Usage Process (Case)

1 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.

2 SVM-LSM usage process (take Wuqi County as an example)

Taking Wuqi County, Yan'an City, Shaanxi Province, China as an example, the developed toolbox was applied to carry out landslide susceptibility assessment. The overall flowchart is shown in Figure 1. First, collect historical landslide data in the study area, and select appropriate landslide influencing factors for subsequent research based on the cause of landslides in the study area. Note: In use, you should ensure that all data are in the UTM coordinate system, otherwise some unknown errors may occur. Since there is no fault distribution in the study area and it is basically not affected by the fault, the distance to fault is not used.

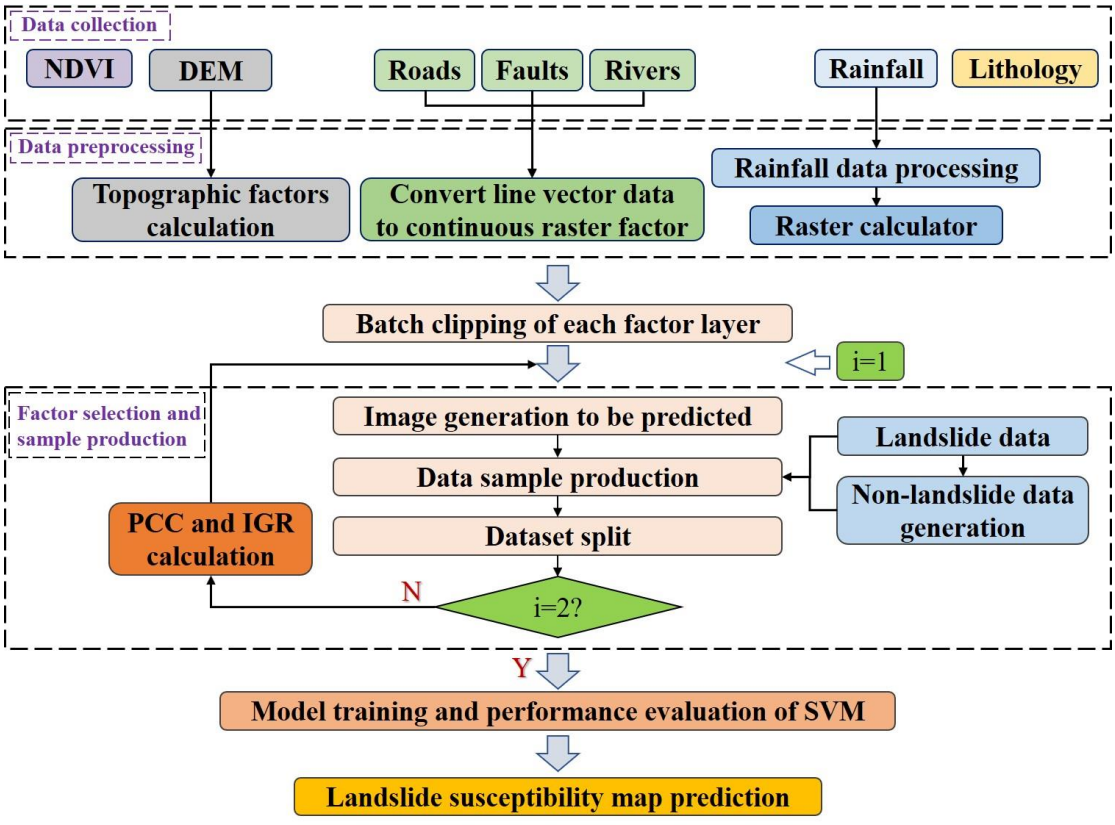

Fig.1 Flowchart for using the SVM-LSM toolbox

1. Topographic Factors Calculation

Use the "1 Topographic Factors Calculation" tool in the "1 Influencing Factor Production" toolbox to automatically calculate other topographic factors based on the

DEM data of the study area, such as slope, aspect, curvature, plan curvature, profile curvature, relief amplitude, surface roughness, topographic wetness index (TWI), etc. Note: DEM data must be in UTM coordinate system. The aspect must be calculated when calculating the plane curvature, the slope must be calculated when calculating profile curvature, surface roughness, or TWI.

[Input]

- DEM (UTM Coordinate System): [*\Case\big_factor\dem.tif*]
- Workspace: [*\Case\big_factor*]
- Selection of calculation factors: select all and named [*slp, asp, cur, plancur, profilecur, SroughnessC, relief, TWI*] respectively.

[Output]

- Slope [*\Case\big_factor\dempro\slp.tif*]
- Curvature [*\Case\big_factor\dempro\cur.tif*]
- Plane Curvature [*\Case\big_factor\dempro\plancur.tif*]
- Profile Curvature [*\Case\big_factor\dempro\profilecur.tif*]
- Relief amplitude [*\Case\big_factor\dempro\relief.tif*]
- Surface Roughness [*\Case\big_factor\dempro\SroughnessC.tif*]
- Topographic wetness index (TWI) [*\Case\big_factor\dempro\TWI.tif*]
- Environment Setting : Parallel Processing Factor: 0

To facilitate subsequent batch clipping, it is recommended to move these data in the "*dempro*" folder to the [*\Case\big_factor*] folder.

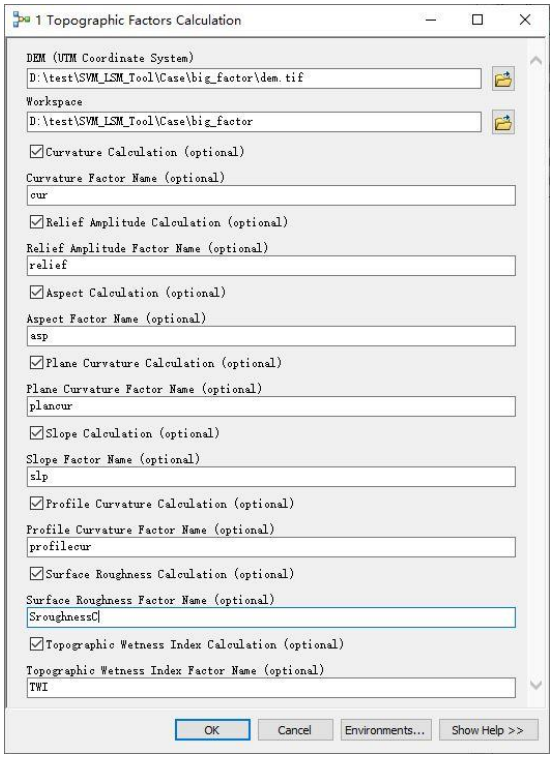

Fig.2 Step 1

2. Convert Line Vector Data to Continuous Raster Factor

Use the "2 Convert Line Vector Data to Continuous Raster Factor " tool in the "1 Influencing Factor Production" toolbox to automatically convert the line vector data of the study area to continuous raster data, such as converting roads to the distance to roads, convert faults to the distance to faults, and convert rivers to the distance to rivers, etc. Among them, the resolution of the generated raster data is 30 m, and the conversion principle is the Euclidean distance. Note: Line vector data must be in UTM coordinate system.

[Input]

- Line Vector Data:

  Roads data (UTM coordinate system): [*\Case\big_factor\roads.shp*]

  Rivers data (UTM coordinate system): [*\Case\big_factor\rivers.shp*]

- Output Path: [*\Case\big_factor*]
- Environment Setting : Parallel Processing Factor: 0

[Output]

Distance to roads [*\Case\big_factor\roads.tif*]

Distance to rivers [*\Case\big_factor\rivers.tif*]

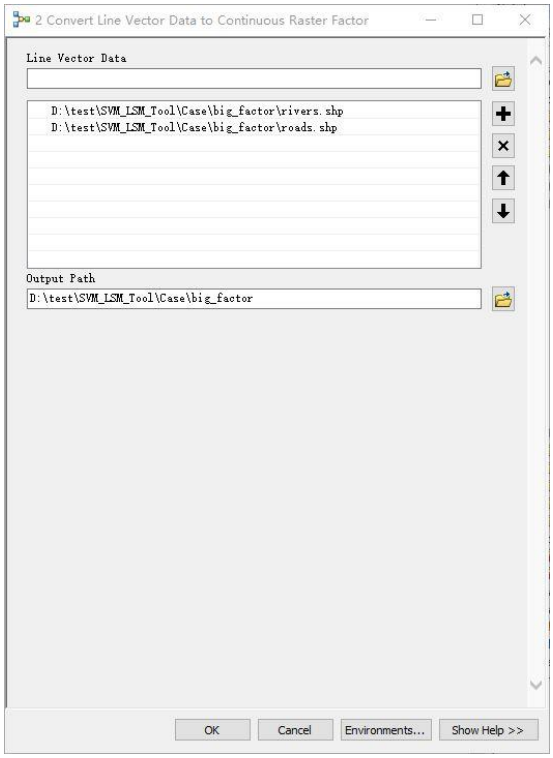

Fig.3 Step 2

3. Rainfall Data Processing (month)

Use the "3 Rainfall Data Processing (Month)" tool in the "1 Influencing Factor Production" toolbox for monthly rainfall data (*.nc4*) downloaded from NASA (*https://gpm.nasa.gov/*) raster data (*.tif*) with a resolution of 30 m. It is still monthly raster data (*.tif*) after conversion.

[Input]

- Rainfall Data (*.nc4*) Folder: [*\Case\big_factor\rainfall*]
- Output Coordinate System: [*WGS_1984_UTM_Zone_49N*]

[Output]

The monthly rainfall file (*.tif*) corresponding to the *.nc4* file. [*\Case\big_factor\rainfall\UTM_30m*].

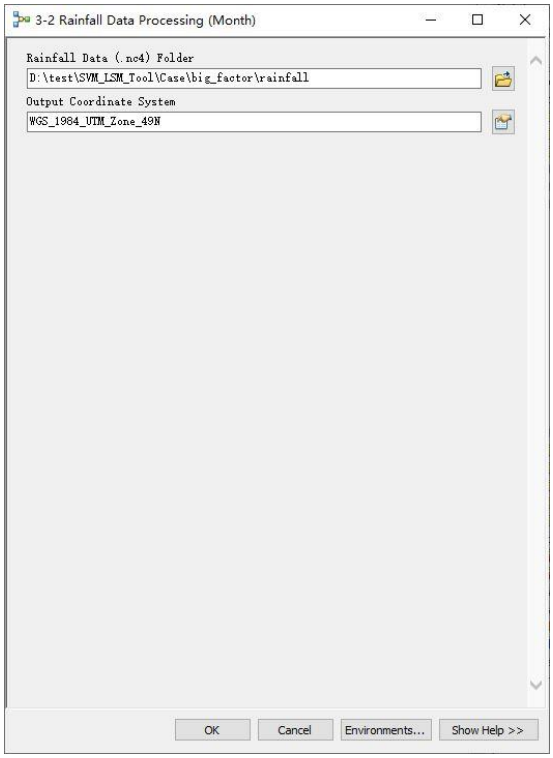

Fig.4 Step 3

4. Annual Rainfall Data

Use the raster calculator to accumulate all *.tif* (strictly use the raster calculate tool operation, do not keyboard input "+", etc.) to get the annual rainfall, named *rainfall.tif*.

[Input]

Converted monthly rainfall raster data.

[Output]

Annual rainfall data [*\Case\big_factor\rainfall.tif*]

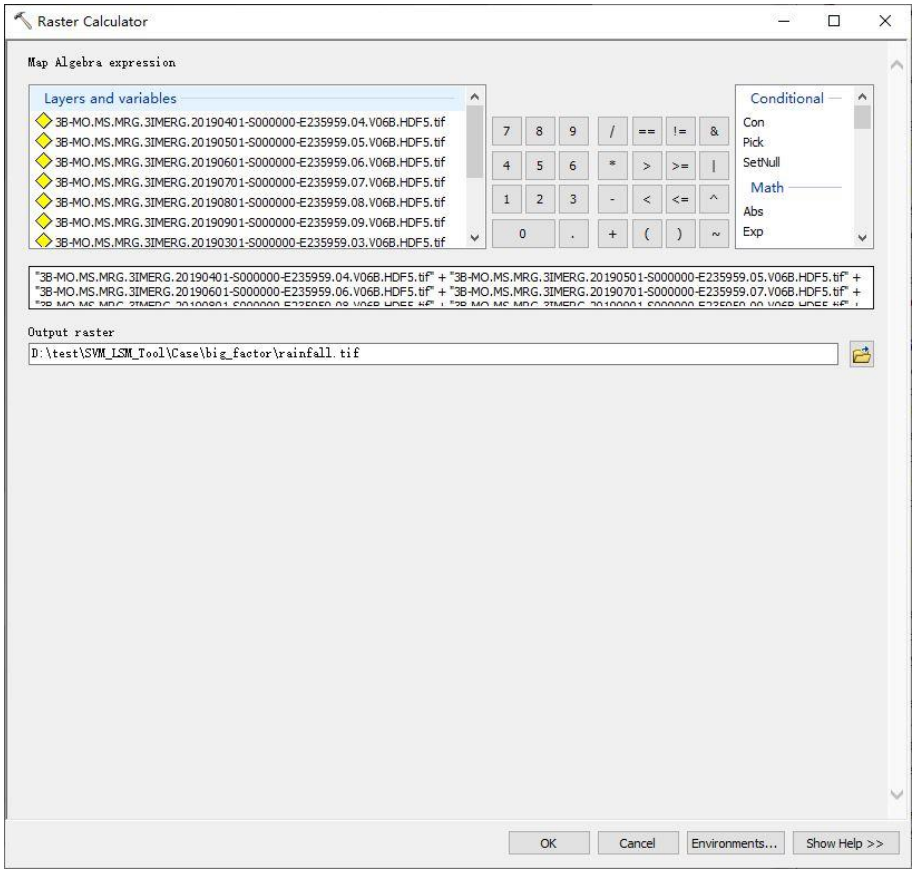

Fig.5 Step4

5. Batch Clipping for each Factor Layer

Use the "4 Batch Clipping for each Factor Layer " tool in the "1 Influencing Factor Production" toolbox. 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 factor is located, and iteratively selects the *.tif* file for clipping automatically. Note: Vector data and raster data must be in UTM coordinate system, and the resolution of raster data must be consistent.

[Input]

- Vector Data of Study Area (UTM Coordinate System): [*\Case\point\study_range.shp*]
- Raster Data Folder (UTM Coordinate System): [*\Case\big_factor*]
- Use Input Features for Clipping Geometry: Unchecked.
- Maintain Clipping Extent: Checked.
- Output Data Folder: [*\Case\factors*]

- Environment Setting : Parallel Processing Factor: 0

[Output]

Batch clipped raster data in the study area. [*\Case\factors*]

[*dem.tif, slp.tif, asp.tif, cur.tif, plancur.tif, profilecur.tif, rivers.tif, roads.tif, lithology.tif, SroughnessC.tif, relief.tif, rainfall.tif, NDVI.tif, TWI.tif*]

Fig.6 Step 5

6. Non-landslide Data Generation

Use the "1 Non-Landslide Data Generation" tool in the "2 Dataset Production and Factor Selected" toolbox to generate non-landslide point data within the vector data 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 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): [*\Case\point\landslide_point.shp*]
- Distance to Landslide Point: 1000 m.

- Vector Data of Study Area (UTM Coordinate System): [*\Case\point\study_range.shp*]
- Number of Points: 789
- Output Folder: [*\Case\point*]
- Output Coordinate System: Unchecked. The default is consistent with the landslide point vector data coordinate system.

[Output]

Non-landslide sample vector data in the study area. [*\Case\point\non_landslide_point.shp*]

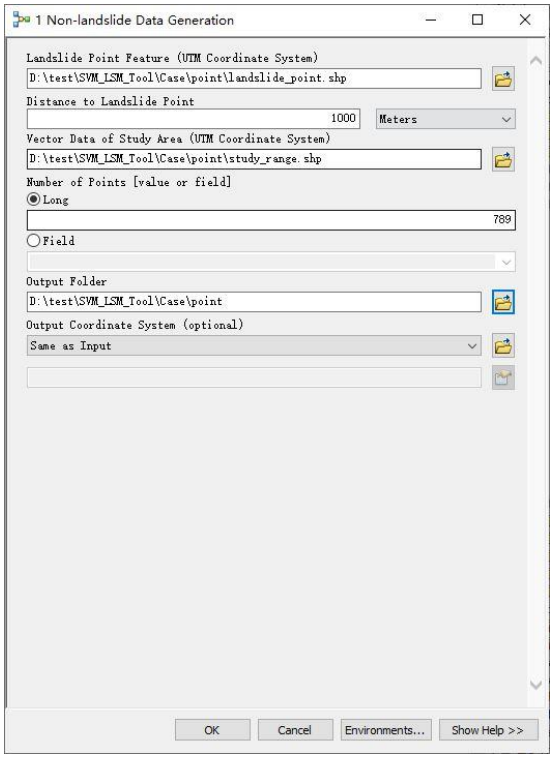

Fig.7 Step 6

7. Image generation to be predicted

Use the "1 Image Generation to be Predicted" tool in the "3 Model Training and Prediction" toolbox, which generates multi-channel raster data for the image to be predicted based on the raster data of each factor layer. It is used for subsequent data sample production and susceptibility map prediction. Note: All factor rasters (*.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: [*\Case\factors*]
- Stacking Factor Layer Order: [*'dem', 'slp', 'asp', 'cur', 'plancur', 'profilecur', 'rivers', 'roads', 'lithology', 'SroughnessC', 'relief', 'rainfall', 'NDVI', 'TWI'*]
- Save Folder of the Image to be Predicted: [*\Case\predict*]

[Output]

The multi-channel raster data of the image is to be predicted in the study area. [*\Case\predict\Factors_14_mapping.tif*]

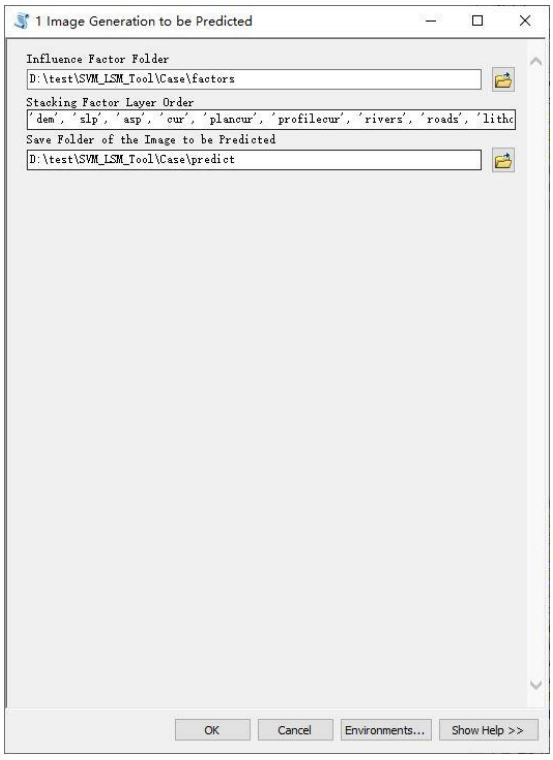

Fig.8 Step 7

8. Data Sample Production

Use the "2 Data Sample Production" tool in the "2 Dataset Production and Factor Selection" toolbox, which 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.

This step is performed twice, for landslide samples and non-landslide samples.

[Input]

- Input Point Feature: the first [*\Case\point\landslide_point.shp*], second [*\Case\point\non_landslide_point.shp*]
- Buffer Distance: 120 m.
- Multi-channel Factor Layer Data: [*\Case\predict\Factors_14_mapping.tif*]
- Data Sample Save Folder: [*\Case\dataset*]
- Sample Label (landslide or non-landslide): the first: *landslide*, second: *non-landslide.*
- Use Input Features for Clipping Geometry: Unchecked.
- Maintain Clipping Extent: Checked.

[Output]

The block data samples of landslide and non-landslide in the study area. In this example, the data size is 15*8*8.

[*\Case\dataset\landslide*, *\Case\dataset\non-landslide*]

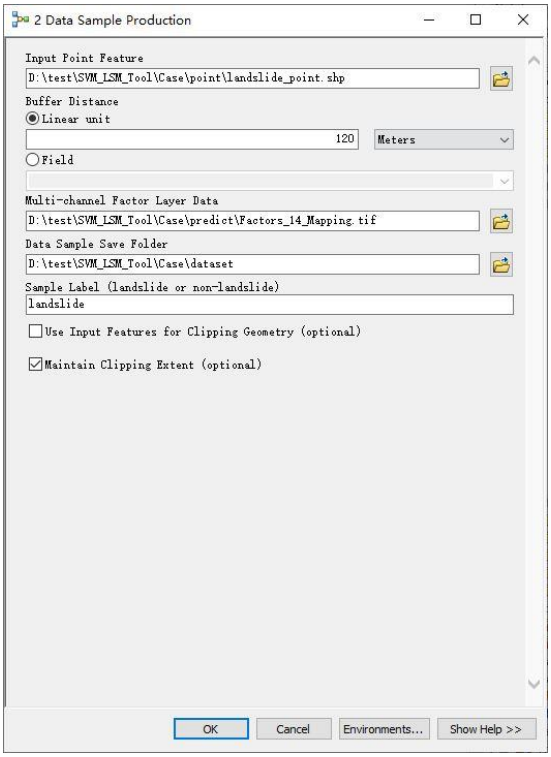

Fig.9 Step 8-1

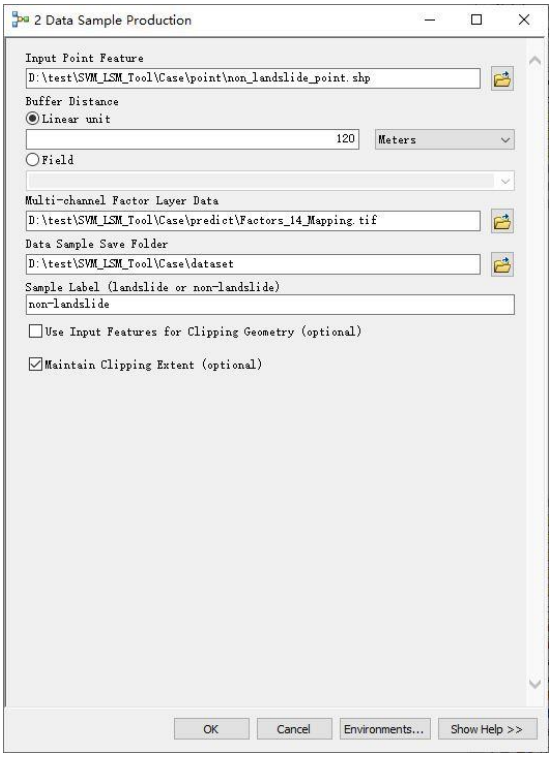

Fig.10 Step 8-2

9. Dataset Split

Use the "3 Dataset Split" tool in the "2 Dataset Production and Factor Selection" toolbox. 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: [*\Case\dataset*]
- Number of Landslides: 789.
- Test Dataset Ratio (e.g. 0.3): 0.3.

[Output]

- [*\Case\dataset.txt*] All sample paths and labels.
- [*\Case\datasettrain.txt*] All training sample paths and labels.
- [*\Case\datasettest.txt*] All test sample paths and labels.

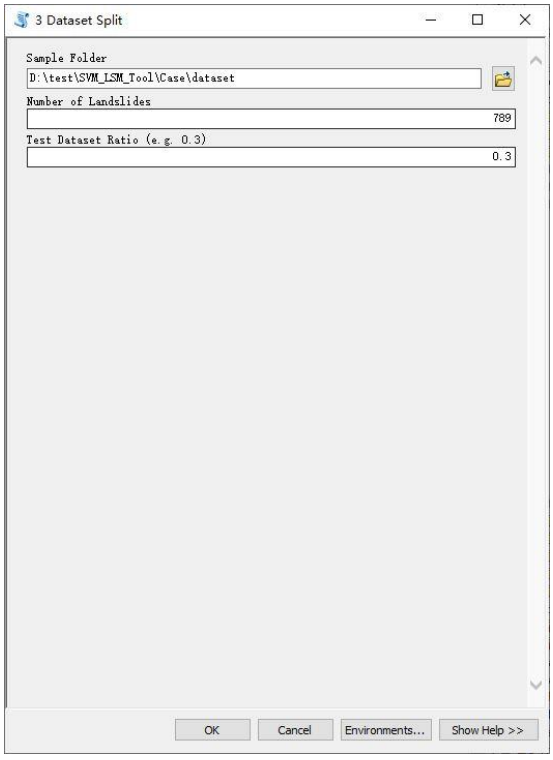

Fig.11 Step 9

10. PCC and IGR Calculation

Using the "4 PCC and IGR Calculation" tool in the "2 Dataset Production and Factor Selection " toolbox. 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 [*\Case\dataset.txt*] file. 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.

[Input]

- Dataset Folder: [*\Case\dataset*]
- Result Save Folder: [*\Case*]
- Input Factor Layer Order: [*'dem', 'slp', 'asp', 'cur', 'plancur', 'profilecur', 'rivers', 'roads', 'lithology', 'SroughnessC', 'relief', 'rainfall', 'NDVI', 'TWI'*]

[Output]

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

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

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

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

Considering the PCC and IGR results comprehensively, the two redundant factors of slope and relief amplitude are removed, and the remaining 12 influencing factors are used for subsequent research.

Fig.12 Step 10

11. Update Image to be Predicted and Samples

According to the selected results in the previous step, the 12 influencing factors are sorted in the order of decreasing information gain ratio, and the image to be predicted and the data samples are regenerated.

Repeat steps 7, 8, and 9. Other options remain the same, except:

- In step 7, the "Stacking Factor Layer Order" is updated to [*'lithology', 'plancur', 'profilecur', 'NDVI', 'TWI', 'asp', 'SroughnessC', 'rivers', 'dem', 'roads' , 'rainfall', 'cur'*]
- In step 8, the "Multi-channel Factor Layer Data" is updated to [*\Case\predict\Factors_12_mapping.tif*], and the "Data Sample Save Folder" is updated to [*\Case\IGR_dataset*]
- In step 9, the "Sample Folder" is updated to [*\Case\ IGR_dataset*]

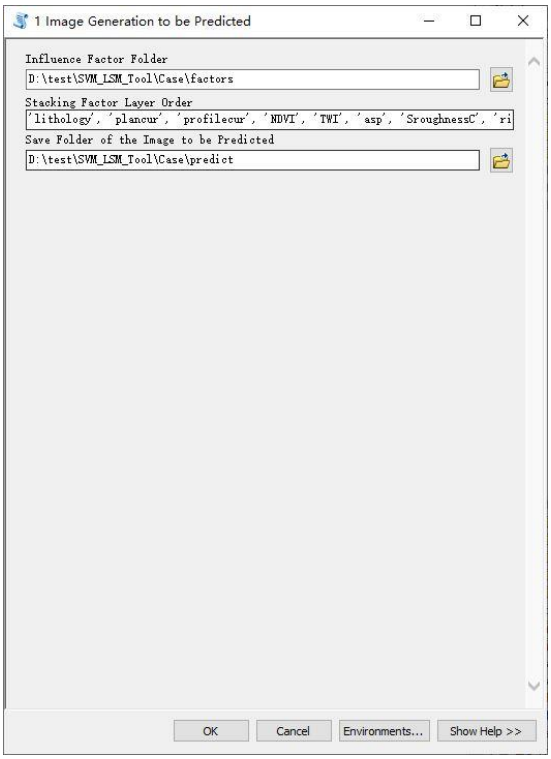

Fig.13 Step 11-1

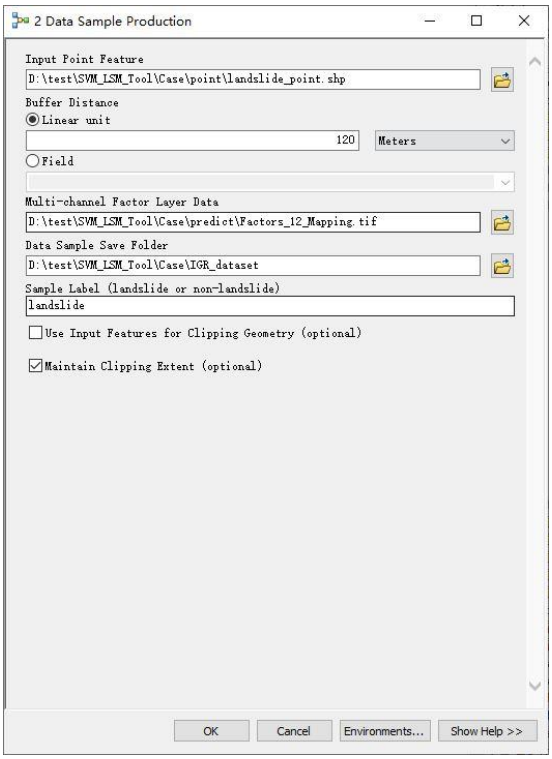

Fig.14 Step 11-2

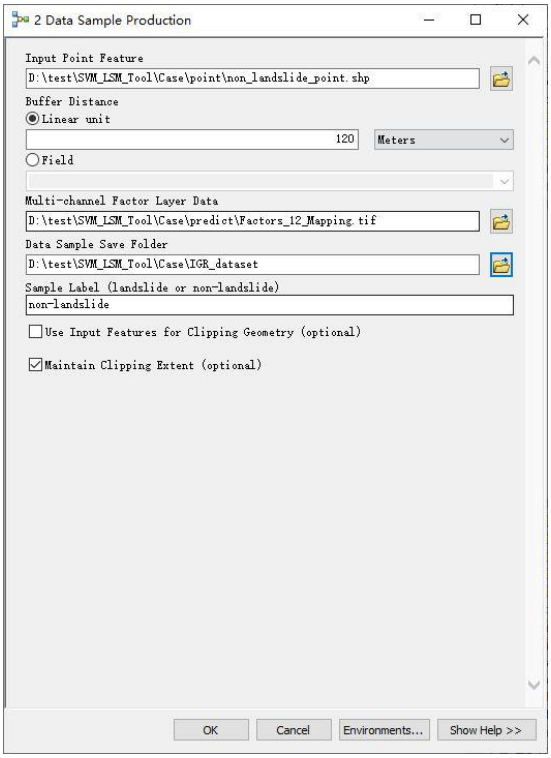

Fig.15 Step 11-3

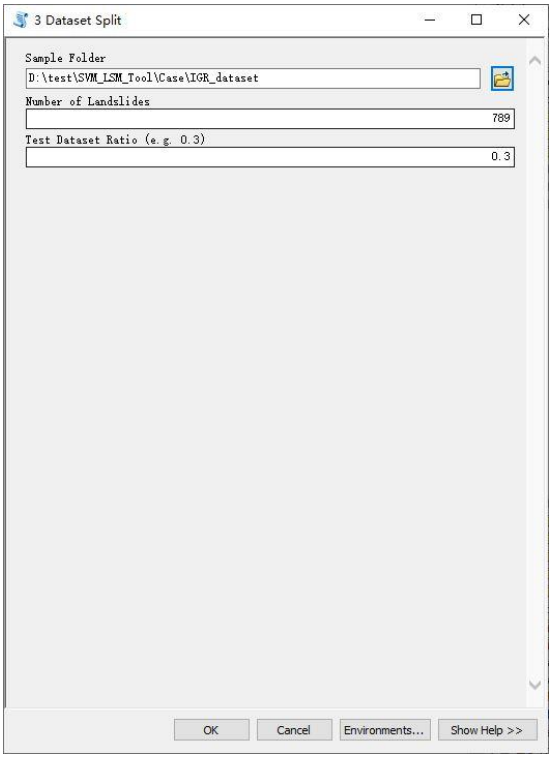

Fig.16 Step 11-4

12. Model Training and Performance Evaluation of SVM

Use the "2 SVM Model Training and Performance Evaluation" tool in the "3 Model Training and Prediction" toolbox, which is used to generate the SVM model under each group of parameters and give the evaluation results 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: [*\Case\IGR_dataset*].
- Model Save Folder: Under this path, different parameter folders will be created with "*g_g value_C_C value*" to save the results. [*\Case\model*]
- gamma Optional Value: 0.01, 0.02, 0.05, 0.08, 0.1, 0.2, 0.5, 0.8, 1, 2, 5.
- C Optional Value: required. 0.01, 0.02, 0.05, 0.08, 0.1, 0.2, 0.5, 0.8, 1, 2, 5.
- Number of Dataset Rows: 8.
- Number of Dataset Columns: 8.
- Number of Dataset Channels: 12.

[Output]

- [*\Case\model*]: save the model and performance evaluation index folder for each group of parameters.
- [*\Case\model\parameter_result_txt.txt*]: save 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.
- [*\Case\model\parameter_result_png.png*] 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.

Take *gamma* as 0.02 and *C* as 2 as an example:

- [*\Case\model\g_0.02_C_2*]: The save path of the SVM result with *gamma* of 0.02 and *C* of 2.
- [*\Case\model\g_0.02_C_2\SVM_g_0.02_C_2.model*]: The SVM model with *gamma* of 0.02, and *C* of 2.
- [*\Case\model\g_0.02_C_2\SVM_train_result_txt.txt*]: The prediction result of SVM model on training dataset with *gamma* is 0.02 and *C* is 2.
- [*\Case\model\g_0.02_C_2\SVM_test_result_txt.txt*]: The prediction result of SVM model on test dataset with *gamma* is 0.02 and *C* is 2.
- [*\Case\model\g_0.02_C_2\evaluate_result.txt*]: Various evaluation indicators of the SVM model on the test dataset, such as confusion matrix, accuracy, precision, F1 value, AUC value etc. with *gamma* is 0.02 and *C* is 2.
- [*\Case\model\g_0.02_C_2\ROC.png*]: The ROC curve and AUC value of the SVM model on the test dataset with *gamma* of 0.02 and *C* of 2.

According to [*\Case\model\parameter_result_txt.txt* and *\Case\model\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.

The optimal model is *g_0.02_C_2* in this example.

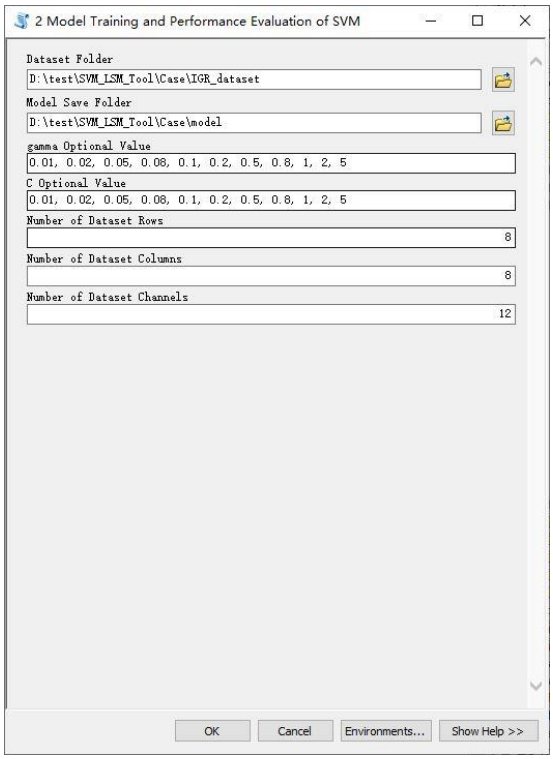

Fig.17 Step 12

13. Landslide Susceptibility Map Prediction

Use the "3 Landslide Susceptibility Map Prediction" tool in the "3 Model Training and Prediction" toolbox, single process or multiprocessing free choice. It can predict 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): [*\Case\predict\Factors_12_Mapping.tif*].
- Vector Data of Study Area (UTM Coordinate System): [*\Case\point\study_range.shp*].
- Optimal Model Folder (.model): [*\Case\predict\model\g_0.01_C_0.8\SVM_g_0.01_C_0.8.model*].
- LSM Output Path: [*\Case\predict\LSM.tif*].
- Number of Dataset Rows: 8.
- Number of Dataset Columns: 8.

- (Multiprocessing parameter) pythonw.exe path: [*C:\Python27\ArcGIS10.8*]

[Output]

Landslide susceptibility map of the study area.[*\Case\predict\LSM.tif*]

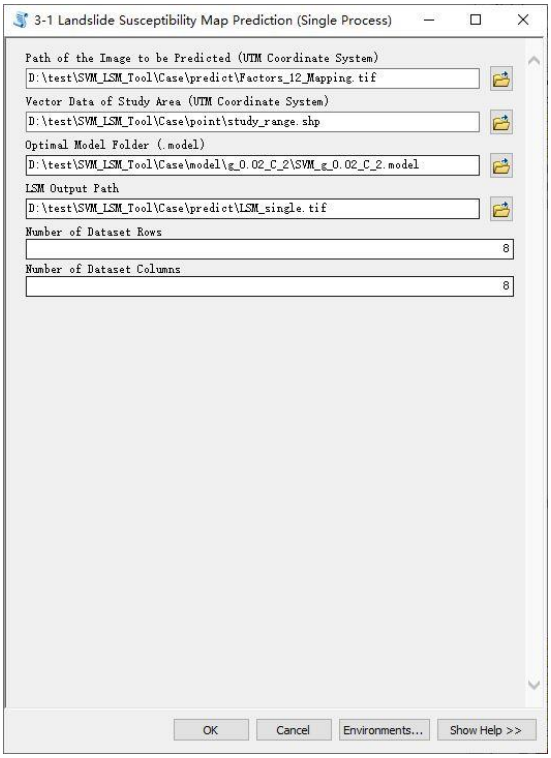

Fig.18 Step 13

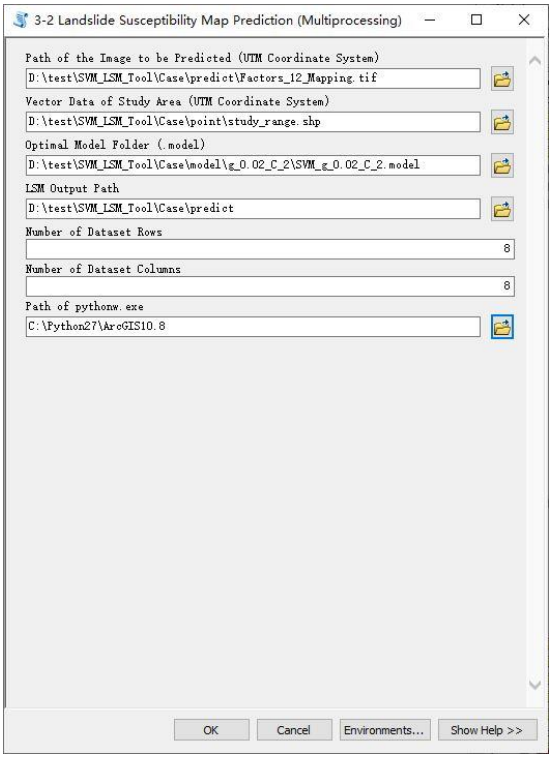

Fig.19 Step 13