

If multiple models are run in the same folder then model output is complied across models into a jpg and a csv. The select and test final model option (Packages drop down, SAHM, Select and Test the final model) can then be used to select from among the fitted models, evaluate these models on the data withheld using the model evaluation split, consolidate evaluation metrics for these best performing models, produce probability surfaces, binary surfaces, MESS and MoD maps. Once Select and Test the Final model has been chosen the user will be given the choice between any Appended Output files that were created for different combinations of model selection splits and response types. A graphic will pop up with the appended output results with the best performing models being marked with their numeric values for each metric. x labels indicate the folder name where all output for each model can be found. Folder output will be left blank for models that failed at some point during execution. If a test\train split was used for model selection evaluation metrics will be shown for each as bars and if a cross validation split was used then barplots showing results from each fold will be shown. Once the desired models are selected then the requested output will be produced for the best performing models.

## Viewing Output

### Spreadsheet

VisTrails has a spreadsheet to view individual outputs side by side. Besides the built in cells for viewing text, webpages, etc SAHM has two specialized cells for viewing model outputs. The output from several model runs can be viewed side by side in individual spreadsheet cells.



****

### SAHM Model Output Viewer Cell

This widget is for viewing the textual and graph output from R. It contains tabs for the model output report, response curves, AUC plot, and calibration graph, confusion matrix and map of residuals. Within any of the graphs you can zoom in or out with the mouse wheel and pan by moving the mouse while holding down the left mouse.



### SAHM Spatial Output Viewer Cell

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 presense 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

**Spatial Output Viewer Toolbar:**

When panning you can zoom by holding a right click and moving the mouse up for zoom out or down for zoom in.

When using the Zoom rectangle a right click rectangle selection will zoom in, a left will zoom out.

Reference, 2000, National water quality inventory—1998 report, U.S. Environmental Protection Agency Report EPA–841–F–00–006, 45 p.

# Glossary

Glossary Term Glossary definition.

References

A. Baddeley and R. Turner (2005). Spatstat: an R package for analyzing spatial point patterns. Journal of Statistical Software 12 (6), 1-42. ISSN: 1548-7660. URL: www.jstatsoft.org

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

Dormann CF, McPherson JM, Araujo MB, Bivand R, Bolliger J, et al. 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30:609–28

*Elith, J.*, *Leathwick, J.R.* & *Hastie, T.* (*2008*) *A working guide to boosted regression trees*. Journal of Animal Ecology, *77*, *802*–*813*.

Elith J, Kearney M, Phillips S (2010) The art of modeling range-shifting species. Methods Ecol Evol 1:330–342

Freeman, Elizabeth (2007). PresenceAbsence: An R Package for Presence-Absence Model

Evaluation. USDA Forest Service, Rocky Mountain Research Station, 507 25th street,

Ogden, UT, USA.

Grosjean, Ph. (2011). SciViews: A GUI API for R. UMONS, Mons, Belgium. URL

http://www.sciviews.org/SciViews-R.

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.

S original by Trevor Hastie & Robert Tibshirani. Original R port by Friedrich Leisch,

Kurt Hornik and Brian D. Ripley. (2011). mda: Mixture and flexible discriminant

analysis. R package version 0.4-2. <http://CRAN.R-project.org/package=mda>

Robert J. Hijmans & Jacob van Etten (2011). raster: Geographic analysis and modeling

with raster data. R package version 1.9-41. <http://CRAN.R-project.org/package=raster>

Timothy H. Keitt, Roger Bivand, Edzer Pebesma and Barry Rowlingson (2011). rgdal:

Bindings for the Geospatial Data Abstraction Library. R package version 0.7-4.

<http://CRAN.R-project.org/package=rgdal>

Leathwick JR, 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

A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News

2(3), 18--22.

*Miller ME, Hui SL, Tierney WM. 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 133:225–245.

Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5

(2), <http://cran.r-project.org/doc/Rnews/>.

*R Development Core Team (2011). R: A language and environment for statistical*

*computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN*

*3-900051-07-0, URL* <http://www.R-project.org/>*.*

*Greg Ridgeway (2010). gbm: Generalized Boosted Regression Models. R package version*

*1.6-3.1. http://CRAN.R-project.org/package=gbm*

*Terry Therneau and original Splus->R port by Thomas Lumley (2011). survival: Survival*

*analysis, including penalised likelihood.. R package version 2.36-10.*

*http://CRAN.R-project.org/package=survival*

Appendix A: SAHM VisTrails Tutorial, modeling cheatgrass in Rocky Mountain Park.