Spatial Data Modeller

Don Sawatzky[1], Gary Raines[1], and Graeme Bonham-Carter[2]
2008

1. U.S. Geological Survey (retired)
2. Geological Survey of Canada (retired)

Contact Gary Raines at garyraines@earthlink.net with questions.

Introduction

Spatial Data Modeller, SDM, is a collection of tools for adding categorical maps with interval, ordinal, or ratio scale maps to produce a predictive map of where something of interest is likely to occur. **Some critical tools will not work correctly with ArcGIS 9.3. A version for ArcGIS 9.3 is in preparation (Nov., 2008).The SDM errors in ArcGIS 9.3 have to do with changes in Band Collection Statistics and intersection of points in rasters. These errors can be subtle; so tools such as Calculate Weights, Area Frequency, NN Input Files, MS Large and MS Small compute but do not provide the correct values in some situations.** All of the tools have help files that include references to publications discussing the applications of the methods implemented in the tool. Several of the tools create output rasters, tables, or files that require the user to enter a name. Default values are provided in most cases to serve as suggestions of the style of naming that has been found useful. These names, following ArcGIS conventions, can be changed to meet the user's needs. To make all of the features of SDM work properly it is required that several Environment parameters are set. See the discussion of Environment Settings below for the details. The Weights of Evidence, WofE, and Logistic Regression, LR, tools addresses area as the count of unit cells. It is assumed in the WofE and LR tools that the data has spatial units of meters. If your data has other spatial units, these WofE and LR tools may not work properly.

All of the tools can be used interactively from the ArcGIS Desktop. Most of the tools can be used in models as a mechanism for repeated running and documentation of the model. Demonstration models are included that utilize the Carlin data, which is provided in the Carlin_Demo folder in the SDM folder. There is a Carlin_Demo.mxd in the Carlin_Demo folder. Saving this MXD to the users Workspace folder and changing the Environment to have the users Workspace and Scratch Workspace folder provides a full setup for testing the demonstration models and tools. These demonstration models help document the applications of the tools and serve for self-training. A demonstration tool, NN Tool, is designed to run part of the neural net process and may serve as a useful general use tool for applying neural networks.

SDM requires the Spatial Analyst extension, which is used to prepare data for use in SDM models.

For the computations to work correctly in SDM, the Regional Settings of the computer must be set to English (United States). This is so the point (.) will represent the decimal point, not a comma (,).

Evidential Rasters

The layers or maps used for evidence for all of the methods except Fuzzy Logic must be integer rasters. Fuzzy Logic can work with real valued, positive rasters. The WofE and LR tools use a DBF table that is jointed to the evidential layers; so this requires that evidential layers must be integer rasters. The Neural Net tools use the Combine tool to generate the unique combinations of the evidence rasters, and Combine only works as desired with integer rasters for inputs.

Guidance on Projections

All layers used with SDM should have the same projection with meter units.

The WofE, LR, and Neural Net methods require that point Shapefiles (the training points) be intersected with the evidence rasters to generate tables needed for training. In addition, the Area Frequency tool, which can be applied to all the model posterior probability or membership response themes, intersects point Shapefiles with the response theme. These tools as implemented in SDM do not deal with projection on the fly. **Therefore, it is essential that all evidence rasters, training Shapefiles, and the study-area rasters have the same projections for these three methods.** Projection on the fly in ArcGIS determines whether a layer requires projection on the fly based on the name of the projection. Consequently, **it is essential that all the projection names be the same.**

With some projections such as the Finland Zonal projections and some of the Australian national projections, the spatial reference or name of the projection of the grid can change during processing because of the nature of how ESRI handles projections in the Grid Engine. GRIDs use the workstation format for the coordinate system information, which is not as flexible or complete as the projection engine for Shapefiles (desktop, features, etc).

Projection names are lost except for UTM, State Plane, British National Grid, New Zealand Map Grid, RSO, UPS. Everything else will lose its name, and end up with the map projection as part of the name. The properties (eastings, northings, etc) of the changed projection are all the same, but the name is changed. This name change causes projection on the fly to be done. Projection on the fly causes problems with many of the raster geoprocessing tools including those of SDM.

In addition for WofE and LR methods and any use of the Area Frequency tool, SDM assumes that the units of the projection are meters. These tools all make measurements of area. This limitation is due to the historical development of these tools and limitations of programming time.

One way to assure that all the projections are the same is to use the following process while creating the layers to be used by SDM.

1. Create an empty Data Frame and define the projection of this empty Data Frame the desired projection to be used in the modeling process.
2. Set the following Environmental parameters
    a. General Tab
        i. Workspace
        ii. Scratch Workspace
        iii. Output Coordinate System – Select an ESRI provided metric projection. For WofE and LR, SDM assumes the units of all the layers are meters.
        iv. Extent – This must be set later when the study area layer has been created
    b. Raster Analysis Settings

   i. Cell Size – This must be set for SDM, but it might not be need at this stage in project development
   ii. Mask – This must be set when the study area layer has been created.
3. Create a Shapefile that defines the Study Area. Create it with the projection desired for the study. It is simpler if you use a predefined ArcGIS projection.
   a. Convert the Study-Area Shapefile to an integer raster with the cell size of the smallest cell size to be used in the SDM project.
   b. Set the Mask and Cell Size in the Raster Analysis Settings of the Environment.
   c. It is a good practice to set the Cell Size the same as that of the Study Area.
4. Convert the vector layers to be used as evidence to integer rasters. During this process, some like to create the integer rasters to the same cell size, the smallest to be used in the SDM project as defined by at least one of the evidence rasters. It is not necessary that all the evidence rasters have the same cell size, so if computation time and data volume is an issue in your study, you might select a cell size appropriate to each evidence rasters.
5. In some of your evidence is already rasters, it is necessary to assure that their projections are appropriate.
   a. For ordered rasters, one way to assure that the projection is the desired projection is in this MXD with the Output Projection set to the desired projection in the Environment, use the Int Spatial-Analyst geoprocessing tool to create a new integer raster from your source raster. This process will also clip the new raster to the Study-Area mask.
   b. For categorical or free rasters similar to the process for ordered rasters, use the Lookup (sa) tool to create a new integer raster that maintains the categorical attribute. As with the Int process above, this will clip the new raster to the Study-Area mask.
   c. Any Spatial-Analyst tool that creates a new raster can be used in this MXD with the Output set to the desired projection in the Environment.
6. Assemble the training points into Shapefiles. Often they are in files in other formats, other projections, or with other points.
   a. One way to create the desired Shapefiles is to get the source files into your MXD, select the subset desired, and Export the selected points to a Shapefile in the projection of the Data Frame.
   b. Any process that creates a point Shapefile with the correct projection can be used.
7. The following process will provide a test that everything has been done correctly:
   a. Create a new Data Frame. Do not define the projection.
   b. Add one of the layers to be used, such as the Study-Area raster. This will define the projection of Data Frame.
   c. Add the other desired layers. If any of these layers have a different projection, then ArcGIS will issue a warning about it having a different projection. If this warning occurs, then something was done incorrectly in the above steps.

   d. One common way that causes the warning to be issued is that the name of the projection is somehow different. Also, if a different datum is used, this warning can occur. To avoid this error always use ESRI provided projections and their associated names.
   e. The projection names can also be checked in the metadata using ArcCatalog or viewing the metadata in ArcGIS.
   f. Experienced users will, of course, see many alternative ways to deal with projections. The guidance provided above is one of many ways to prepare the data.
8. If some of the layers in the modeling still have different projection names, but all of the projection parameters are the same, the final fix is to change the projections in ArcCatalog. Using ArcCatalog, the properties of layers with different named projections, can be changed to the standard to be used in the modeling. Typically, this will be necessary to assure that the projection of the point Shapefiles used for training or validation have the same projection as the raster layers. This process does not change the projections, but simply gets the names the same because the parameters are already the same.

Citation of SDM

In citing this software, please use a reference in the following form:

Sawatzky, D.L., Raines, G.L. , Bonham-Carter, G.F., and Looney, C.G., 2008, Spatial Data Modeller (SDM): ArcMAP 9.2 geoprocessing tools for spatial data modelling using weights of evidence, logistic regression, fuzzy logic and neural networks. http://arcscripts.esri.com/details.asp?dbid=15341.

Acknowledgements

This software has evolved through several generations, which have had input from several people and organizations. The development of this version was supported with funds from the U.S. Geological Survey. The evolution of this software starts with Frits Agterberg whose work at the Geological Survey of Canada set the foundation for much of the mathematics implemented in these tools. More recently, Ryan DeBruyn of ESRI has made significant contributions in solving specific programming problems in ArcGIS. We would also like to thank Vesa Nykanen of the Geological Survey of Finland and Carlos de Souza Filho of the University of Campinas, Brazil, for beta testing this software and many useful suggestions for improvements. In addition, we thank Steve Kopp of ESRI for stimulating us to create this Geoprocessing version of SDM and providing support from ESRI to answer questions and solve problems during the development.

Standard Software Disclaimer

All of our software is covered by this disclaimer:

While the *producers of Spatial Data Modeller Tools* make every effort to deliver high quality products, we do not guarantee that our products are free from defects. Our software is provided “as is," and you use the software at your own risk.

We make no warranties as to performance, merchantability, fitness for a particular purpose, or any other warranties whether expressed or implied.

No oral or written communication from or information provided by the *producers of Spatial Data Modeller Tools* shall create a warranty.

Under no circumstances shall the *producers of Spatial Data Modeller Tools* be liable for direct, indirect, special, incidental, or consequential damages resulting from the use, misuse, or inability to use this software, even if the *producers of Spatial Data Modeller Tools* has been advised of the possibility of such damages.

These exclusions and limitations may not apply in all jurisdictions. You may have additional rights and some of these limitations may not apply to you.

Installation

Before installing a new version of Spatial Data Modeller, it is best to delete the SDM folder from your computer. Installation simply requires extraction of everything in the zip file. The zip-file contents contain a relative path; so everything will extract to a folder named SDM, Spatial Data Modeller. The SDM folder can be extracted to any location on your computer. A typical place would be to put the SDM folder in the root directory, typically C:\.

The SDM folder contains two toolboxes and their associated scripts and documentation, Spatial Data Modeller Toolbox and Spatial Data Modeller Demonstration Models. The Spatial Data Modeller Toolbox contains tools sets for Fuzzy Logic, Neural Networks, Weights of Evidence, and Utilities. The Spatial Data Modeller Demonstration Models contains example models of Fuzzy Logic, Neural Networks, Logistic Regression, and Weights of Evidence. These models complement the documentation to demonstrate how the tools may be used. These demonstration models do not show, however, the complete spectrum of applications of the tools.

The SDM folder contains, also, the Carlin_Demo folder, which contains data and an map document for use in the Demonstration Models and further experimentation with the tools. These data are useful for training.

Also in the SDM folder are three LYR files useful for symbolization of models. The names of these files indicate their application. See the relevant demonstration model to see applications of these files. See the Known Issues section below for a bug associated with the use of the LYR files.

Options Settings

When using these tools it is useful to change some default settings of ArcGIS. In the Tools/Options window select the Geoprocessing tab. Check the "Overwrite the outputs of Geoprocessing operations" and "Log Geoprocessing operations to a history model". The overwrite option is useful because it allows experimentation without changing the names of outputs. The Log is useful to provide a history of what has been done. See the ESRI hyperlink below for more information on the use of the Log tools, which are shown on the Results tab in ArcGIS.

http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?id=624&pid=618&topicname=Working_with_results

SDM now takes advantage of this geoprocessing history to report and record as messages important modeling parameters. These parameters include the environment setting, the study area, prior probability, an overall CI value from Agterberg-Cheng and other important information.

ArcGIS by default limits the number of records in a raster to 65,536. This limit is often too small with the neural net tools where combination rasters of evidence are created. In the Tools/Options window, select the Raster tab. The limit of 65,536 records can be increased to a value more appropriate.

Environment Settings

The Environmental Settings can also be set of this page. Setting them in the Tool Options makes these Environmental Settings the defaults. Setting the Environment by right clicking on ArcToolbox sets the environment for an MXD. As discussed in the SDM tools documentation, the Workspace, Scratch Workspace, Extent, Cell size, and Mask must be set for many of the SDM tools to work properly. Set the Cell size to that of the Mask. It is also useful to set the Output coordinate systems (General Tab) the same as the metric coordinate system used for the modeling. Note SDM does not support geodatabases; so the Workspace nor Scratch Workspace cannot be a geodatabase.

Language Settings

These tools require that the Language settings have the period for the decimal symbol, not the comma that is common outside of the United States. This feature is controlled from the Regional and Language Options in the Control Panel. On the Regional tab, select Customize to verify that the Decimal symbol is a period.

Symbolization

A standard way to symbolize many of the response themes, such as the Weights-of-Evidence posterior probability raster, is with increasing warm colors indicating a high value. This can be achieved by selecting the full-spectrum red-green-blue color ramp available in the ESRI styles and then inverting this ramp. An alternative is provided in the SDM.style file in the SDM folder. This style has a full spectrum bright color ramp; so the additional step of inverting is unnecessary. This style can be made available for symbolization from the ArcGIS tool menu, Styles, and selecting the Style Manager. Adding the SDM style to your list makes this style available. It is convenient to copy the

Full Spectrum color ramp from this added style and paste it into the Color Ramp folder on your User default style. This can all be done in the Style Manager. This new blue-green-red full-spectrum color ramp will be available after the ESRI red-green-blue ramp.

Enhancements

Release March 3, 2008

New features and enhancements have been added to this release of Spatial Data Modeller.

1. SDM now takes advantage of the geoprocessing history to report and record as messages important modeling parameters. These parameters include the environment setting, the study area, prior probability, an overall CI value from Agterberg-Cheng and other important information.
2. Documentation for many of the tools has been clarified or enhanced based on feedback from users.
3. Calculate Weights: A new method, Unique, has been added to the list of methods. This method will create a weights table with the Gen_Class, Weights, and WStd columns empty. The user can then edit these columns for an Expert or Fuzzy Weights of Evidence. Remember the Gen_Class defines the classes in Logistic Regression and the Weights define the classes for Weights of Evidence.
4. It is no longer required that all evidence rasters have the same cell size as the Study Area. In order to get accurate weights, it is best that the Study Area raster still have a cell size of the smallest cell size of the evidence rasters.
5. Agterberg-Cheng CI Test: This tool has been modified to explicitly address areas of Missing Data.
6. Neural Net Input Files: A new method has been developed to create the Class and Train files that is much faster. Now a set of unique conditions as large as 100,000 can be processed in a few minutes.
7. Create Missing Data: This tool is useful to explicitly define the Missing Data Class for evidence. This is a Utility tool in SDM.
8. Bugs in the Area Frequency tool have been fixed.
9. A new utility tool, Contact Proximity, has been created to create an integer raster of distance to a contact between two raster units.
10. A new utility tool, View SDM Text Files, has been developed to facilitate viewing within ArcGIS of text and neural net DTA files.
11. The Gamma fuzzification tool has been added to Fuzzy Membership toolset. This tool is similar to the Near tool, but it has a slightly different shape.

Release April, 2008

New features and enhancements have been added to this release of Spatial Data Modeller.

1. The Unit Area default of 1 has been removed. It is now necessary to enter a value with all WofE and LR tools. It is critical that users select a Unit Area that is appropriate to their specific task. If a user wants to set a default value, this can be defined in the Script Properties on the Parameters tab.
2. Unit Area problem: It was discovered that the Unit Area was not being properly used in Calculate Weights, Calculate Response, Area Frequency, and Agterberg-

Cheng. This error occurred when the Unit Area had a value other than 1. To the best of our knowledge, this error occurred during the development of the ArcGIS 9.1version. This error has now been corrected. Results will now be correct for any Unit Area, and the results might be different than in earlier releases for ArcGIS 9.1 and 9.2.

3. Several new utility tools have been provided.
    a. Get SDM Parameters: This tool reports the Environment and other SDM parameters. This tool along with each tool reporting on the Environment and SDM parameters provide a mechanism for monitoring these parameters throughout the modeling process.
    b. View Floating Raster Vat: Tool lists the values of floating rasters
4. The WofE and LR tools have been revised to increase the speed and precision. Now probabilities less than 1 part in 1,000,000 will be properly handled.

Release June, 2008

1. Calculate Weights has been modified to deal with special cases of the number of training points. The modifications are summarized in the table and Notes below.

Table Type	No. TPs = 0	No. TPs = MaxNo. TPs	Other No. TPs	Gen_Class
Ascending[3]	Wts = Null[1]	Wts for MaxTPs – 0.01	Calc Wts[2]	Calc Wts
Descending[3]	Wts = Null	Wts for MaxTPs – 0.01	Calc Wts	Calc Wts
Categorical[4]	Wts = Null	Wts for MaxTPs – 0.01	Calc Wts	If Gen_Class = 99 with zero TPs, then Wts for TPs = 0.01; Otherwise Calc Wts

No. TPs = Number of Training Points
MaxNo. TPs = total training points available for the study area

Notes
Null is represented as a zero in a DBF table
Calc Wts indicates that weights are calculated with the conventional rules for the method using the training points that are found.

For the Cumulative Table types (Ascending and Descending), the conventional method is the maximum contrast with an acceptable confidence defines the break class between the two binary classes. The Gen_Class, Weight, and W_Std for those classes up to and including the break class are 2 and the Wplus and S_Wplus of the break class. For those classes after the break class, the values are 1 and the Wminus and S_Wminus of the break class.

In the case where all of the Training Points are all in one class, this class will be the break class and the number of training points will be the maximum number of training points minus 0.01

For the Categorical Table types, the conventional method is all classes that have acceptable absolute value of confidence are given the Gen_Class, Weight, and W_Std of their class value, Wplus, and S_Wplus. Those classes that do not have acceptable

absolute value of confidence are grouped into a Gen_Class of 99 or some other appropriate number if 99 is a valid class. If training points occur in this Gen_Class 99, then the weight is normally calculated for this group of classes. If, however, no training points occur in this Gen_Class 99, then the training points are 0.01 and the Weight and W_Std are calculated.

2. Grand WofE is a new tool that combines Calculate Weights, Response, and Logistic Regression into one simple tool requiring minimal input from the user. Besides making the weights of evidence and logistic-regression methods even simpler to apply than the individual tools, Grand WofE is designed to use in web applications. As with all of the WofE tools, the Environment must be set and the training Shapefile must be provided. Based on a user input of confidence, Ordered data are calculated using the Cumulative Ascending and Descending weighting methods. If in the rare case where an Ordered evidence raster produces a valid cumulative ascending and descending weights tables, multiple sets of WofE- and LR-models and associated tables will be created.
    a. Why not use Grand WofE for all weights-of-evidence modeling. If a standard model is what is desired, then use Grand WofE. There are situations where it might be of interest to use some other generalization procedure, such as a binary generalization of categorical data with some units treated as missing data. A generalization based on maximum confidence instead of maximum contrast might be desired. There might not be sufficient training sites and an expert weights-of-evidence model might be desired. If Logistic Regression was the method desired, it might be interesting to not generalize the ordered data. All of these reasons and others might be reasons to use the individual tools and not Grand WofE.
    b. Because Grand WofE has a variable number of outputs, it is necessary to add these outputs in a different way. When the tool is complete, instructions on add the outputs are provided in green text telling the user to copy some text with Control-C and then paste it into the Add Data names window with a Control-V. These instructions and the list of outputs to add are also available in the Results tab.
3. In Logistic Regression and Grand WofE, a required input is the data type. In the past, the inputs were Ordered and Free for Categorical data. In this release, the user can still enter these two types, or now the user can enter O or o, F or f. In addition, in place of Free, the user can now enter Categorical or C or c.
4. The WofE and LR tools have been revised to increase the speed and precision. Now probabilities less than 1 part in 1,000,000 will be properly handled.
5. Two new evidence rasters, Sbint and Fake_Data, have been added to the Carlin_Demo.mxd to demonstrate additional aspects of WofE tools.
    a. Sbint is an integer raster of the stream sediment antimony data that has not been reclassified. It is simply the parts per million values of the antimony data. This data set is to emphasize that it is not required nor always beneficial to reclassify the evidence. The reclassifications used in the Carlin_Demo are simply done to speed up processing for demonstration purposes.

   b. Fake_Data is an artificial data set created to demonstrate what happens in Grand WofE when valid weights tables are created by both the Cumulative Ascending (CA) and Cumulative Descending (CD) methods. Use this Fake_Data as evidence in Grand WofE with a confidence of 0.7 to demonstrate that two separate sets of models will be created. Each set of models will use one of the valid Ordered evidence rasters.
6. Comments about each evidence layer have been added to the Comments block in the General tab of the properties window for the rasters. This information should provide some guidance for the nongeologist about how to use the evidence in a WofE or Fuzzy Logic model.
7. In order to help address some problems users have had with projections, the projection must be defined in the Environment. On the General Tab, set the Output Coordinate System to be the same as the projection used for the evidence rasters. The Utility Get SDM Parameters can be used to test that all of the necessary environmental variables are changed from the defaults.

Release July, 2008

1. Two tools, Partition NNInput Files and Combine NNOutput Files, have been added to the neural net toolset to get around the 200,000 limit of GeoXplore.
2. The Get SDM Parameters tool now properly reports the Unit Area.
3. Training Site Reduction has an improved algorithm so it will process the Shapefile significantly faster.

Release Mid July, 2008

1. A new utility model has been added that is useful for proximity to lines. This model, Line Proximity, calculates a new raster of proximity to selected lines, that is linear features. The lines might be faults, fold axis, rivers, or any other type of line feature that proximity to is evidence for some process. This is distinct from Contact Proximity, which is proximity to a boundary in a raster representation of a polygon feature.
2. A new utility model has been added to test for where two symbolized response rasters have class difference. The tool is Rank Differences.
3. A new fuzzification model has been added that uses the classes defined by the symbolized classes in the table of contents. If for example, an evidence raster is symbolized with 5 classes, then this tool would produce a fuzzy membership raster with memberships from 0.2 to 1
4. The Add Bearings to Linear Features tool has been repaired.
5. Incomplete testing as of June, 2008, of the SDM tools in ArcGIS 9.3 indicates that the WofE and LR tools work correctly in ArcGIS 9.3. We believe all of the tools will work without modification in ArcGIS 9.3. Please notify us if you find problems with ArcGIS 9.3.

Release September, 2008

1. The TOC Fuzzification tool has been fixed so the fuzzy membership is now properly scaled from 0 to 1.

2. A new tool has been added to the Neural Network toolbox. The tool is called Color Composite. It is designed to be used to create a color composite from 3 patterns from the Fuzzy Neural Network. This is often a useful way to see combinations or mixtures of the memberships in three patterns.
3. The Carlin_Demo.MXD now has a hill shade raster of the DEM for the Carlin study area. By viewing a response theme with transparency and the hill shade, the response theme and hill shade can be seen together to determine locations. If it is desired to better maintain the colors of the response theme, view the hill shade on top with transparency and the response theme below.

Release November, 2008

1. An error has been found in the Neural Net Input Files tool. The error involves how the training sites are filtered so that only one training site is reported in the Train.dta file for a unique condition. The tool now filters the site so when there are multiple training Shapefile points in a unique condition, the point with the maximum fuzzy membership is kept. If there are multiple training sites with the same fuzzy membership then the first one found is kept. Typically, the fuzzy membership of a "Deposit" site is greater than those of a "Not Deposit" site; so this rule favors "Deposit" sites. It is not believed this change will produce significant changes in the results if a small number of carefully selected sites were used. If, however, a large number of sites were used for training and many of them are in the same unique condition, then different results can be obtained with this corrected train.dta file.
2. A change in the format of the Train.dta file has been made. In the data records, record 5 and beyond, in the Train.dta file, columns 2 and 3 are ignored by GeoXplore. Therefore, in column 2 we have put the FID number of the point Shapefile record selected. If the selected record is from a "Deposit" Shapefile, the FID is recorded in column 2. If the selected record is from a "Not Deposit" Shapefile, the number entered is 1000 + FID. In column 3 we have put the value attribute number of the unique condition record for the Train.dta record. When training in GeoXplore, the classification of the Train.dta record number, column 1 in the Train.dta evidence records, is reported. Selecting this record number in the Train.dta file provides the FID number of the site and the value attribute of the unique condition that produced a classification result. Sometimes when a record or group of records do not classify well in GeoXplore training, insights can be gained by knowing which records. Sometimes such records are not appropriate to the desired training goals.
3. A new Home Page for future SDM developments is being implemented at the University of Campinas by Prof. Carlos Roberto de Souza Filho, beto@ige.unicamp.br. This site has been maintained to contain historic versions of SDM and related documentation. The site is now being redesigned to contain among other things "step-by-step" tutorials, new validation tools, possibly new neural net tools such as a Self Organizing Map (SOM) neural net, and other tools. The URL for this site is http://www.ige.unicamp.br/sdm/default_e.htm and it is anticipated that the redesign will be operating by January 2009. So check out this site for the future evolution of SDM. The ArcGIS 9.2 version of SDM will be

maintained on the ESRI site (http://arcscripts.esri.com/details.asp?dbid=15341) . It is anticipated that the ArcGIS 9.3 version of SDM will be posted on the University of Campinas home page for SDM evolution.

Known Issues

The user should be aware of these known issues.

Path and Folder Names

Long paths to folders can cause problems. It is better to use short paths to folders, such as C:\MyModel. We have found that models fail to add the results when a long path is used for the workspace. A path to the workspace of 25 characters is known to cause this problem.

Paths and folders can never have blanks in the name. For example, “Gary’s Workspace” is not an acceptable name for ArcGIS folders; however Garys_Workspace is acceptable.

Note SDM does not support geodatabases; so the Workspace nor Scratch Workspace cannot be a geodatabase.

Calculate Weights Overwrite

This tool will not overwrite an existing weights table that is in the table of contents of the open ArcGIS map document. Assuming the intent is to overwrite this file, the work around for this problem is to delete the weights table from the table of contents.

Area-Frequency Lock

The Area Frequency tool creates a DBF table from which several graphs can be made. If the DBF table is removed from the table of contents of ArcGIS, but the graphs are still present in the MXD, a lock will not allow overwriting the existing DBF. This lock problem is a known problem in ArcGIS.

The workaround for DBF overwrite problem is to write a new file when running this tool.

Memory Management

The GeoXplore, Area Frequency, and Site Reduction tools all require large arrays be stored in memory. On some computers these large arrays can cause failure of the tool.

The GeoXplore.exe fails sometimes when Fuzzy Classification is started after doing a Fuzzy Neural Net training. If the GeoXplore tool is closed and restarted, then Classification can be done. This failure causes the NN Demonstration Model to fail because all three neural net methods must be completed by GeoXplore for this model to finish. In normal neural net operations this is not an problem because the NN Tool should be used. The NN Demonstration Model is simply a training model that documents the full processing scheme. This problem can be fixed by defragmentation of the computer using an aftermarket defragmentation tool.

If the number of training sites gets very large, the Area Frequency and Site Reduction tools can fail in different ways. This error has been seen on some computers when 11,000 training sites were used with these tools. On other computers, these tools computed for over 5 hours and did not complete. Huge number of training sites are generally a violation of basic assumption of Weights of Evidence, which assumes the training site is a very small area and the total training sites are small. Therefore, the solution is to use a small number of training sites. When this problem has been encountered and the number of training sites reduced significantly, the response rasters from WofE and LR were unchanged and the Area Frequency and Site Reduction tools worked as expected. We believe that about 1000 training sites is the maximum that works, but we are not sure. We consider significantly more than 1000 training sites excessive.

If the problem occurs with the Site Reduction tool due to a Shapefile with a very large number of points, one solution is to first make a random subset of the points. Create a new Shapefile with this subset. Then apply the thinning with the Site Reduction tool to this new Shapefile. In a test of Shapefiles with 500, 1000, and 5000 points where both thinning and random subsets were done simultaneously, it took, respectively, 3 seconds, 23, seconds, and 7 minutes for the tool to complete. The above work around should address the problem. The thinning algorithm is very slow whereas the random subset process if very fast, so the two-step process suggested above should solve the problem.

GeoXplore

The neural nets are limited to about 200,000 unique conditions. This limitation is due to array sizes in GeoXplore.exe. Two new tools have been added to the Neural Net toolset to deal with this problem.

Layer Statistics

In the SDM folder are three LYR files useful for symbolization of models. The names of these files indicate their application. If applied in models, these LYR files corrupt the layer statistics in the Symbolization window. To undue this corruption if a different symbolization is desired, the raster must first be symbolized by the Stretch Method. The Parameters window is then closed and reopened. Then the statistics will be revised so another Symbolization method can be applied. This is a known bug in ArcGIS 9.2

Iteration Demonstration Models

Most of these models using iteration run properly if run in the Edit mode. If these models are simply opened in order to run, the incrementation of the outputs does not function properly and the results are not always added to the table of contents. This problem is suppose to be fixed in ArcGIS 9.3.

Limit to Number of Rasters

ArcGIS sometimes appears to misbehave when a large number of rasters (more than 50) have been added to an MXD. With tools such as Grand WofE and when doing a lot of

experimentation, a large number of rasters can be created quickly. The solution is to either start a new MXD or to delete some unneeded rasters.

This problem might be due to the Geoprocessing History becoming over full. It is worthwhile going to My Toolboxes and emptying the geoprocessing history toolbox. Also, the Temp folder in C:\Documents and Setting\User Name\Local Settings can also become over full and should periodically be emptied.

Logistic Regression

This tool still uses a unique conditions table. The tool is limited to 6000 unique conditions. This is because of array sizes in SDMLR.exe. We are investigating recreating SDMLR.exe and increasing the array sites.

This problem can easily be encountered if the ordered data are not binary, such as using the unique method in Calculate Response.

The only solution at this point is to take the data into a statistical package, such as SAS, to do the logistic regression and bring the results back to ArcGIS. The Logistic Regression tool in the Scratch Workspace creates all of the necessary files.

Area Frequency

If a WofE or LR model is formed incorrectly with a huge number of training sites, for example 11,000, the area frequency tool can fail or continue to calculate for hours. This is because this tool creates a complex table and it has to deal with too many training sites.

A clue that you have created the potential for this problem is the Studentized Contrast is huge, such as 90. A value of 7 or 8 might occur in a well formed model, but greater values should be suspect.

The problem is corrected easily by creating a proper training set, which is typically a small number of sites. We believe the tool will handle as many as 1000 training sites without problem; but the user should really consider if so many training sites area required. Training sites are points that represent areas. See Poli and Sterlacchini in Natural Resources Research, Vol. 16, No. 2, June 2007, pages 121-134 for an analysis addressing this problem of points representing areas. In our experience to date, considering the map scale of the evidence and the nature of the training sites has always lead to the conclusion that a single point is sufficient to represent the features to be modeled and the number of such points is always a few tens to a few hundreds of points.