

Sample data for simulation in Jing-Jin-Ji region.

1. Sample data of vegetation: The model training and assessment data were obtained from field surveys and publications. They are organized shp format in arcgis 10.3.
2. Vegetation map: The vegetation map was completed in 2007 based on field survey data.
3. Input variable: Geospatial variables, including elevation, slope and aspect, were derived from the 30-m resolution SRTM DEM product (Zhao et al., 2018). These data were resampled to a 500×500 m grid cell size using a nearest-neighbor method in ArcGis 10.3. Climate data including 19 bioclimatic variables of current climate at a 1-km resolution were downloaded from WorldClim (Fick & Hijmans, 2017) at <http://worldclim.org/>. These climate data were also resampled to a 500×500 m grid cell size using a nearest-neighbor method in ArcGIS 10.3. The MYD09A1500M product data (sinusoidal projection, path 4 and row 26, path 4 and row 27, path 5 and row 26, path 5 and row 27) in summer (July 20, 2013) and winter (January 17, 2013), as Modis images, were acquired from the Geospatial data Cloud at <http://www.gscloud.cn/>. Image pre-processing included image subset mosaicking and image clipping according to study area in ENVI 5.2. The land surface albedo in bands 1-7 was directly obtained from the MYD09A1500M product. The vegetation indices was calculated from the land surface albedo in bands 1-7.
4. The classification rules of DT models in this study were with 5 layers, where 40 samples were in the smallest parent node and 10 samples were in the smallest child node.
5. RF model was generated using the default settings in EnMAPBox, with 100 trees, the number of randomly selected features was equal to the square root of the number of all features, and a Gini coefficient was used for the node impurity function.
6. The classification rules of MLC assign every pixel to the class with the highest probability according to the statistics of the Gaussian probability density function.
7. The default settings in EnMAP-Box were applied to the SVM model, where g was from 0.01 to 1000, c was from 0.1 to 1000. Parameters of g and c are tested using grid search with internal performance estimation and those with the best performance are used for the data training.