Hindawi Publishing Corporation
Advances in Meteorology
Volume 2013, Article ID 520803, 12 pages
http://dx.doi.org/10.1155/2013/520803

*Research Article*

# Land Cover Mapping Based on Multisource Spatial Data Mining Approach for Climate Simulation: A Case Study in the Farming-Pastoral Ecotone of North China

**Feng Wu, Jinyan Zhan, Haiming Yan, Chenchen Shi, and Juan Huang**

*State Key Laboratory of Water Environment Simulation, School of Environment, Beijing Normal University, Beijing 100875, China*

Correspondence should be addressed to Jinyan Zhan; zhanjy@bnu.edu.cn

Received 17 May 2013; Accepted 28 June 2013

Academic Editor: Hongbo Su



The land use and land cover change (LUCC) is one of the prime driving forces of climate change. Most attention has been paid to the influence of accuracy of the land cover data in numerous climate simulation projects. The accuracy of the temporal land use data from Chinese Academy of Sciences (CAS) is higher than 90%, but the high-precision land cover data is absent. We overlaid land cover maps from different sources, and the grids with consistent classification were selected as the sample grids. By comparing the results obtained with different decision tree classifiers with the WEKA toolkit for data mining, it was found that the C4.5 algorithm was more suitable for converting land use data of CAS classification to land cover data of IGBP classification. We reset the decision rules with Net Primary Productivity (NPP) and Normalized Difference Vegetation Index (NDVI) as the indicators. The accuracy of the reclassified land cover data was proven to reach 83.14% through comparing with the Terrestrial Ecosystem Monitoring Sites and high resolution images. Therefore, it is feasible to produce the temporal land cover data with this method, which can be used as the parameters of dynamical downscaling in the regional climate simulation.

## 1. Introduction

The land cover change, which plays an important role in the climate system at the global, regional, and local scales, contributes to the climatic change and variability [1]. With the progress of the research on the climatic modeling over the past decade, it has been widely recognized that there is a more urgent need to accurately characterize the land surface as the boundary conditions in the climate modeling [2–6]. The precise contribution of the land cover change to the global climate change remains a controversial but growing concerned issue. Many land cover data of China have been produced in recent years with the remote sensing data. The previous study showed that the result of the precipitation study would be greatly influenced if the accuracy of land cover data is under 80%, and the result may be worse as the accuracy continues to decrease [7]. Unfortunately, neither the overall nor class-specific accuracy of most datasets can meet the common requirements of the regional climate modeling. Therefore, it is necessary to produce the land cover dataset with high accuracy for the climate simulation based on the existing land use dataset, land cover datasets, and some ancillary datasets. These available data with a high level of uncertainty may be improved by combining the different data sources so as to meet the requirement of the climate simulation.

The researches on the climate modeling vary substantially in the spatial and temporal scales. So the temporal land cover datasets are essential to the development of the cohesive climate model. The Chinese Academy of Sciences (CAS) has constructed a land use dataset that includes the data of 1988, 1995, 2000, and 2005 [8–10]. However, there are still no comparisons of land cover datasets at the regional scale, especially in China where the land use is changing drastically due to the rapid economic development and the anthropogenic disturbance. Many studies have indicated that the disagreement among the land cover datasets

primarily resulted from the differences in the sensors, spatial resolutions, algorithms, and classification schemes [11, 12]; among them, the difference in the classification schemes was considered to be the key reason for the disagreement of the land cover datasets and the main obstacle to comparing the data from different land cover datasets [13, 14]. Therefore, great contribution may be made to climate change research if we can take full advantage of the long-term land use datasets from the CAS and use an appropriate method to convert them to the International Geosphere Biosphere Programme (IGBP) land cover classification scheme which consists of seventeen categories (Table 1) and is widely accepted and used in the simulation of climate changes [15, 16].

The decision tree is one of the most powerful classification algorithms to classify land cover type of remote sensing image [17, 18]. The decision tree technique is more suitable for the analysis of the categorical outcomes. Besides, it is easy to interpret, computationally inexpensive, and capable of dealing with the noisy data. Moreover, its prediction model is more understandable to the users. In addition, it can find the significant high-order interactions quickly with the automatic interaction detection, and it can produce more informative outputs [19–21]. The decision tree classifiers include the C4.5/C5.0/J48, NBTree, SimpleCart, REPTree, and BFTree, among which the C4.5/C5.0/J48 classifier is the most popular and powerful one [22, 23]. The C4.5 classifier was selected in this study according to the accuracy assessment to identify the vegetation disaggregation classification in the farming-pastoral ecotone of North China.

The ecotones are recognized as one of the most important objects of the ecological research, since they are unstable and very sensitive to the surrounding environment [24]. Besides, the ecotones are more suitable for the study of the land cover mapping for the climate simulation. The farming-pastoral ecotone has received a lot of attention from the academic community due to its largest area, longest span, and typical characteristics [25]. It involves 9 provinces and 106 counties, with a total area of 654, 564 $km^2$ [26]. The total population in this area is $3.14 \times 10^7$, with the average population density of 47.9 persons per square kilometer. The land use has changed drastically throughout the farming-pastoral ecotone of North China after the widespread and profound economic reform that was initiated in the early 1980s [27, 28], and the current ratio of the cropland, forest land, and grassland is 1.0 : 1.17 : 3.67 (Figure 1). The temperature rise has been more and more obvious in the past 50 years, with an average increase rate of 0.4°C/10a [29]. Therefore, more attention shall be paid to the interaction between the land cover change and climate change during the control of the eco-environment degradation in the ecozone.

This article is organized as follows. Section 1 discussed the significance of the land cover to the climate simulation and introduced the objectives of this study. Section 2 introduced the input and reference data, and Section 3 presented the spatial data mining approach. Section 4 analyzed the results, along with an evaluation of the accuracy and uncertainty of the obtained map in comparison with other land cover maps. Section 5 discussed the findings and concluded.

## 2. Data Preparation

This paper presents an inference rule of spatial data mining to distinguish forest types based on the consistent grids in the data of the International Geosphere-Biosphere Programme Data and Information System (IGBPDIS: https://lpdaac.usgs.gov/products/modis_products_table/mcd12q1 [30], Global Land Cover2000 (GLC2000: http://bioval.jrc.ec.europa.eu/products/glc2000/glc2000.php [31], Multisource Integrated Chinese Land Cover (WESTDC: http://westdc.westgis.ac.cn/) [32], and UMD (http://www.landcover.org/) land cover data in 2000. The classification rule was first rectified so as to improve the accuracy in 2000. Then the land use data of 1988, 1995 and 2005 were converted to the land cover data according to this inference rule.

In this paper, we also used the land use database developed by the Chinese Academy of Sciences (CAS). The data are available during four periods, that is, year 1988, year 1995, year 2000, and year 2005. A hierarchical classification system of 25 land cover classes was applied to the data. The data team also spent considerable time validating the precision of the interpretation of TM images and land cover classification by extensive field surveys (ground validation). The validation result indicated that the average precision of the interpretation reached 95% [33]. The 1 km land use map of China was derived from the 1 : 100,000 land use database. It includes two kinds of data; one was geocoded with the greatest-area method (i.e., if a cell has more than one possible code or it contains two or more polygons, the code of the polygon with the greatest area in the cell is used). The other was geocoded with area percentage grid method, in which each cell can be divided into 25 layers to record the area of each type [10]. Besides, the vegetation map can provide the reference information of vegetation since the change of forest categories is slight in the short term. The vegetation map of China reflects detailed information on the distribution of vegetation and includes the horizontal and vertical zones of 11 vegetation groups, 54 vegetation types, 135 biome units, and 796 subbiome units [34].

The mapping of land cover data in 2000 based on the data mining is a benchmark of the long-term land cover dataset. It is necessary to collect the ancillary data due to the absence of other data series. Data of physical geography include information on terrain slope and information on vegetation property variability. Information on the terrain slope and the plain area proportion were derived from DEM data covering the entire China at the scale of 1 : 250,000. These data were provided by the Data Center for Resources and Environmental Sciences Chinese Academy of Sciences. The meteorological data, including the annual temperature and annual precipitation, were acquired from China Meteorological Bureau. The NDVI dataset came from the Pathfinder dataset of Earth Resources Observation System (EROS); it was extracted from the NOAA/AVHRR-NDVI images. The spatial resolution of the images is 1 km × 1 km, and their temporal resolution is 15 days. In order to guarantee the data quality, all the data have all been preprocessed with the internationally accepted reliable approach [35]. Besides, in order to eliminate the noise caused by the cloud pollution

Table 1: Types and descriptions of IGBP land cover classification scheme.

| Code | Type | Descriptions |
|---|---|---|
| 1 | Evergreen needle leaved forest | Lands dominated by trees with a per cent canopy cover >60% and height exceeding 2 m. Almost all trees remain green all year. Canopy is never without green foliage. |
| 2 | Evergreen broad leaved forest | Lands dominated by trees with a per cent canopy cover >60% and height exceeding 5 m. Almost all trees remain green all year. Canopy is never without green foliage. |
| 3 | Deciduous needle leaved forest | Lands dominated by trees with a per cent canopy cover >60% and height exceeding 2 m. Consists of seasonal needle leaved tree communities with an annual cycle of leaf-on and leaf-off periods. |
| 4 | Deciduous broad leaved forest | Lands dominated by trees with a per cent canopy cover >60% and height exceeding 2 m. Consists of seasonal broad leaved tree communities with an annual cycle of leaf-on and leaf-off periods. |
| 5 | Mixed forests | Lands dominated by trees with a per cent canopy cover >60% and height exceeding 2 m. Consists of tree communities with interspersed mixtures or mosaics of the other four forest cover types. None of the forest types exceeds 60% of the landscape. |
| 6 | Closed shrublands | Lands with woody vegetation less than 2 m tall and with shrub-canopy cover >60%. The shrub foliage can be either evergreen or deciduous. |
| 7 | Open Shrublands | Lands with woody vegetation less than 2 m tall and with shrub canopy cover between 10–60%. The shrub foliage can be either evergreen or deciduous. |
| 8 | Woody savannas | Lands with herbaceous and other understorey systems and with forest canopy between 30–60%. The forest cover height exceeds 2 m. |
| 9 | Savannas | Lands with herbaceous and other understorey systems and with forest canopy between 10–30%. The forest cover height exceeds 2 m. |
| 10 | Grasslands | Lands with herbaceous types of cover. Tree and shrub cover is less than 10%. |
| 11 | Permanent wetlands | Lands with a permanent mixture of water and herbaceous or woody vegetation that cover extensive areas. The vegetation can be present in either salt, brackish, or fresh water. |
| 12 | Croplands | Lands covered with temporary crops followed by harvest and a bare soil period (e.g., single and multiple cropping systems). Note that perennial woody crops will be classified as the appropriate forest or shrubs land cover type. |
| 13 | Urban and built up | Land covered by buildings and other man-made structures. Note that this class will not be mapped from the AVHRR imagery but will be developed from the populated places layer that is part of the digital chart of the world. |
| 14 | Cropland/natural vegetation mosaic | Lands with a mosaic of croplands, forest, shrublands, and grasslands in which no one component comprises more than 60% of the landscape. |
| 15 | Snow and ice | Lands under snow and/or ice cover throughout the year. |
| 16 | Barren or sparsely vegetated | Lands of exposed soil, sand, rocks, or snow and never have more than 10% vegetated cover during any time of the year. |
| 17 | Water bodies | Oceans, seas, lakes, reservoirs, and rivers. Can be either fresh or salt water. |

and the atmospheric influence, we have also smoothed the time-series NDVI data with the Savitzky-Golay smoothing filtering method [36]. The NPP data during 1985–1999 came from the remote sensing data of NOAA/AVHRR and that during 2000–2010 came from the NPP product of MODIS.

## 3. Methodology

The working procedure of the classification is as follows. First, based on the definition of mosaics type, we produced the cropland/natural vegetation mosaics data using the grid area percentage dataset in CAS land use system. Then other types of land use except for forest and woods were achieved by utilizing grid maximum area mapping with two subclassification definition between the CAS and IGBP. Thereafter, we checked out and determined the grids whose types were consistent with the forest and woods among the WESTDC, UMD, GLC, and IGBPDIS land cover data; at the same time we identified the boundary of the forest and woods, which were consistent with CAS land use, that generated them into the sample data. Finally, we realized the conversion of forest types of IGBP scheme with the C4.5 classifier (Figure 2).

*3.1. Mapping the Land Use Types to Determinate the Land Cover Classification.* The land use types were first transformed into the land cover types. It is easy to transform some land use types, for example, 3 classes of developed and mosaic lands, 2 classes of artificial lands, and 1 class of water among the IGBP land cover classification.

It only needs to transform from many to one or one to one (Table 2). For example, the paddy land and dry land in the land use map of CAS are explicit and correspond to the cropland class definition in the IGBP, so it only needs to aggregate

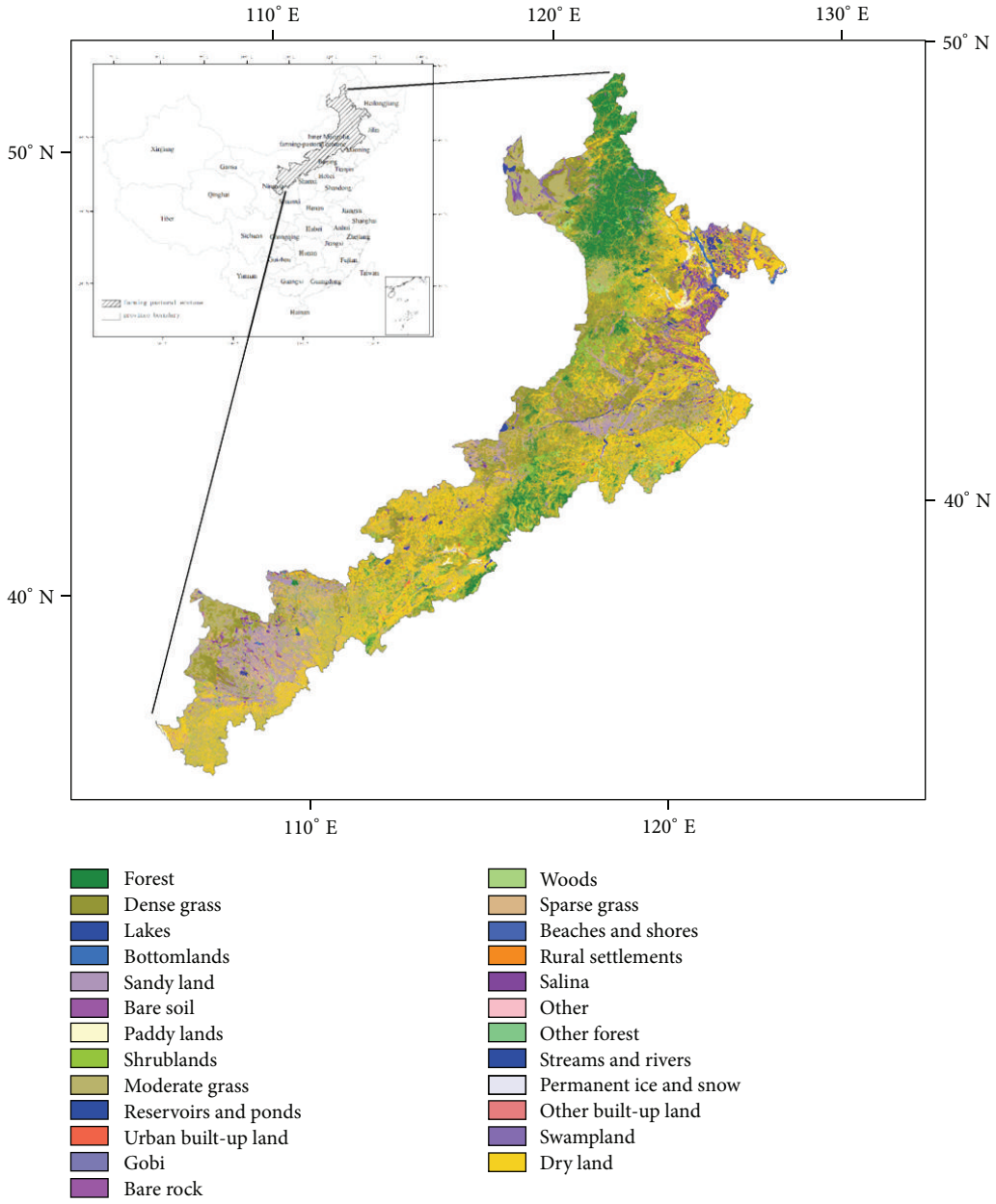


Figure 1: The location and 2000 years' land use map of farming-pastoral ecotone in North China.

them into cropland with the binary grid. It is more feasible to judge the land cover classification of cropland/natural vegetation mosaic with the area percentage grid data of Paddy lands, dry lands, forest, shrublands, among which no single type comprises more than 60% of the landscape. The land cover of cropland/natural vegetation mosaic is mainly located in the Inner Mongolia, Liaoning, Hebei, Shaanxi, Shanxi provinces, with a total area of about 730,00 km$^2$ in 2000 (Figure 3). The 8 classes of land cover types including the IGBP10-IGBP17 were transformed, which account for nearly half of the total land area. In addition, there is a little savanna in China, which is convenient to judge based on the temperature and land use type. However, the 8 classes of vegetation (forest, shrubs, and herbaceous vegetation), the leaf attributes (evergreen and deciduous), and the leaf types (broadleaved and coniferous) are difficult to determine because we lack the information of vegetation.

*3.2. Selecting the Spatial Agreement Samples of Vegetation for Data Mining.* The closed forest and other forest classes are arbor forest classes in land use classifications of CAS. They do not concretely specify the forest type information. However, this provides an accurate boundary for the forest; therefore, we need an inference rule to transform between forest in the land use classification system and IGBP forest categories: evergreen needle-leaf forest, evergreen broadleaved

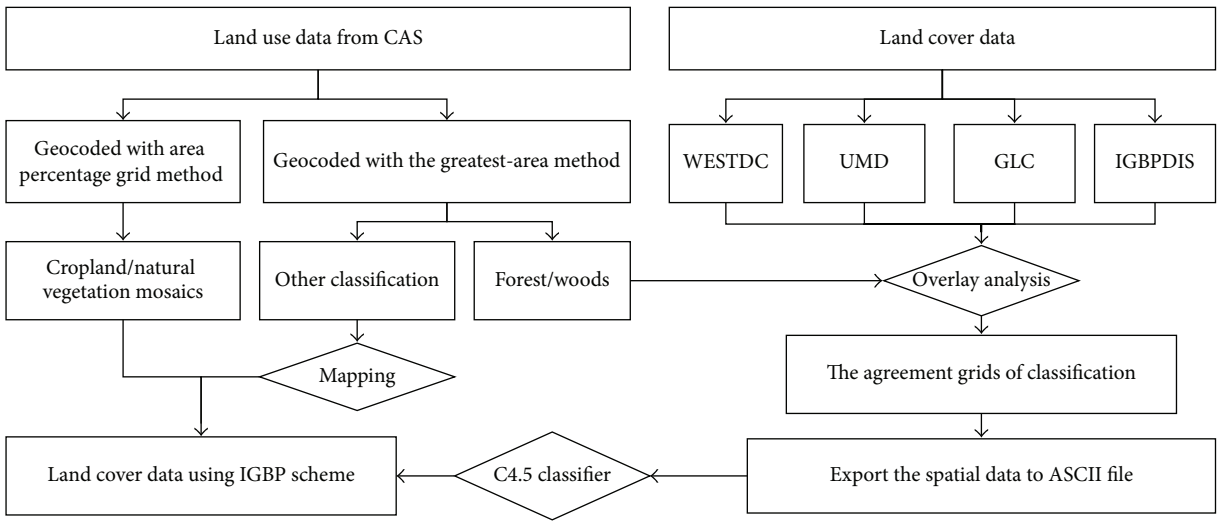


Figure 2: The work flow of mapping based on multisource spatial data mining approach.

Table 2: Comparison between the CAS land use classification scheme and IGBP land cover classification scheme.

(a)

| | Paddy land | Dryland | Stream and rivers | Lakes | Reservoirs and ponds | Permanent ice and snow | Beach and shores | Bottomland | Dense grass | Moderate grass |
|---|---|---|---|---|---|---|---|---|---|---|
| IGBP12 | ☆ | ☆ | | | | | | | | |
| IGBP15 | | | | | | ☆ | | | | |
| IGBP17 | | | ☆ | ☆ | ☆ | | ☆ | ☆ | | |
| IGBP10 | | | | | | | | | ☆ | ☆ |

Note: ☆ stands for match.

(b)

| | Urban built up | Rural settlement | Other built up | Sandy land | Gobi | Salina | Swamp land | Bare soil | Bare rock | other |
|---|---|---|---|---|---|---|---|---|---|---|
| IGBP11 | | | | | | | ☆ | | | |
| IGBP13 | ☆ | ☆ | ☆ | | | | | | | |
| IGBP16 | | | | ☆ | ☆ | ☆ | | ☆ | ☆ | ☆ |

Note: ☆ stands for match.

forest, deciduous needle-leaf forest, deciduous broadleaved forest, and mixed forest based on the ancillary data in 2000.

The degree of overlap between any two land cover classes based on the feature definitions of the classification schemes was used to select the sample grids among the IGBPDIS, WESTDC, UMD, and GLC data [37]. The degree of agreement for each grid was determined by the overlap metric, which indicates the feature-based similarity among different land cover products. If the classes of the two products are identical or mostly overlapped for a given grid, then the grid will be assigned a value of 1, which indicates that the two classes of the different classification schemes completely agree with each other. Otherwise, the grid will be assigned a value of 0. Finally, the agreement and disagreement maps will be created over the entire region, which highlight the areas that have a high confidence of classification (Figure 4). In other words, the sample grids could be selected from the agreement degree maps.

In this study, the method improves the classification results by further applying the data mining technique and using the ancillary information. The detailed DEM data, NDVI, NPP, and meteorological data were utilized as the ancillary information to separate the vegetation classes, which have very different ecological characteristics. The vegetation types are closely related to the physiographic factors and meteorological conditions. The topography at every grid could be described by landform classes (e.g., hill, slope, depression etc.) by processing the raw elevation data, and the meteorological data of observation could be interpolated to 1 km grid cell. Therefore, these datasets could be expressed with the 1 km grid data. The additional information sources were used to refine the result of the C4.5 classifier. We overlapped land cover maps and these ancillary data and

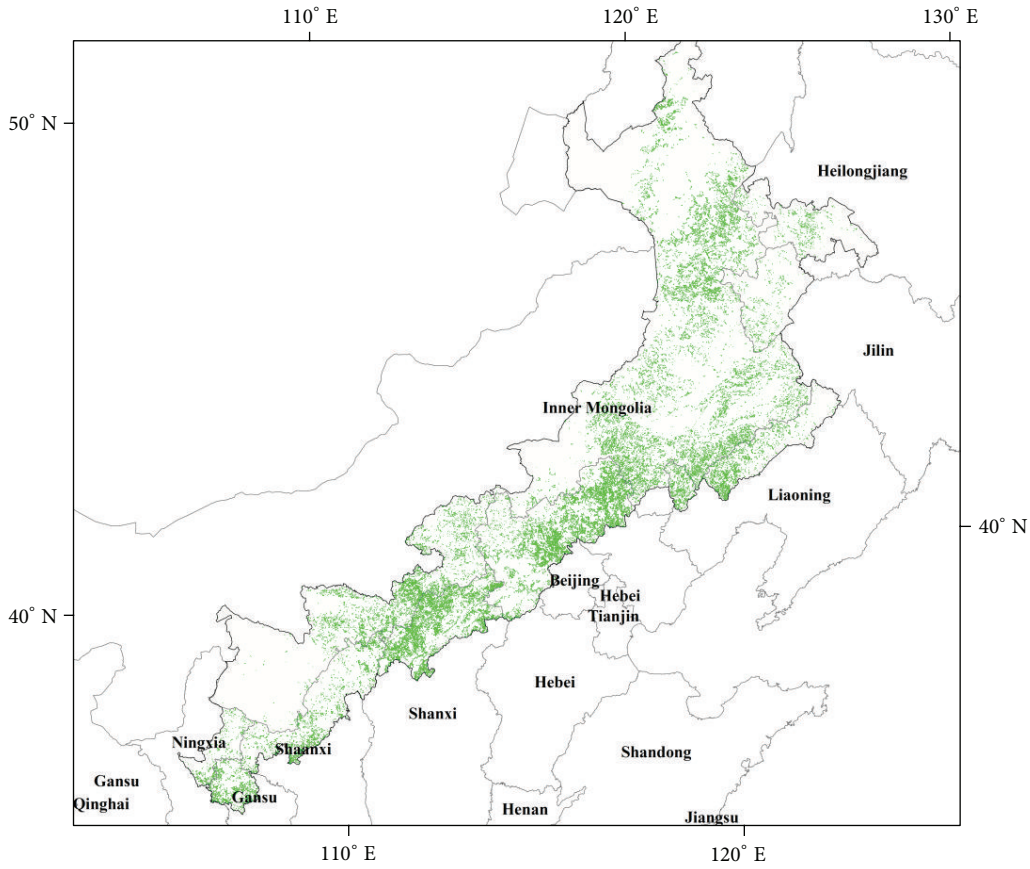


Figure 3: The distribution of cropland/natural vegetation mosaics in 1 km grid all over North China.

sampled the dataset for ASCII text format with the ArcInfo WorkStation toolkit. Thereafter the dataset for the training and testing the classifier of data mining in the WEKA toolkit was constructed.

*3.3. Constructing the Classification Method to Identify Vegetation Types.* Many classification methods have been proposed by researchers in the fields of machine learning, pattern recognition, and statistics. In this study, we focused on the classification methods to convert the forest and grassland classification to the IGBP land cover scheme. In this case, the hidden and valuable knowledge discovered in the related ancillary databases was summarized in the decision tree structure. This classification with the decision tree technique can be performed without complicated computation, and this method can be used for both the continuous and categorical variables. We found that the C4.5 classifier achieved the highest accuracy among these methods for the land cover identification. The classifier was developed on the basis of the decision tree learning, which is a heuristic, one-step lookahead (hill climbing), nonbacktracking search through the space of all possible decision trees. The specific principles of this classifier are as follows. First, the initial sample data were recursively partitioned into subgroups. Then the gain values of all the attributes of the sample data were calculated, according to the numerical value of which the attributes used in the classification were selected. Next, the attribute with the largest gain value was used in the logical test, and each test forms a branch, and the subsets of samples (training data) satisfying the outcomes at the child nodes were moved to the corresponding child nodes. Thereafter, this process runs recursively on each child node until the needed leaf nodes were obtained. Finally, the decision tree was modified according to the relevant empirical knowledge. The C4.5 classifier is one of the decision tree families that can produce both decision tree and rule sets; the C4.5 classifier uses two heuristic criteria to rank the possible tests, that is, the information gain that uses the attribute selection measure, which minimizes the total entropy of the subset $S_i$ and the default gain ratio that divides the information gain by the information provided by the test outcomes. The algorithm of gain ration is described as the function Gain ($A$), which was shown as follows.

(1) The attribute with the highest information gain is selected.

(2) $S$ contains $S_i$ tuples of the class $C_i$ ($i = 1, \ldots, m$). $m$ means the number of classification.

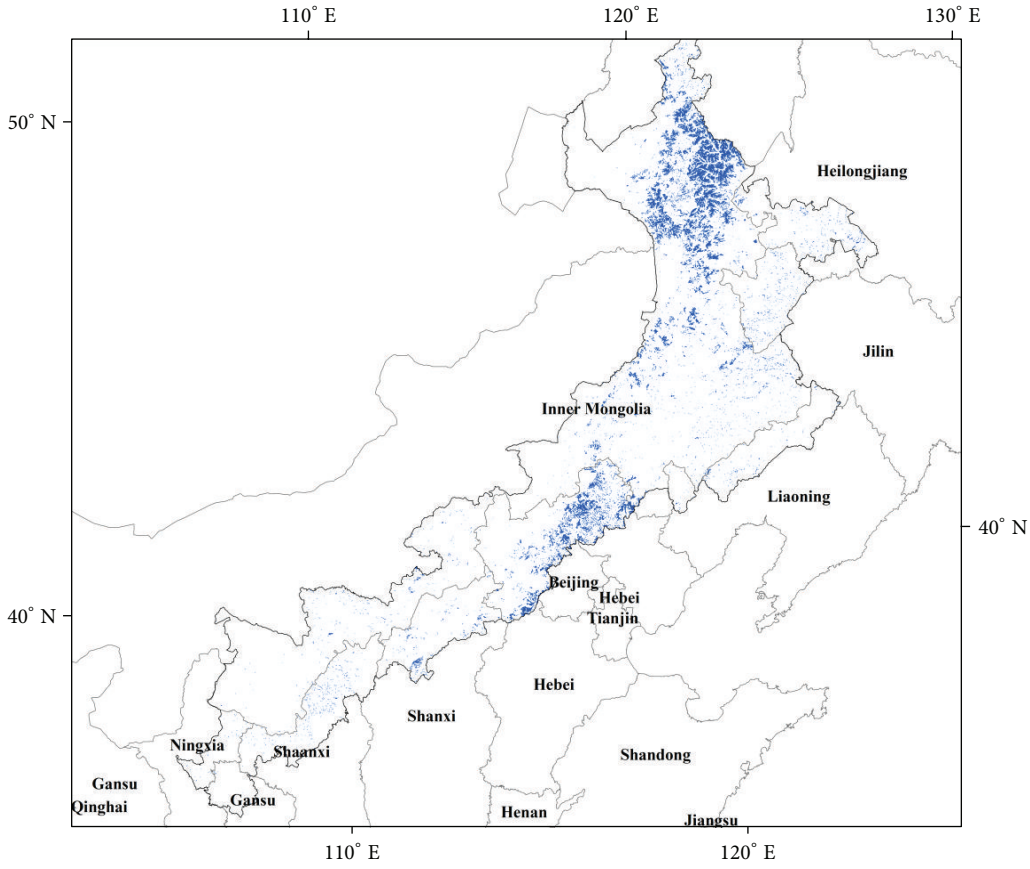


Figure 4: The agreement grids of classification among GLC, UMD, IGBPDIS, and WESTDC.

(3) The information measure or expected information is required to classify any arbitrary tuple:

$$I(S_1, \ldots, S_m) = -\sum_{i=1}^{m} \frac{S_i}{S} \log_2 \frac{S_i}{S}. \quad (1)$$

(4) Entropy of attribute $A$ with values $\{a_1, a_2, \ldots, a_v\}$ was calculated.

$$E(A) = \sum_{j=1}^{v} \frac{S_{1j} + \cdots + S_{mj}}{S} I\left(S_{1j}, \ldots, S_{mj}\right). \quad (2)$$

(5) The information gain means how much can be gained by branching on the attribute $A$:

$$\text{Gain}(A) = I(S_1, S_2, \ldots, S_m) - E(A). \quad (3)$$

The attribute $A$ contains the DEM, longitude, latitude, annual temperature, annual precipitation, NPP, NDVI, and other ancillary spatial data. We calculated the gain ratio to select the attributes that can be used to generate the ancillary information of classification (Table 3). There are about 35396 sample cells of the closed forest and other forest. The gain ratio for the training dataset was calculated, the biggest value of which is 0.27, indicating that NDVI-12 is the most suitable to be the attribute for the forest categories. The forest was further divided into two subcategories according to the NDVI-12 and NDVI-3; that is, the forest with the NDVI-12 reaching 0.53 and NDVI-3 reaching 0.39 was categorized into the evergreen forest, while the forest with the NDVI-12 below 0.53 and NDVI-3 below 0.39 was categorized into the deciduous forest. Although the gain ratio of DEM and temperature is higher than that of the NPP, it is difficult to distinguish the forest type according to them. Therefore, we distinguished the broadleaved, the needleleaved, and mixed forest according to the NPP. The NPP of the broadleaved forest was more than 445, and that of the needle-leaved forest was less than 297, and the forests with the middle NPP value was categorized into the mixed forest.

The accuracy of different classifiers was compared with the WEKA toolkit. We reset the decision tree rule using the NPP and NDVI according to the aforementioned information. The WEKA toolkit is a collection of machine learning algorithms for data mining tasks. It contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. It is also very suitable for developing new machine learning schemes.

## 4. Result and Discussion

*4.1. Evaluating the Accuracy of the Land Cover Classification.* Using the method mentioned previously, a Serving Climate

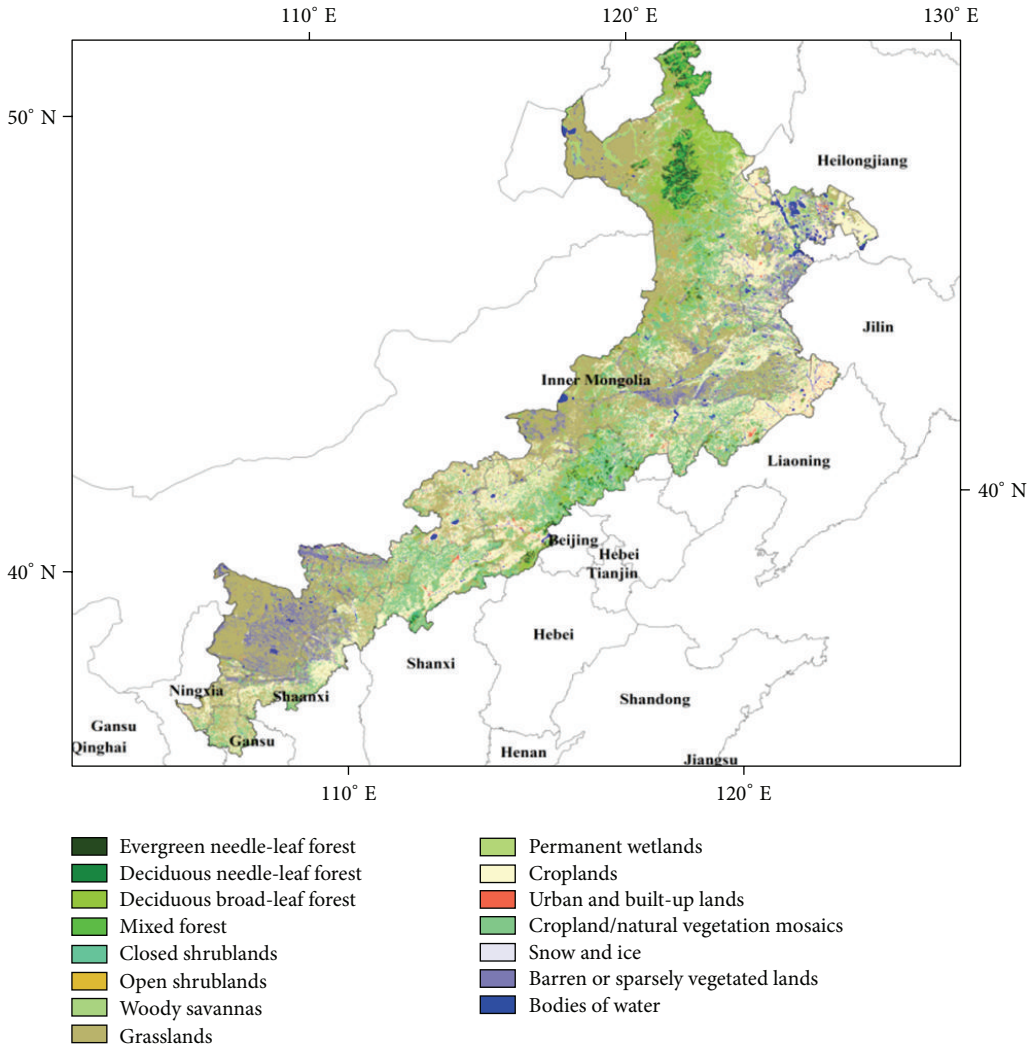


Figure 5: The transformed land cover map over farming-pastoral ecotone in 2000.

Simulation Land Cover (SCSLC) map was generated with a decision rule based on multisource spatial data mining in the farming-pastoral ecotone of North China (Figure 5). To analyze the characteristics of this map, we compared the area of each land cover class in this map with other three popular land cover maps, that is, the WESTDC map, UMD map, and GLC map. The overall areas of each land cover class in the four maps were shown according to the same classification (Table 4). It is notable that the SCSLC map using the C4.5 classifier is similar to the WESTDC map, but there is remarkable increase in the cropland/natural vegetation mosaics and the corresponding decrease in grassland. We also found that the accuracy of the GLC map and UMD map is lower than that of the SCSLC and WESTDC. The GLC map ignores the urban and built-up land, and the UMD ignores the water bodies in the farming-pastoral ecotone of North China, but the two kinds of land cover types are vital to the climate simulation.

Throughout the classification process, the accuracy of the classification maps was assessed by a set of 35396 sample points selected with the stratified random sampling method; these sampling points were randomly selected for each of the classes in the first generated classification map in this research. For each map, a confusion matrix was created, and the accuracy was measured. The use of measurements such as the overall accuracy, Kappa statistics, producer's accuracy and user's accuracy have been quite common and have been explained in detail in numerous publications. The confusion matrix is constructed with the land cover data using the decision rule and the large scale land cover mapping with the integration of multisource information, which is recognized as the real data. The result indicated that an overall accuracy of 88.62% was achieved, which suggested that it gained about a 17.62% increase in accuracy in comparison to the WESTDC map (Table 5).

In addition, we drew the Receiver Operating Characteristic (ROC) curve of each forest classification decision rule using the WEKA. The true positive rate (sensitivity) is plotted in the false positive rate (1-Specificity) function for different cut-off points in the ROC curve. Each point in the ROC

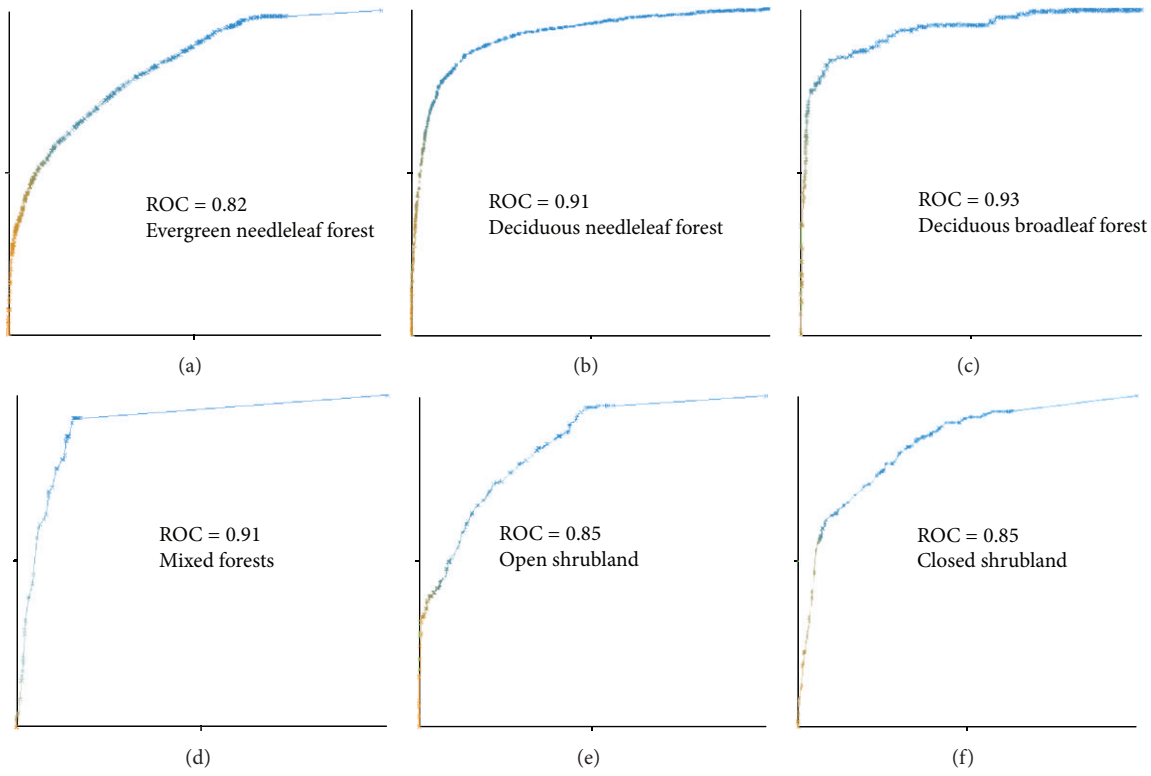


Figure 6: The ROC curve value of different vegetation class rule.

Table 3: The attribute gain ratio value for constructing decision tree.

| Name | Gain ratio | Rank | Description |
|---|---|---|---|
| X | 0.03 | 8 | Rectangular coordination of longitude |
| Y | 0.22 | 2 | Rectangular coordination of latitude |
| PA | 0.12 | 6 | 0.1 mm annual precipitation |
| TA | 0.20 | 3 | 0.1°C annual accumulated temperature |
| DEM | 0.15 | 4 | Elevation |
| LFM | 0.11 | 7 | Landform type |
| NDVI-3 | 0.22 | 2 | Normalized differential vegetation index in March |
| NDVI-12 | 0.27 | 1 | Normalized differential vegetation index in December |
| NPP | 0.14 | 5 | (g C/m$^2$/year) net primary productivity |

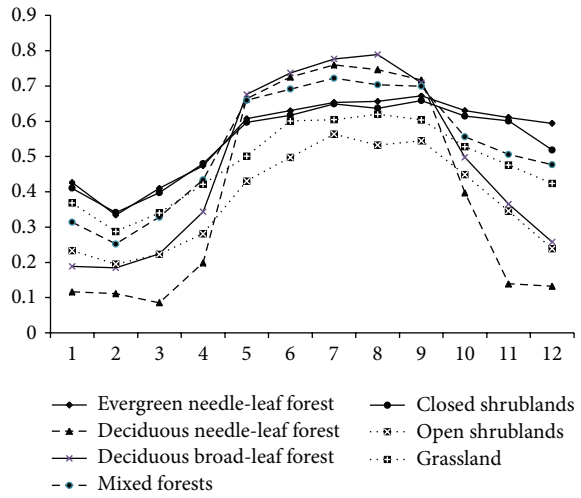


Figure 7: The NDVI characteristics of different vegetation types.

curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. A test with the perfect discrimination (no overlap in the two distributions) was carried out on the ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). The closer to the upper left corner the ROC curve is, the higher the overall accuracy of the test is. The area under ROC curve (AUC) for evergreen needleleaved forest, deciduous needleleaved forest, deciduous broadleaved forest, mixed forest, open shrub land, and closed shrub land were 0.82, 0.91, 0.93, 0.91, 0.85, and 0.85, respectively (Figure 6). The biggest value of AUC was assigned to the evergreen broadleaved forest, indicating that the result gained by the evergreen broadleaved forest should be better than the other four models.

*4.2. Validation with the Ground Reference Data.* It is difficult to carry out the validation of the large-scale map for all land cover types in all regions due to the lack of reference data

Table 4: Comparison of different classification among the IGBPDIS, SCSLC, WESTDC, and UMD products (unit $km^2$).

| Class name | WESTDC | SCSLC | GLC | UMD |
|---|---|---|---|---|
| Evergreen needle leaved forest | 13174 | 11532 | 19250 | 9075 |
| Deciduous needle leaved forest | 312 | 4309 | 25702 | 10819 |
| Deciduous broad leaved forest | 36330 | 32120 | 33971 | 67687 |
| Mixed forests | 6680 | 3434 | 123 | 3529 |
| Closed shrublands | 7219 | 4878 | 16212 | 24049 |
| Open shrublands | 4198 | 4945 | | 184674 |
| Grasslands | 267639 | 234562 | 325539 | 265344 |
| Permanent wetlands | 18731 | 18284 | 10055 | |
| Croplands | 195069 | 161682 | 163214 | 7038 |
| Urban and built up | 10274 | 9059 | | 5182 |
| Cropland/natural vegetation mosaic | 7419 | 84276 | | |
| Barren or sparsely vegetated | 47482 | 46324 | 29255 | 48860 |
| Water bodies | 11739 | 10861 | 2945 | 9 |

Table 5: The confusion matrix for the vegetation classification from land use type to land covers scheme.

| Class | EN | DN | DB | MF | CS | OS | Classified total | Number correct | Accuracy | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | Producer's | User's |
| EN | 87920 | 452 | 121 | 389 | | | 9754 | 8792 | 90.14 | 90.86 |
| DN | 990 | 5346 | 327 | 213 | | | 5985 | 5346 | 89.32 | 84.62 |
| DB | 720 | 72 | 4783 | 412 | | | 5339 | 4783 | 89.59 | 87.70 |
| MF | 7130 | 448 | 223 | 4956 | | | 6340 | 4956 | 78.17 | 83.02 |
| CS | | | | | 3268 | 123 | 3391 | 3268 | 96.37 | 89.95 |
| OS | | | | | 365 | 4222 | 4587 | 4222 | 92.04 | 97.17 |
| **Ref. total** | **96760** | **6318** | **5454** | **5970** | **3633** | **4345** | **35396** | **31367** | | |
| Overall classification accuracy = 88.62% | | | | | | | | Overall Kappa statistics = 0.86 | | |

EN: evergreen needle leaved forest; DN: deciduous needle leaved forest; DB: deciduous broad leaved forest; MF: mixed forests; CS: closed shrublands; OS: open shrublands).

that can represent the "true" land cover. Gong performed the validation of a global land cover map using the ground-truth sample land cover data from the global flux site [38]. In this study, the accuracy of the input land use data was high and had been validated in 2000. So we only needed to validate the accuracy of the forest type and grassland type. The ground reference data, which came from multiple sources such as the field investigations, Terrestrial Ecosystem Monitoring Sites (TEMS), and 2 samples from the high-resolution images obtained via Google Earth, were used to validate the land cover products (Table 6). The results showed that the overall accuracy of the SCSLC map was 83.14%, which was much higher than that of the GLC land cover map (68%) and the UMD land cover map (52%).

In addition, the temporal characteristics are also very important to the validation of the information of the vegetation type. We compared the temporal NDVI value of the transformed land cover data to analyze the characteristics of different forest types. We evaluated the dataset according to phenological traits of vegetation which are closely related to the temperature as well as the elevation. The vegetation dynamics represents some important short-term and long-term ecological processes. The continuous temporal observations of the land surface parameters with the satellite can reveal their seasonal and annual development. In this study, we used vegetation indices of classified forests to characterize the state and dynamics of vegetation. In most cases, different types of vegetation have different phenological patterns. The NDVI value of the deciduous broadleaved forest is the highest in the four types of vegetation, and that of open shrublands is the lowest. The statistical curve from the classified land cover maps showed that the evergreen land cover had no remarkable change during the study period. However, the deciduous forest had a single peak in the sliding curve of NDVI in a year (Figure 7). This may be because the deciduous broadleaved forests were mainly located in the temperate zone, while the needleleaved forests were mainly in a cold-temperate zone or on mountains in a temperate zone.

## 5. Conclusions and Discussions

The information of the land cover is of great importance to the research of the global change science. The impacts of human activities such as the land cover change on regional and the global climate can be studied with climate modeling techniques. The land cover datasets, which are often derived from the remote sensing images, have been widely used to describe the physical surface conditions in the land surface

Table 6: The ground sample sites for validation over the farming-pastoral ecotone of North China.

| Longitude | Latitude | Station | Land cover types |
|---|---|---|---|
| 123.01°E | 51.78°N | Huzhong | Temperate coniferous forests |
| 121.56°E | 50.83°N | DaXinAnLing | Cold coniferous forests |
| 127.53°E | 45.38°N | MaoErShan | Temperate deciduous forest |
| 127.09°E | 42.40°N | ChangBaiShan | Temperate mixed forest |
| 119.94°E | 49.33°N | HuLunBeiEr | Temperate meadow steppe |
| 116.32°E | 44.13°N | XiLinGeLe | Temperate grassland |
| 117.45°E | 43.50°N | XiLinHaoTe | Leymus chinensis steppe |
| 115.99°E | 41.27°N | Google earth | Evergreen needle leaved forest |
| 115.61°E | 40.60°N | Google earth | Temperate mixed forest |
| 111.72°E | 40.61°N | ShaErQin | Grassland |
| 124.91°E | 41.82°N | QinYuan | Deciduous broad leaved forest |

schemes of climate models. But the accuracy of these datasets is still not high enough to meet the requirement of the climate simulation.

This paper has described the significance of the research on the use of data mining classification techniques for the land cover classification. The study significantly improved the vegetation classification accuracy of the land cover in North China by employing the data mining technique to the different satellite-derived land cover data of China, higher-precision land use data, and other ancillary spatial data. By computing the gain value of attributes for the vegetation classification, the results showed that the special monthly NDVI information is the most important, and temperature was more sensitive to the local land cover changes than precipitation. The method is used to classify the vegetation classes such as the closed forest, shrubland, and grassland with the exclusive spectral feature parameters.

The accuracy of the land cover classification is assessed by comparing the classification result with some reference data that is believed to have accurately reflected the true land cover. In this study, we found that the accuracy of the C4.5 classifier was 88.96%, which was higher than others, including NBTree, SimpleCart, REPTree, and BFTree. Besides, we calculated the confusion matrix and ROC value of the vegetation classification. The Kappa factor was 0.87, and the ROC value almost reached 0.90 on the whole, but the ROC value of the deciduous broadleaved forest was only 0.74. The validation all over China showed that the overall accuracy of the land cover map was 83.14%, which was higher than that of other land cover maps and met the requirement of the climate simulation for the accuracy over 80%. Therefore, the results have the potential to improve modeling accuracy for the land surface processes over China and can be used as the parameters of dynamical downscaling in the regional climate simulation.

In summary, the classifier developed in this study can be used to rapidly convert the high resolution CAS land use types into the land cover types for the climate simulation with the regional climate model. Besides, the time-series NDVI and NPP data retrieved from the remote sensing data can be used to rapidly produce the high resolution time-series vegetation data and realize the dynamic input of the parameters of the regional climate model, which can greatly improve the accuracy of the regional climate simulation. In addition, the results may provide support to other researches of the land surface science.

## Acknowledgments

This research was supported by the National Basic Research Program of China (973 Program) (no. 2010CB950904). Data supports from projects of the National Natural Science Foundation of China (no. 71225005) and the Exploratory Forefront Project for the Strategic Science Plan in IGSNRR, CAS are also appreciated.

## References

[1] K. Hibbard, A. Janetos, D. P. Van Vuuren et al., "Research priorities in land use and land-cover change for the Earth system and integrated assessment modelling," *International Journal of Climatology*, vol. 30, no. 13, pp. 2118–2128, 2010.

[2] P. H. Verburg, K. Neumann, and L. Nol, "Challenges in using land use and land cover data for global change studies," *Global Change Biology*, vol. 17, no. 2, pp. 974–989, 2011.

[3] E. Sertel, A. Robock, and C. Ormeci, "Impacts of land cover data quality on regional climate simulations," *International Journal of Climatology*, vol. 30, no. 13, pp. 1942–1953, 2010.

[4] J. Jin, S. Lu L, and L. Norman Miller, "Impact of land use change on the local climate over the Tibetan plateau," *Advances in Meteorology*, vol. 2010, Article ID 837480, 6 pages, 2010.

[5] G. B. Bonan, S. Levis, S. Sitch, M. Vertenstein, and K. W. Oleson, "A dynamic global vegetation model for use with climate models: concepts and description of simulated vegetation dynamics," *Global Change Biology*, vol. 9, no. 11, pp. 1543–1566, 2003.

[6] J. J. Feddema, K. W. Oleson, G. B. Bonan et al., "Atmospheric science: the importance of land-cover change in simulating future climates," *Science*, vol. 310, no. 5754, pp. 1674–1678, 2005.

[7] J. Ge, J. Qi, B. M. Lofgren et al., "Impacts of land use/cover classification accuracy on regional climate simulations," *Journal of Geophysical Research*, vol. 112, Article ID D05107, 2007.

[8] Z. Gao, J. Liu, and D. Zhuang, "The research of Chinese land-use/land-cover present situations," *Journal of Remote Sensing*, vol. 3, no. 2, pp. 134–138, 1999 (Chinese).

[9] J. Liu, M. Liu, D. Zhuang, Z. Zhang, and X. Deng, "The spatial pattern analysis of land use change of China," *Science in China D*, vol. 32, no. 13, pp. 1031–1040, 2002 (Chinese).

[10] J. Liu, M. Liu, D. Zhuang, Z. Zhang, and X. Deng, "Study on spatial pattern of land-use change in China during 1995–2000," *Science in China D*, vol. 46, no. 4, pp. 373–384, 2003.

[11] A. T. K. Tchuenté, J. Roujean, and S. M. de Jong, "Comparison and relative quality assessment of the GLC2000, GLOBCOVER, MODIS and ECOCLIMAP land cover data sets at the African continental scale," *International Journal of Applied Earth Observation and Geoinformation*, vol. 13, no. 2, pp. 207–219, 2011.

[12] I. McCallum, M. Obersteiner, S. Nilsson, and A. Shvidenko, "A spatial comparison of four satellite derived 1 km global land cover datasets," *International Journal of Applied Earth Observation and Geoinformation*, vol. 8, no. 4, pp. 246–255, 2006.

[13] M. Herold, R. Hubald, and A. D. Gregorio, "Translation and evaluating land cover legends using the UN land cover classification system (LCCS)," GOGC-GOLD Report 43, GOGC-GOLD, Florence, Italy, 2009.

[14] J. Dong, D. Zhuang, Y. Huang, and J. Fu, "Advances in multi-sensor data fusion: algorithms and applications," *Sensors*, vol. 9, no. 10, pp. 7771–7784, 2009.

[15] P. J. Sellers, D. A. Randall, G. J. Collatz et al., "A revised land surface parameterization (SiB2) for atmospheric GCMs. Part I: model formulation," *Journal of Climate*, vol. 9, no. 4, pp. 676–705, 1996.

[16] H. Gao and G. Jia, "Spatial and quantitative comparison of satellite-derived land cover products over China," *Atmospheric and Oceanic Science Letters*, vol. 5, no. 5, pp. 426–434, 2012.

[17] P. Mahesh, "Ensemble learning with decision tree for remote sensing classification," *World Academy of Science, Engineering and Technology*, vol. 26, pp. 258–260, 2007.

[18] M. R. Rahman and S. K. Saha, "Multi-resolution segmentation for object-based classification and accuracy assessment of land use/land cover classification using remotely sensed data," *Journal of the Indian Society of Remote Sensing*, vol. 36, no. 2, pp. 189–201, 2008.

[19] M. Simard, S. S. Saatchi, and G. De Grandi, "The use of decision tree and multiscale texture for classification of JERS-1 SAR data over tropical forest," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 5, pp. 2310–2321, 2000.

[20] C. Liu, P. Frazier, and L. Kumar, "Comparative assessment of the measures of thematic classification accuracy," *Remote Sensing of Environment*, vol. 107, no. 4, pp. 606–616, 2007.

[21] A. M. Niccolai, A. Hohl, M. Niccolai, and D. O. Chadwick, "Decision rule-based approach to automatic tree crown detection and size classification," *International Journal of Remote Sensing*, vol. 31, no. 12, pp. 3089–3123, 2010.

[22] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: a comparison of regression analysis, decision tree and neural networks," *Energy*, vol. 32, no. 9, pp. 1761–1768, 2007.

[23] M. K. Ghose, R. Pradhan, and S. S. Ghose, "Decision tree classification of remotely sensed satellite data using spectral separability matrix," *International Journal of Advanced Computer Science and Applications*, vol. 1, no. 5, pp. 93–101, 2010.

[24] M. Kamel, "Ecotone classification according to its origin," *Pakistan Journal of Biological Sciences*, vol. 6, no. 17, pp. 1553–1563, 2003.

[25] J. Liu, J. Gao, S. Lv, Y. Han, and Y. Nie, "Shifting farming-pastoral ecotone in China under climate and land use changes," *Journal of Arid Environments*, vol. 75, no. 3, pp. 298–308, 2011 (Chinese).

[26] H. Zhao, X. Zhao, T. Zhang, and R. Zhou, "Boundary line on agro-pasture zigzag zone in North China and its problems on eco-environment," *Advance in Earth Sciences*, vol. 17, no. 5, pp. 739–747, 2002 (Chinese).

[27] J. Liu, J. Gao, B. Geng, and L. Wu, "Study on the dynamic change of land use and landscape pattern in the farming- pastoral region of Northern China," *Research of Environmental Sciences*, vol. 20, no. 5, pp. 148–154, 2007 (Chinese).

[28] J. Dong and X. Xu, "Land use change especially alternations of farming and grazing in typical agro-pastoral transitional zone in 1988–2000: a case study in Chifeng City of Inner Mongolia," *Journal of Geo-Information Science*, vol. 11, pp. 413–420, 2009.

[29] J. Fan, Y. Zhang, and G. Li, "Climate change in the middle of farming-grazing zone of Northern China," *Advance in Climate Change Research*, vol. 3, no. 2, pp. 91–94, 2007 (Chinese).

[30] M. A. Friedl, D. K. McIver, J. C. F. Hodges et al., "Global land cover mapping from MODIS: algorithms and early results," *Remote Sensing of Environment*, vol. 83, no. 1-2, pp. 287–302, 2002.

[31] E. Bartholomé and A. S. Belward, "GLC2000: a new approach to global land cover mapping from earth observation data," *International Journal of Remote Sensing*, vol. 26, no. 9, pp. 1959–1977, 2005.

[32] Y. Ran, X. Li, L. Lu, and Z. Y. Li, "Large-scale land cover mapping with the integration of multi-source information based on the Dempster-Shafer theory," *International Journal of Geographical Information Science*, vol. 26, no. 1, pp. 169–191, 2012.

[33] J. Liu, M. Liu, H. Tian et al., "Spatial and temporal patterns of China's cropland during 1990–2000: an analysis based on Landsat TM data," *Remote Sensing of Environment*, vol. 98, no. 4, pp. 442–456, 2005.

[34] X. Hou, *Vegetation Map (1:1, 000, 000) in China*, Science Press, Beijing, China, 2001.

[35] R. S. Defries and J. R. G. Townshend, "NDVI-derived land cover classifications at a global scale," *International Journal of Remote Sensing*, vol. 15, no. 17, pp. 3567–3586, 1994.

[36] J. Chen, P. Jönsson, M. Tamura, Z. Gu, B. Matsushita, and L. Eklundh, "A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky-Golay filter," *Remote Sensing of Environment*, vol. 91, no. 3-4, pp. 332–344, 2004.

[37] O. Ahlqvist, "Extending post-classification change detection using semantic similarity metrics to overcome class heterogeneity: a study of 1992 and 2001 U.S. National Land Cover Database changes," *Remote Sensing of Environment*, vol. 112, no. 3, pp. 1226–1241, 2008.

[38] P. Gong, "The accuracy evaluation for global land cover map based on the global flux site," *Progress in Natural Science*, vol. 19, pp. 754–759, 2009 (Chinese).