Integrating Landsat TM Imagery and See5 Decision-Tree Software for Identifying Croplands: A Case Study in Shunyi District, Beijing

Jinling Zhao[1, 2], Dongyan Zhang[1], Dacheng Wang[1], and Wenjiang Huang[1,*]

[1] Beijing Research Center for Information Technology in Agriculture, Beijing Academy of Agriculture and Forestry Sciences, Beijing 100097, P.R. China
zhaojl@nercita.org.cn
yellowstar0618@163.com
[2] Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing 100101, P.R. China

Abstract. As an important natural resource, cropland plays a key role in ensuring food safety. In this study, an integral method combining Landsat TM imagery and See5 decision-tree software was developed to identify croplands by taking Shunyi District, Beijing as the study area. Considering the specific topographic conditions, vegetation types and variable climate environment as well as growth period of the study area, texture variables, band ratios, digital elevation model (DEM) and its derived slope and aspect were added into decision tree classification. Finally, the cropland distribution map of Shunyi District was derived combining See5 decision tree classification software and NLCD mapping tools integrated in the ERDAS environment. An accuracy evaluation shows that the overall accuracy is 88.48% and 91.85% using GPS sample points and statistical data, separately. The result shows that it is feasible to identify croplands using See5 decision-tree classification tool based on the Landsat TM imagery.

Keywords: Band ratios, cropland identification, Landsat TM, See5 decision-tree classification, texture analysis.

1 Introduction

As an important agricultural resource to ensure food security, it is very essential to derive the real and accurate utilization information and spatial distribution of croplands to a great degree. Base on such information, farmers can arrange their crop planting and decision-makers in the departments of agriculture can make an optimum planning for various kinds of crop planting [1]. With the increasing demands, it is urgent to find out an effective way for identifying and evaluating the croplands especially on a large scale. However, traditional labor-intensive and high-cost consuming methods were mainly used in the past. In order to detect the planting conditions of croplands, surveyors must physically survey the information, which is

* Corresponding author.

Z. Gong et al. (Eds.): WISM 2011, Part I, LNCS 6987, pp. 251–258, 2011.

time consuming and subject to considerable errors. Therefore, the mapping of cropland information is very essential for knowing the planting area, spatial distribution, land cover and land use, field management, etc.

As a result, to derive related information of croplands is very necessary. With the increasing advances of earth-observing satellite technology, remote sensing technology has been widely used to monitor growth conditions, yield estimation, drought/cold damage, etc [2-4]. However, in previous studies moderate spatial resolution imagery such as MODIS, AVHRR, ERTS-1were primarily used to identify and extract cropland information [5-7]. However, due to the influence of subpixel heterogeneity and mixed pixel for coarse resolution sensors and narrow swath, such as MODIS, AVHRR, they may result in significant errors in cropland area estimation. Therefore, those remotely sensed images with higher resolution have to be utilized in order to derive more accurate cropland information.

The primary objective of this study was to evaluate the potential of the Landsat TM imagery to identify cropland information by integrating See5 decision-tree software in Guangxi Province, China. In addition, we also want to evaluate the classification efficiency and accuracy of See5 classifier by field survey data and statistical data.

2 Description of the Study Area

Shunyi District is located in the north-east suburbs of Beijing, about 30 kilometers from the centre, at latitude 40°00′-40°18′ North and longitude 116°28′-116°58′ East. The total area is 1,021 km^2 and it has a population of 593,000, of which 419,000 are permanent agricultural residents (Beijing Statistical Office, 2001). Shunyi has a warm temperate wet continental monsoon climate. Average annual temperature is 11.5 °C, that in January 4.9 °C, and in July 25.7 °C. The lowest temperature in January is -19.1 °C and the highest in July 40.5 °C. The frost-free period lasts around 195 days. Annual sunshine duration is 2,750 hours, average annual relative humidity about 50%. Average annual precipitation is about 625 mm, of which 75% falls in summer. It has a fertile soil ranging from sandy to loamy soils. Fig. 1 shows the spatial location and administration divisions of Shunyi District.

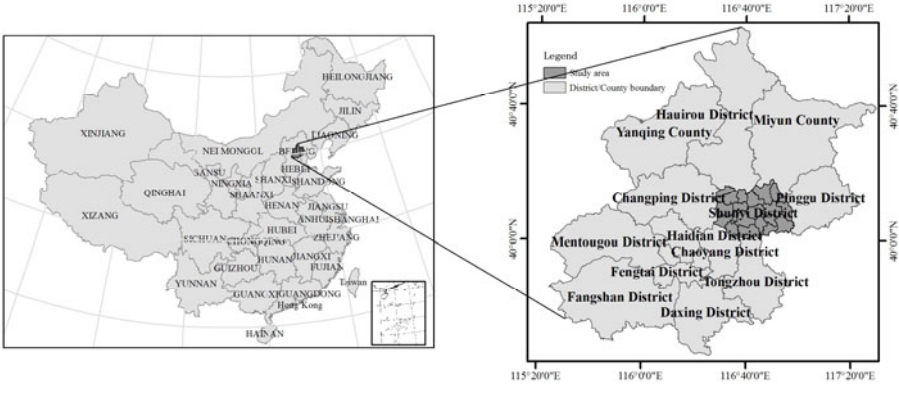

Fig. 1. Spatial location and administration divisions of Shunyi District, Beijing

3 Methodology

3.1 Data Sources and Preprocessing

We acquired Landsat TM imagery of the study which retain as much of the original radiometric and geometric properties as possible and they were just systematically processed by radiometric and geometric corrections. Therefore, accurate radiometric and geometric corrections must be firstly performed. In addition, topographic relief exerts a great influence on identifying croplands and topographic correction must be also performed. The commercial ERDAS Imagine software was used to co-register and orthorectify the Landsat TM images. The orthocorrection model of Leica Photogrammetry Suite (LPS) was used by combining with ASTER 30 m resolution digital elevation model (DEM). As a result, the root mean squared errors (RMSE) for all images were less than 0.5 pixels, and the nearest neighbor method was used for image resampling.

3.2 Classification Method

For indentifying cropland information, the machine learning Windows XP software package See5 was used to generate classification trees from the Landsat TM imagery, ancillary data sets, and field measurements [8]. The reason we chose this algorithm is that it is freely available for time limited testing. In addition, it is also an improved version of the most popular classification algorithm C4.5 and has a cross-validation function built-in. In order to complete the croplands identification, the National Land Cover Dataset (NLCD) Mapping Tool must be jointly utilized, which was designed by MDA Federal, Inc., for the United States Geological Survey. The tools developed for use within the ERDAS Imagine 8.7 software environment, and are for use with the Rulequest Research Cubist and See5 software packages above version 1.12 of cubist, and above version 1.18 of See5.

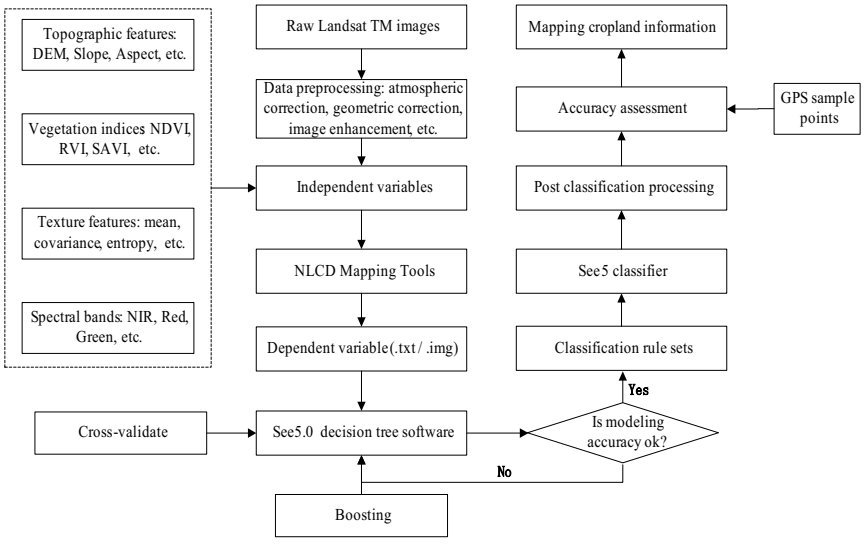

Fig. 2. Overall flow chart for identifying croplands integrating NLCD mapping tools and See5 decision-tree software

Based on the original Landsat TM image and DEM data, a variety of independent variables were derived. When they were input into the NLCD tool, the dependent variables required by See5 software was obtained. In the hierarchical tree structure, each split in the tree results in two branches (Fig. 2). The algorithm searches for the dependent variable that, if used to split the population of pixels into two groups, explains the largest proportion of deviation of the independent variable. At each new split in the tree, the same exercise is conducted and the tree is grown until it reaches terminal nodes, or leaves, each leaf representing a unique set of pixels. Every leaf has a land cover class assignment.

3.3 See5 Decision-Tree Software

Data mining software of See 5 with boosting technique can build decision tree quickly and improve the precision of miscible classes. See5 (Windows 2000/Xp/Vista/7) is sophisticated data mining tools for discovering patterns that delineate categories, assembling them into classifiers, and using them to make predictions. See5 data mining software is based on the C5.0 algorithm. Fig. 3 is the user interface of See5 and NLCD integrated in the ERDAS image processing environment. Independent variables including band ratios and texture were input See5 software. Then, it constructs a decision tree with the default values of all options. Classifiers constructed by See5 are evaluated on the training data from which they were generated. Finally, a decision tree is constructed and the accuracy is evaluated by cross-validation to select the available variables which have more contribution to classification.

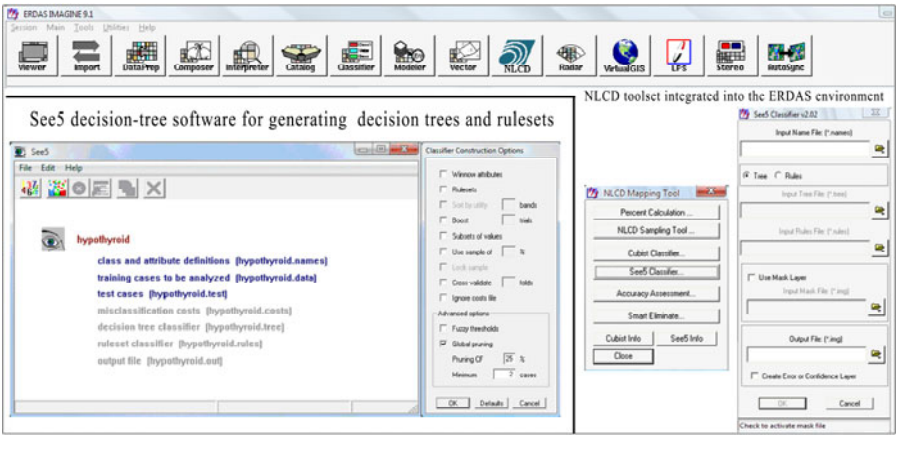

Fig. 3. The user interface of See5 and NLCD integrated into the ERDAS image processing environment

3.4 Derivation of Independent Variables

Compared with traditional supervision and unsupervised classification techniques, decision trees are strictly nonparametric and do not require assumptions regarding the statistical properties of the input data. They combine spectral tone, spatial texture, topographic data and line and sample location. In this study, four kinds of

Table 1. Derived independent variables for constructing decision tree rule sets

Vegetation Indices/Band Ratios	Equation [a]	Reference
Normalized Difference Vegetation Index (NDVI)	$NDVI=(R_{NIR}-R_{RED})/(R_{NIR}+R_{RED})$	[9]
Ratio Normalized Difference Vegetation Index (RNDVI)	$RNDVI=(R_{NIR}-R_{RED})/(R_{NIR}+R_{RED})*(R_{NIR}/R_{RED})$	[10]
Green normalized difference vegetation index (GNDVI)	$GNDVI= (R_{NIR}-R_{Green})/(R_{NIR}+R_{Green})$	[11]
Transformed NDVI (TNDVI)	$TNDVI = [(NIR-R)/(NIR+R)+1]^{1/2}$	[12]
Soil Adjusted Vegetation Index (SAVI)	$SAVI=[(1+L)\times(R_{NIR}-R_{RED})]/(R_{NIR}+R_{RED}+L)$	[13]
Difference Vegetation Index (DVI)	$DVI= R_{NIR}-R_{RED}$	[14]
Ratio Vegetation Index (RVI)	$RVI= R_{NIR}/R_{RED}$	[15]
Infrared percentage vegetation index (IPVI)	$IPVI= R_{NIR}/(R_{NIR}+R_{RED})$	[16]
Transformed vegetation index (TVI)	$TVI=(NDVI+0.5)^{1/2}$	[12]
Renormalized difference vegetation index (RDVI)	$RDVI= (NDVI\times DVI)^{1/2}$	[14]

[a] where NIR denotes crop reflectance in the near infrared band (Landsat TM-Band4), RED denotes crop reflectance in the red band (Landsat TM-Band3), L = constant (taken as 0.5).

Table 2. Derived texture variables of GLCM for generating decision tree rule sets

GLCM variable	Formulas[b]	Description
GLCM Mean	$Mean=\sum_{i,j=0}^{N-1} i\left(p_{i,j}\right)$	It calculates the mean based on the reference pixels i.
Entropy (ENT)	$ENT=\sum_{i,j=0}^{N-1} p_{ij}\left(-\ln p_{ij}\right)$	Entropy is usually classified as a first degree measure, but should properly be a "zeroth" degree!
Contrast (CON)	$CON=\sum_{i,j=0}^{N-1} p_{ij}(i-j)^2$	1) If i and j differ by 1, there is a small contrast, and the weight is 1. 2) If i and j differ by 2, contrast is increasing and the weight is 4. The weights continue to increase exponentially as (i-j) increases.
COR (Correlation)	$COR=\sum_{i,j=0}^{N-1} p_{ij}\left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}}\right]$	The Correlation texture measures the linear dependency of grey levels on those of neighbouring pixels.
HOM (Homogeneity)	$HOM=\sum_{i,j=0}^{N-1} \frac{P_{i,j}}{1+(i-j)^2}$	Homogeneity weights values by the inverse of the Contrast weight, with weights decreasing exponentially away from the diagonal.

[b] where N is the number of levels specified under quantization, P_{ij} value is the probability value from the GLCM.

independent variables were derived: topographic data, texture features, vegetation indices and raw spectral bands. In accordance with the available spectral bands, vegetation indices as well as topographic and texture independent variables were calculated as shown in Table 1 and Table 2. The Gray Level Co-occurrence Matrix (GLCM) and associated texture feature calculations are image analysis techniques. Given an image composed of pixels each with an intensity (a specific gray level), the GLCM is a tabulation of how often different combinations of gray levels co-occur in an image or image section. Texture feature calculations use the contents of the GLCM to give a measure of the variation in intensity (a.k.a. image texture) at the pixel of interest. Texture datasets are from grave-level co-occurrence matrix (GLCM): Mean, Entropy, Contrast, Correlation and Homogeneity.

4 Results

4.1 Identified Croplands

In the study area, land under field crops, vegetables and flowers which account for most of the agricultural land were assigned to cropland, and all other land cover types were classified as non-cropland. It must be mentioned that perennial tree crops (including fruit trees) which present very little percentage of the agricultural land were classified into non-crop classes because they have similar spectral reflectance features with the forest. Therefore, it is very difficult to partition the tree crops from forest without the support of other ancillary data.

An exclusion method was applied to the Landsat TM classification data to produce the crop/non-crop map. At the initial stage, non-vegetation area including water, construction land, and bare land was derived. Then, the forest map was produced. After extracting the forest and no-vegetation land, the rest land was treated as potential crop area. Fig. 4 shows the original Landsat TM image and cropland identification result. As shown in this figure, the cropland distributes almost everywhere in the study area besides the Southwestern part and Northeastern corner because of built-up areas and forest.

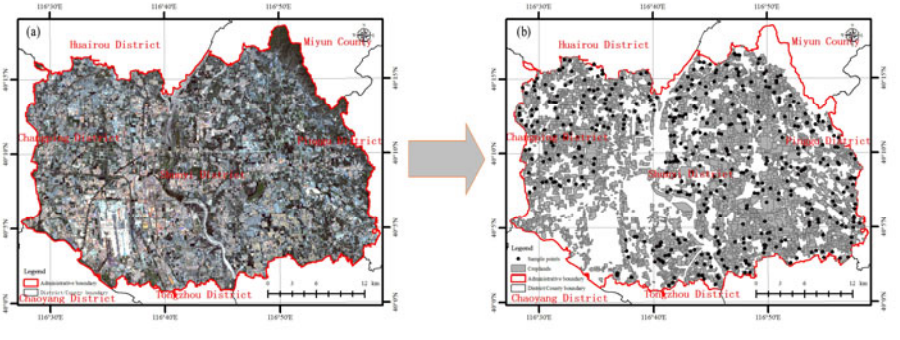

Fig. 4. (a) is the original Landsat TM false color composite image by Band 1, Band 2 and Band3, and (b) is the identified croplands in Shunyi District

4.2 Accuracy Assessment

The classification result based on remote-sensing images must be evaluated by other ancillary dataset such as field survey data and statistical data. Here, 495 sample points was used to validate the cropland classification result. The result shows that 454 points are classified as cropland and the overall accuracy is 88.48% (438/495*100%). This means that the accuracy of cropland thematic map could satisfy the minimum identification requirements and are acceptable [17]. In addition, the total acreage of cropland was also used to evaluate the classification. The Landsat TM-based cropland acreage is 42,864.56 ha compared with 46,666.67 ha of the statistical data from Shunyi Branch of Beijing Municipal Bureau of State Land and Resources, and the overall accuracy is 91.85%.

5 Conclusion

Compared with traditional labor-intensive manner, remote sensing investigations can get accurate and real-time information especially in the large-scale regions. Even though the Landsat TM imagery cannot derive accurate cropland acreages because of its coarse spatial resolution, the spatial distribution trend of croplands is relatively useful to mater the cropland layout, cropping system and planting acreage, etc. for farmers and decision-makers. In addition, the accuracy can be improved with the help of advanced identification methods and various kinds of ancillary variables such as texture, color, DEM data. In the near future, with the launch of increasing sensors, the spatial, spectral and temporal resolution of remotely sensed images will be further improved and the cropland information can be accurately identified.

Acknowledgements. This work was financially supported by the Postdoctoral Science Foundation of Beijing Academy of Agriculture and Forestry Sciences, the National Natural Science Foundation of China (41071276), the Program of Ministry of Agriculture (200903010), and the Special Funds for Major State Basic Research Project (2007CB714406).

References

1. Hao, Y., Lal, R., Owens, L.B., Izaurralde, R.C., Post, W.M., Hothem, D.L.: Effect of Cropland Management and Slope Position on Soil Organic Carbon Pool at the North Appalachian Experimental Watersheds. Soil Till. Res. 68, 133–142 (2002)
2. Abdel-Rahman, E.M., Ahmed, F.B.: The Application of Remote Sensing Techniques to Sugarcane (Saccharum spp. hybrid) Production: A Review of the Literature. Int. J. Remote Sens. 23, 3753–3767 (2008)
3. Rao, P.V.K., Rao, V.V., Venkataratnam, L.: Remote Sensing: A Technology for Assessment of Sugarcane Crop Acreage and Yield. Sugar Tech. 4, 97–101 (2002)
4. Rudorff, B.F.T., Batista, G.T.: Yield Estimation of Sugarcane based on Agrometeorological Spectral Models. Remote Sens. Environ. 33, 183–192 (1990)
5. Carlson, R.E., Aspiazu, C.: Cropland Acreage Estimates from Temporal, Multispectral ERTS-1 Data. Remote Sens. Environ. 4, 237–243 (1975-1976)

6. Xiao, X.M., Liu, J.Y., Zhuang, D.F., Frolking, S., Boles, S., Xu, B., Liu, M.L., Salas, W., Moore III, B., Li, C.S.: Uncertainties in Estimates of Cropland Area in China: a Comparison between an AVHRR-derived Dataset and a Landsat TM-derived Dataset. Global Planet. Change 37, 297–306 (2003)
7. Lobell, D.B., Asner, G.P.: Cropland Distributions from Temporal Unmixing of MODIS Data. Remote Sens. Environ. 93, 412–422 (2004)
8. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1993)
9. Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W.: Monitoring Vegetation Systems in the Great Plains with ERTS. In: Proceedings of the Third Earth Resources Technology Satellite-1 Symposium, pp. 301–317. NASA, Greenbelt (1974)
10. Gong, P., Pu, R., Biging, G.S., Larrieu, M.R.: Estimation of Forest Leaf Area Index using Vegetation Indices Derived from Hyperion Hyperspectral Data. IEEE Trans. Geosci. and Remote Sens. 41, 1355–1362 (2003)
11. Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N.: Use of a Green Channel in Remote Sensing of Global Vegetation from EOS-MODIS. Remote Sens. Environ. 58, 289–298 (1996)
12. Tucker, C.J.: Red and Photographic Infrared linear Combinations for Monitoring Vegetation. Remote Sens. Environ. 8, 127–150 (1979)
13. Huete, A.R.: A Soil Adjusted Vegetation Index (SAVI). Remote Sens. Environ. 25, 295–309 (1988)
14. Roujean, J.L., Breon, F.M.: Estimating PAR Absorbed by Vegetation from Bidirectional Reflectance Measurements. Remote Sens. Environ. 51, 375–384 (1995)
15. Jordan, C.F.: Derivation of Leaf Area Index from Quality Measurements of Light on the Forest Floor. Ecology 50, 663–666 (1969)
16. Crippen, R.E.: Calculating the Vegetation Index Faster. Remote Sens. Environ. 34, 71–73 (1990)
17. Lucas, I.F.J., Frans, J.M., Wel, V.D.: Accuracy Assessment of Satellite Derived Land-cover Data: A review. Photogramm. Eng. Rem. Sens. 60, 410–432 (1994)