1 ArcGIS and ArcGIS Pro Configuration

Tips: ArcGIS Pro install dependent libraries are easier to ArcGIS

ArcGIS Pro：python3 (ArcGIS Pro2.5 corresponds to python3.6)

ArcGIS：python2.7

1.1 Version requirements

This toolbox requires **ArcGIS version no less than 10.1**, and ArcGIS Pro has no version requirements.

1.2 ArcGIS Pro Installs Additional Python Libraries

Start menu——ArcGIS——Python Command Prompt

Similar to Anaconda installing python libraries:

pip install scikit-learn --user

pip install seaborn --user

1.3 ArcGIS Installs Additional Python Libraries

1. Firstly, check whether the Python2.7 installation path was modified when installing ArcGIS, if it is modified, it should be reinstalled, and the Python2.7 installation path remains the default.

Check whether the default: if the "*C:\python27*" path exists, the path has not been modified, you can proceed to the following steps.

2. Open *cmd* and go to the *C:\Python27\ArcGIS10.X\Scripts* path and use:

pip install scikit-learn

pip install seaborn

2 Using of SVM-LSM Toolbox

2.1 Toolbox Installation

In the ArcGIS "Catalog" interface, right click "Folder Connections" (Figure 1), and find the toolbox under the save position this toolbox to add.

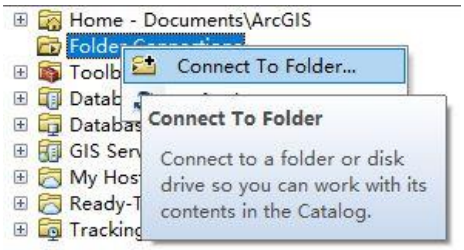

Fig.1 Add Toolbox

2.2 Toolbox Description

This toolbox aims to provide an easy-to-use user-friendly toolbox for landslide susceptibility mapping based on support vector machines. The toolbox includes "1 Influencing Factor Production", "2 Dataset Production and Factor Selection", and "3 Model Training and Prediction" three sub-toolbox and 11 tools, as shown in Figure 2. The implementation of this toolbox is based on *Arcpy* and Model Builder. You need to install *scikit-learn* and *seaborn* in the environment of ArcGIS or ArcGIS Pro Python. The specific method of using the toolbox below takes supporting case data as an example.

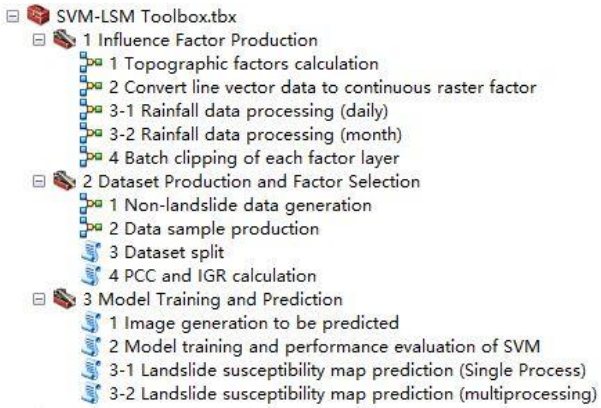

Fig.2 SVM-LSM toolbox

2.2.1 Influencing Factor Production

(1) Topographic Factors Calculation

[Function]

Topographic factors (such as slope, aspect, curvature, plan curvature, profile curvature, relief amplitude, surface roughness, topographic wetness index (TWI), etc.) are automatically calculated according to DEM in the study area. It is noted that DEM must be UTM coordinate system. These factors can be calculated selectively according to the needs of users, but the aspect must be calculated when calculating the plane curvature, and the slope must be calculated when calculating profile curvature, surface roughness, or TWI.

[Input]

- DEM (UTM Coordinate System): required. DEM raster data of the study area (UTM coordinate system)
- Workspace: required. Output the directory of file stores. The tool will automatically generate a "*dempro*" folder under this path to save data.
- Factor Selection: optional. According to demand, select factors that need to be calculated and customize a factor name, and vice versa. This factor name is corresponding to the output's name, respectively. Note: the aspect must be calculated when calculating the plane curvature, and the slope must be calculated when calculating profile curvature, surface roughness, or TWI.

[Output]

Slope, aspect, curvature, plan curvature, profile curvature, relief amplitude, surface roughness, and topographic wetness index (TWI) raster data were calculated by DEM.

In order to facilitate subsequent batch clipping, it is recommended to move data in the "*dempro*" folder to the upper level folder.

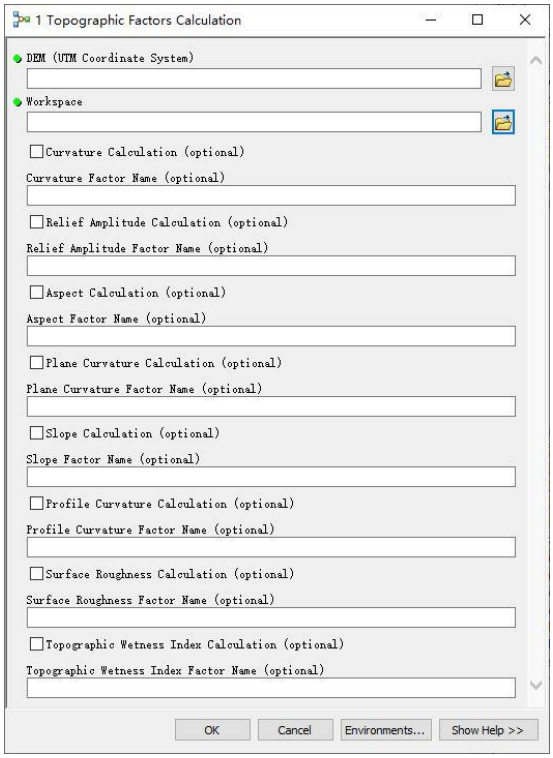

Fig.3 Topographic Factors Calculation

(2) Convert Line Vector Data to Continuous Raster Factor

[Function]

Converted line vector data to continuous raster data automatically in the study area. Such as converting roads to the distance to roads, converting faults to the distance to faults, and converting rivers to the distance to rivers. Among them, the generated raster data resolution is 30 m, and the conversion principle is Euclidean distance. Note: The line vector data must be UTM coordinate system.

[Input]

- Line Vector Data: required. Line vector data to be calculated, such as:

  Roads_UTM: roads data of study area (UTM coordinate system). If this item is not calculated, you do not need to be added.

  Faults_UTM: faults data of study area (UTM coordinate system). If this item is not calculated, you do not need to be added.

  Rivers_UTM: rivers data of study area (UTM coordinate system). If this item is not calculated, you do not need to be added.

- Output Path: required. Directory of the output file. The output file is saved directly under the folder after generation.

[Output]

Raster data of Distance to roads, distance to faults, and distance to rivers.

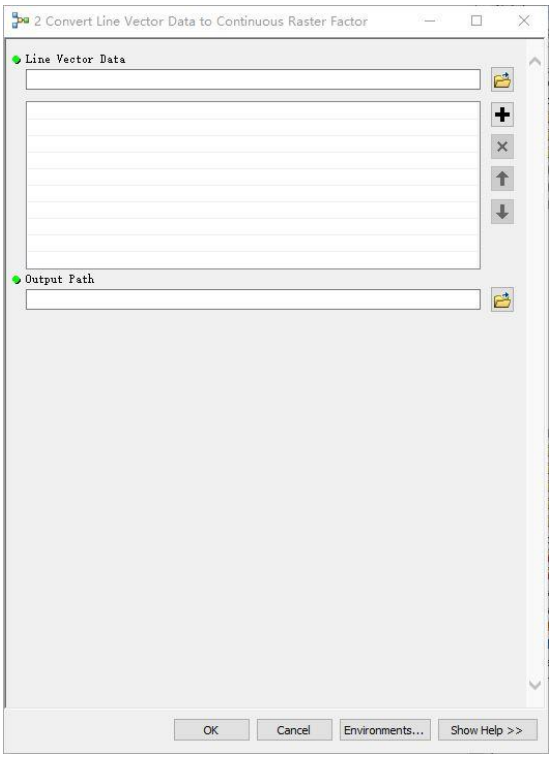

Fig.4 Convert Line Vector Data to Continuous Raster Factor

(3) Rainfall Data Processing

[Function]

This tool converted monthly or daily rainfall data (*.nc4*) downloaded from NASA (*https://gpm.nasa.gov/*) to raster data (*.tif*) with a resolution of 30 m. After the conversion, it is still monthly or daily raster data (*.tif*).

[Input]

- Rainfall Data (*.nc4*) Folder: required. The data folder of the monthly or daily rainfall data downloads from NASA.
- Output Coordinate System: required. For the converted *.tif* data coordinate system, it is recommended to select UTM coordinate system.

[Output]

The rainfall raster data (*.tif*) corresponds to the *.nc4* file, and create a new folder named "*UTM_30m*" under the input path.

The raster calculates tool should be used to stack data as required in subsequent use. The annual rainfall is used in this case. Therefore, the raster calculates tool is used to

stack all *.tif* get annual rainfall. (Strictly use the raster calculate tool operation, do not keyboard input "+", etc.)

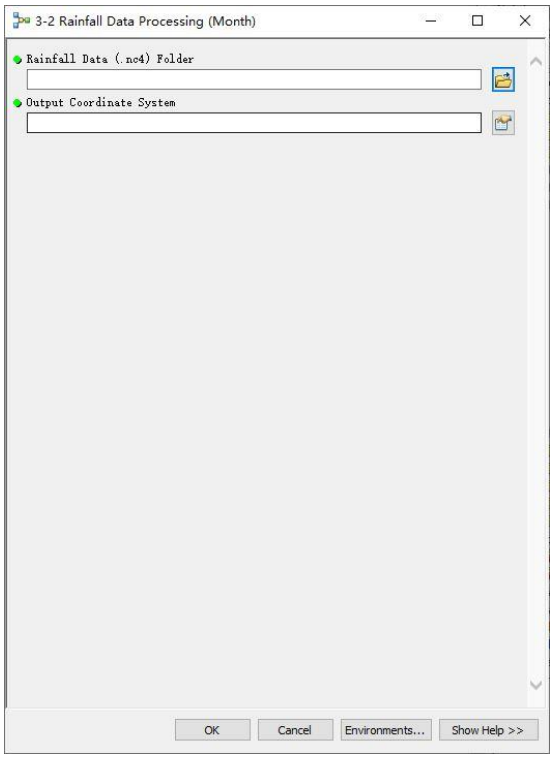

Fig.5 Rainfall Data Processing (month)

(4) Batch Clipping for each Factor Layer

[Function]

This tool is used to batch clip the raster data of each factor layer according to the vector data of the study area to obtain the factor layer data of the study area. This tool only needs to give the folder where the raster is located, and automatically iterates to select the *.tif* file for clipping. Note: Vector data and raster data must be UTM coordinate systems, and raster data's resolution must be consistent.

[Input]

- Vector Data of Study Area (UTM Coordinate System): required. Select the study area vector data (*.shp*) file, and note that the vector data has been projected to UTM coordinate system.
- Raster Data Folder (UTM Coordinate System): required. The folder where the raster data needs to be clipped. The *.tif* file will be automatically identified and iterated to clip. Note that all raster data have been projected into UTM coordinate system and have a consistent resolution.

- Use Input Features for Clipping Geometry: optional. Not recommended! Checked - Uses the geometry of the selected feature class to clip the data. The pixel depth of the output may be increased; therefore, you need to make sure that the output format can support the proper pixel depth. Unchecked - Uses the minimum bounding rectangle to clip the data.
- Maintain Clipping Extent: optional. Strongly recommended, be sure to check! Checked - Adjusts the number of columns and rows and resamples pixels to exactly match the clipping extent specified. Unchecked - Maintains the cell alignment as the input raster and adjusts the output extent accordingly.
- Output Data Folder: required. The clipped factor layer data in the study area is saved in the folder. The output file is saved directly under the folder after generation.

[Output]

Batch clipped raster data in the study area.

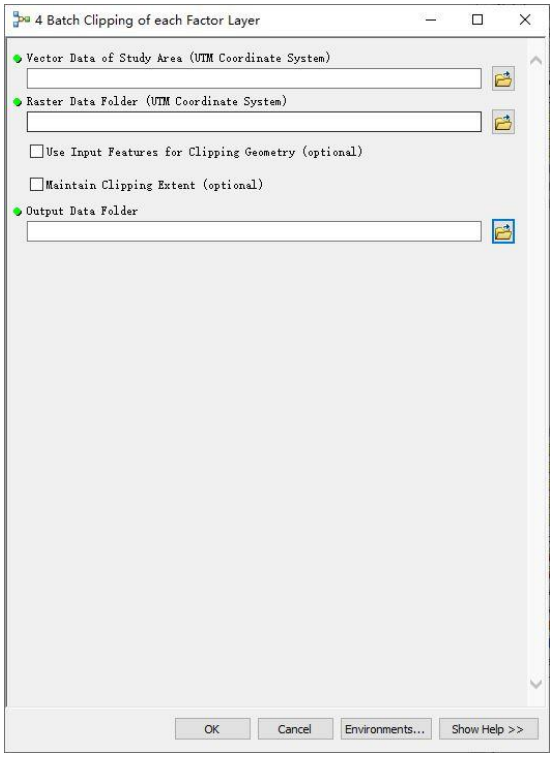

Fig.6 Batch Clipping for each Factor Layer

2.2.2 Dataset Production and Factor Selection

(1) Non-landslide Data Generation

[Function]

This tool is used to generate non-landslide point data within the vector layer of the study area. Principle: randomly select the same number of non-landslide sample points outside a certain buffer area for a given landslide sample point. Note: The vector data of the study area and the vector data of the landslide points must be in the UTM coordinate system, and the obtained non-landslide point vector data should be consistent with the landslide point vector data coordinate system by default.

[Input]

- Landslide Point Feature (UTM Coordinate System): required. Select the landslide point vector data (*.shp*) file of the study area, and note that the vector data has been projected into the UTM coordinate system.
- Distance to Landslide Point: required. Select non-landslide sample points outside the range of the landslide sample points buffer with a given distance.
- Vector Data of Study Area (UTM Coordinate System): required. Select the study area vector data (*.shp*) file, and note that the vector data has been projected to the UTM coordinate system.
- Number of Points: required. The number of landslide points, that is, the number of non-landslide points generated.
- Output Folder: required. Output directory of non-landslide point vector data. The output file name of the non-landslide point vector data is *non_landslide_point.shp* by default.
- Output Coordinate System: optional. The Coordinate system of non-landslide point vector data, the default is consistent with the landslide point vector data coordinate system.

[Output]

Vector data of non-landslide sample points in the study area.

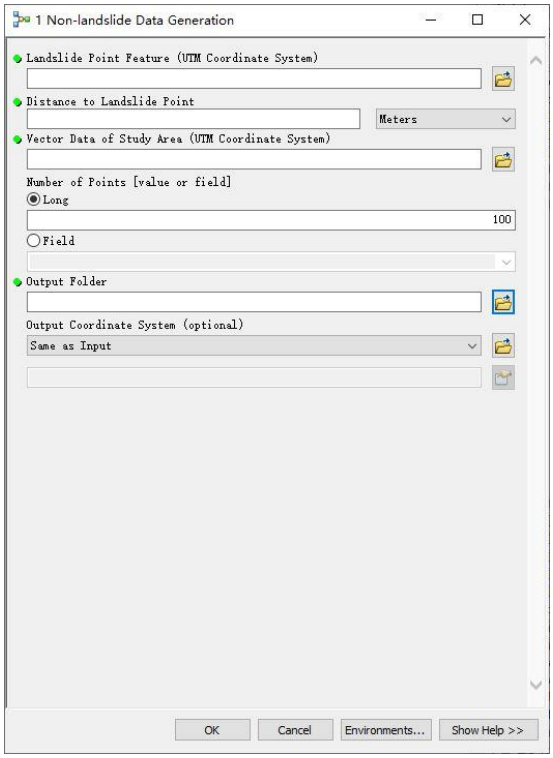

Fig.7 Non-landslide Data Generation

(2) Data Sample Production

[Function]

This tool generates multi-channel block sample raster data from vector point data. Principle: Use vector point data (*.shp*) to make buffers and clip multi-channel raster data (*.tif*) one element by one, get a single clipping result of each element, and name it with the "*FID*" value. Note: Both vector point data (*.shp*) and multi-channel raster data (*.tif*) are in the same UTM coordinate system. The multi-channel block sample raster data is consistent with the multi-channel raster data (*.tif*) coordinate system by default.

[Input]

- Input Point Feature: required. Select the landslide point vector data (*.shp*) or non-landslide point vector data (*.shp*) in the study area, and note that the vector data has been projected to the UTM coordinate system.
- Buffer Distance: required. The landslide sample will generate based on the landslide sample point and buffer with a given distance.
- Multi-channel Factor Layer Data: required. The combined multi-channel raster data for each factor layer was used.

- Data Sample Save Folder: required. Save path of landslide and non-landslide sample datasets.
- Sample Label (landslide or non-landslide): Required. The name of the folder name of the dataset is saved.
- Use Input Features for Clipping Geometry: optional. Not recommended! Checked - Uses the geometry of the selected feature class to clip the data. The pixel depth of the output may be increased; therefore, you need to make sure that the output format can support the proper pixel depth. Unchecked - Uses the minimum bounding rectangle to clip the data.
- Maintain Clipping Extent: optional. Strongly recommended, be sure to check! Checked - Adjusts the number of columns and rows and resamples pixels to exactly match the clipping extent specified. Unchecked - Maintains the cell alignment as the input raster and adjusts the output extent accordingly.

[Output]

The block data samples of landslide and non-landslide in the study area.

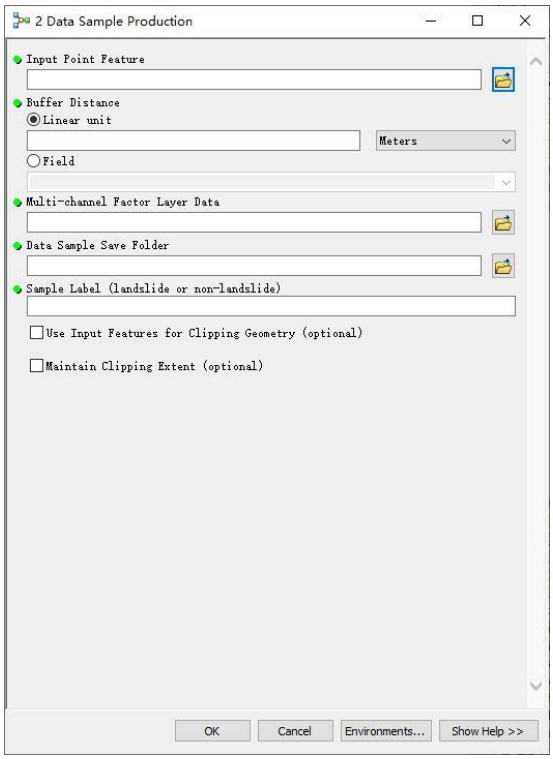

Fig.8 Data Sample Production

(3) Dataset Split

[Function]

This tool is based on the number of landslide points and the data samples generated in the previous step, dividing the training dataset and the test dataset according to the ratio of the test dataset. and saving the division results in a *.txt* file.

[Input]

- Sample Folder: required. Select the save path of the previous landslide and non-landslide samples.
- Number of Landslides: required. Number of landslide points.
- Test Dataset Ratio (e.g. 0.3): required. The ratio of the test dataset to the total samples, 1 minus this value is the ratio of the training set to the total samples.

[Output]

Generate three text files, all sample paths and labels (*XXX.txt*), all training sample paths and labels (*XXXtrain.txt*), and all test sample paths and labels (*XXXtest.txt*) respectively.

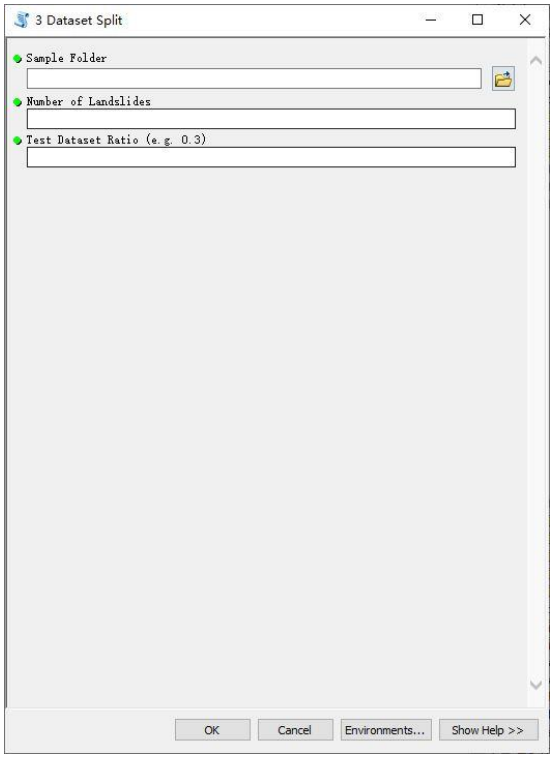

Fig.9 Dataset Split

(4) PCC and IGR Calculation

[Function]

This tool calculates the Pearson correlation coefficient (PCC) and information gain ratio (IGR) of each influencing factor layer based on the generated data samples and

all sample paths and label files. The correlation between the factor layers represented by the PCC is between [-1, 1], and factors with greater correlation should be considered to be eliminated. The IGR represents the contribution of each factor layer to the occurrence of landslides. If its value is greater than 0, it means that it contributes to the occurrence of landslides. The larger the value, the greater the contribution.

PCC: $$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}},$$

where *r* represents the PCC, $X_i$and$Y_i$represent the *i*-th factor layer, respectively, and $\bar{X}$and $\bar{Y}$represent the average value of *X* and *Y*, respectively.

IGR: $$IGR(D,A) = \frac{IG(D|A)}{Splitinfor_A(D)},$$

where *D* be the label of a landslide influencing factor value in the training set; that is, whether it is a landslide or not. And *A* is a landslide influencing factor value in the training set. $IG(D|A)$ is the information gain of the label *D* corresponding to the landslide influencing factor value *A*, and $Splitinfor_A(D)$ is the information entropy of the label *D* corresponding to the landslide influencing factor value *A*. which is:

$$IG(D|A) = Infor(D) - Infor(D|A),$$

$$Splitinfor_A(D) = -\sum_{i=1}^{m}\frac{|D_i|}{|D|} \times log_2\left(\frac{|D_i|}{|D|}\right),$$

where $Infor(D|A)$ is conditional entropy and $Infor(D)$ is information entropy; the calculation formula is as follows:

$$Infor(D|A) = \sum_{j=1}^{n}\frac{|D_j|}{|D|} \times Infor(D_j),$$

$$Infor(D) = -\sum_{j=1}^{n} p(x_j) \times log_2 p(x_j).$$

[Input]

- Dataset Folder: required. Select the save path of the previous step of landslide and non-landslide samples.
- Result Save Folder: required. Save path of PCC and IGR result.
- Input Factor Layer Order: required. The stacking order of factor layer raster data. It needs to be consistent with "*Factor Layer Stacking Order*" in "*Image*

Generation to be Predicted"

[Output]

Generate six files [*PCC.txt*, *PCC.png*, *PCC.svg*, *IGR.txt*, *IGR.png*, *IGR.svg*].

[**.txt*] PCC or IGR txt file.

[**.png*]PCC or IGR png file.

[**.svg*] PCC or IGR editable vector file.

After the results are obtained by this tool, the PCC and IGR results should be considered comprehensively, and redundant factors should be removed.

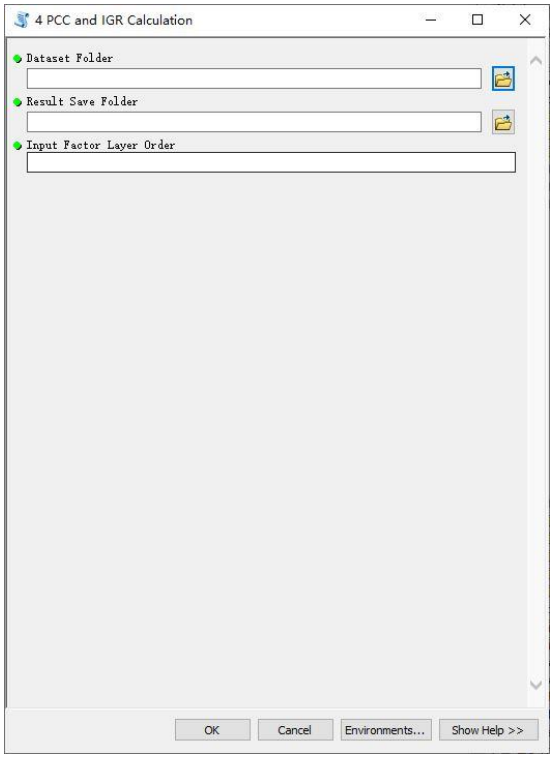

Fig.10 PCC and IGR Calculation

2.2.3 Model Training and Prediction

(1) Image Generation to be Predicted

[Function]

This tool generates multi-channel raster data of the image to be predicted based on the raster data of each factor layer. It is used for subsequent data sample production and prediction of susceptibility maps. Note: All factor data (*.tif*) are in the same UTM coordinate system. The obtained multi-channel raster data is consistent with the coordinate system of each factor raster data (*.tif*) by default.

[Input]

- Influencing Factor Folder: required. Select the directory where the raster data of each influencing factor in the study area is located, and note that all raster data have been projected to the UTM coordinate system.
- Stacking Factor Layer Order: required. The order in which to stack the factor layer raster data. It needs to be consistent with the name of each factor layer selected in the previous step.
- Save Folder of the Image to be Predicted: required. The folder where the multi-channel raster data is saved. This tool will automatically create a raster file of "*Factors_[Number of selected factor layers]_mapping.tif*" in the folder.

[Output]

The multi-channel raster data of the image to be predicted in the study area, the file name is "*Factors_[Number of checked factor layers]_mapping.tif*".

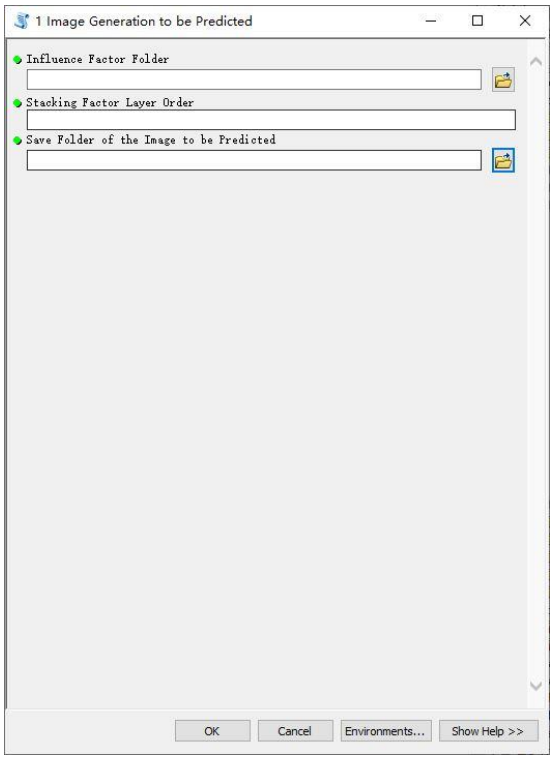

Fig.11 Image Generation to be Predicted

(2) Model Training and Performance Evaluation of SVM

[Function]

This tool is used to generate the SVM model under each group of parameters and give the evaluation result of the model performance. In this tool, the default SVM kernel function is radial basis function (RBF), the parameters to be adjusted are

gamma and penalty factor *C*, and the parameter adjustment method is a grid search algorithm.

[Input]

- Dataset Folder: required. Select the save path for the generated landslide and non-landslide samples.
- Model Save Folder: required. Select the save path of the SVM model trained by each group of parameters. Under this path, different parameter folders will be created with "*g_g value_C_C value*" to store the results.
- gamma Optional Value: required. The optional value of the *gamma* parameter in the SVM radial basis kernel function.
- C Optional Value: required. The optional value of the *C* parameter in the SVM radial basis kernel function.
- Number of Dataset Rows: required. The number of rows in the dataset sample.
- Number of Dataset Columns: required. The number of columns in the dataset sample.
- Number of Dataset Channels: required. The number of channels in the dataset sample.

[Output]

- [*parameter_result_txt.txt*]: Store the AUC value of the SVM model under each group of parameters on the test dataset, the test dataset accuracy (*Test_acc*), the training set accuracy (*Train_acc*), and the difference between the two. The results can be used for optimal model selection.
- [*parameter_result_png.png and parameter_result_png.svg*]: txt file drawing display. In the figure, the size of the circle represents the AUC value. The larger the circle, the higher the AUC value and the better the accuracy of the model. The circular color represents the accuracy difference between the training dataset and the test dataset. If it exceeds 0.5, it is represented by 0.5. The greater the difference, the higher the degree of overfitting of the model and the worse the generalization performance.
- [*\g_0.02_C_0.5*]: The save path of the SVM result in *gamma* of 0.02 and *C* of

0.5.

- [*\g_0.02_C_0.5\SVM_g_0.02_C_0.5model*]: The save path of the SVM trained model result with *gamma* of 0.02 and *C* of 0.5.
- [*\g_0.02_C_0.5\SVM_train_result_txt.txt*]: The prediction result of the SVM model on the training dataset with *gamma* is 0.02 and *C* is 0.5
- [*\g_0.02_C_0.5\SVM_test_result_txt.txt*]: The prediction result of the SVM model on the test dataset with *gamma* is 0.02 and *C* is 0.5
- [*\g_0.02_C_0.5\evaluate_result.txt*]: Various evaluation indicators of the SVM model on the test set, such as confusion matrix, accuracy, precision, F1 value, AUC value, etc. with *gamma* is 0.02 and *C* is 0.5
- [*\g_0.02_C_0.5\ROC.png*]: The ROC curve and AUC value of the SVM model on the test dataset with *gamma* is 0.02 and *C* is 0.5

According to [*parameter_result_txt.txt and parameter_result_png.png*], the model with a higher AUC value and the smaller difference is selected as the optimal model for susceptibility map prediction.

Fig.12 Model Training and Performance Evaluation of SVM

(3) Landslide Susceptibility Map Prediction

This function includes two tools, namely "Landslide Susceptibility Map Prediction

(Single Process)" and "Landslide Susceptibility Map Prediction (Multiprocessing)". Single process and multiprocessing can be used in ArcGIS and ArcGIS Pro, and since the python 2.7 installed in ArcGIS is generally 32-bit, in the face of a large amount of data, the 32-bit Python environment has extremely limited use of memory resources.

[Function]

This tool predicts the landslide susceptibility map of the study area based on the optimal model obtained in the previous step. The obtained susceptibility map coordinate system is consistent with the input image to be predicted.

[Input]

- Path of the Image to be Predicted (UTM Coordinate System): required. Select the directory where the multi-channel raster data to be predicted in the study area is located.
- Vector Data of Study Area (UTM Coordinate System): required. Used to clip the generated susceptibility map to ensure that the results are within the study area.
- Optimal Model Folder (.model): required. The optimal model save path was obtained in the previous step.
- LSM Output Path: required. The generated landslide susceptibility map saves a file (*.tif*).
- Number of Dataset Rows: required. The number of rows in the dataset sample.
- Number of Dataset Columns: required. The number of columns in the dataset sample.
- (multiprocessing parameter) pythonw.exe path: required. Generally, ArcGIS Pro is "*C:\Python27\ArcGIS10.8*", ArcGIS Pro is "*ArcGIS Pro installation path\bin\Python\envs\arcgispro-py3*".

[Output]

Landslide susceptibility map of the study area.

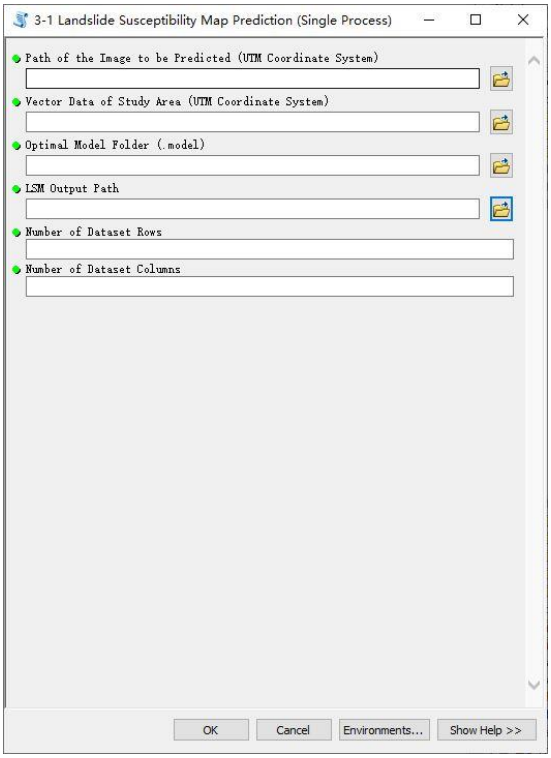

Fig.13 LSM Prediction (Single Process)

Fig.14 LSM Prediction (multiprocessing)

2.3 Data preparation and case data description

[*Data*] is the original data that matches this case, and the data description is shown in Table 1.

Table 1 Data description

Data Preparation	Data Format	Case Data
Study area boundary	polygon feature (.shp)	\Data\point\study_range.shp
Historical landslide data in the study area	point feature（.shp）	\Data\point\landslide_point.shp
DEM	30 m raster（.tif）	\Data\big_factor\dem.tif
faults	line feature（.shp）	\Data\big_factor\faults.shp
lithology	30 m raster（.tif）	\Data\big_factor\lithology.tif
roads	line feature（.shp）	\Data\big_factor\roads.shp
rivers	line feature（.shp）	\Data\big_factor\rivers.shp
NDVI	30 m raster（.tif）	\Data\big_factor\NDVI.tif
monthly rainfall	NC4 files（.nc4）	\Data\big_factor\rainfall*.nc4

Note: All data must be projected to the UTM coordinate system.

[*Case*] is the paper data generated by using the toolbox, all case data in this manual take ArcGIS software as an example, the data description is as follows:

- *Case**big_factor*: saving the original data and the influencing factor data that are not batch clipped.
- *Case**point*: save the vector data of the study area and the vector files of landslide and non-landslide points.
- *Case**factors*: save the clipped data of each factor layer in the study area.
- *Case**dataset*: save block datasets generated without factor selecting, including landslide (*landslide*) and non-landslide samples (*non-landslide*).
- *Case**IGR_dataset*: save the block datasets generated after factor selecting, including landslide (*landslide*) and non-landslide samples (*non-landslide*).
- *Case**model*: save the generated models and model evaluation results with different parameters.
- *Case**predict*: save the generated images to be predicted and the predicted susceptibility map.
-