# Estimating impervious surface percent in plain river network regions using a refinement CART analysis

SHE Yuanjian
School of Earth Sciences and Engineering
Hohai University
Nanjing, China
sheyj@hhu.edu.cn

LI Xiaoning
School of Earth Sciences and Engineering
Hohai University
Nanjing, China
lixiaoning158@163.com

***Abstract*—The rapid expansion of impervious surface has become a major factor affecting ecosystem health of the high density river network. In this paper, the ensemble learning of CART analysis was used to estimate impervious surface percent(ISP) through Variable Precision Rough Sets (VPRS). First, Landsat TM and ALOS imagery were utilized to construct the ISP predictive model; then, in order to get the best attribute variables of CART decision tree, VPRS was adopted to extract optimum feature subset from multi-source feature sets. Results illustrate the validity of this ensemble learning, and prove that this method can obtain higher accuracy than the traditional single CART method. However, in the initial estimation results, ISP's high value area was underestimated relatively seriously. It was found that there is an intensive relationship between the Temperature Vegetation Dryness Index (TVDI) and ISP. The increase of ISP will cause significant increase of local TVDI. Then post-processing rules extracted from the relationship was used to improve results. According to the verified results, the combination of VPRS reduction and post-processing rule in CART algorithm has higher analysis precision than the traditional single CART learning algorithm. The root mean square error between estimated ISP value and reference ISP is 10.0% and the correlation coefficient is 0.89. The method is viable for the estimation of the ISP in plain river network regions.**

***Index Terms*—Impervious Surface Percent (ISP), CART, Variable Precision Rough Sets (VPRS), Temperature Vegetation Dryness Index (TVDI ), Plain river network region**

## I. INTRODUCTION

Impervious surfaces are manmade features through which water cannot infiltrate into the soil. Impervious surface are the features such as rooftops, roads, driveways, sidewalks, and parking lots and others. Impervious surface percent is the percent of impervious surface in an area as large as a drainage basin or as small as an area[1]. Previous researches show that the increase of impervious surface would increase the flood occurrence frequency, cause the disruption of the habitat of aquatic plants, and the decrease of the supply of groundwater. Furthermore, a large amount of pollutants from impervious surface goes into rivers along with the surface runoff, polluting the rivers and reducing the health water supplies, which has become a major factor affecting ecosystem health of the high density river network.

Among estimating ISP models, CART analysis, which has all the advantages that decision trees analysis algorithms have, has become a hot topic of the research. Nevertheless, it is a kind of weak learning algorithm and the classification accuracy depends upon if the construction and pruning of the decision tree is reasonable. Although Nathaniel[2] and Ma Xuemei[3] made a comprehensive use of multi-source features, they did not optimized the feature set; Liao Mingsheng[4] introduced the technology of Boosting in the CART analysis, but he just used the spectral features of the remote sensing images. When the training samples have much noise, the accuracy of the above two methods will greatly decrease. YANG[5] sets up many evaluation models by combining multi-features. Then the model with the least error and the highest correlation index will be selected. This method eliminates data redundancy to some degree, but it needs more expert knowledge.

A Variable Precision Rough Set (VPRS)[6] model approach has been widely used in the area of artificial intelligence, especially in data mining and knowledge discovery, is applied for attributes reduction. Then the Temperature Vegetation Dryness Index(TVDI)[7] is used to get post-processing rule for the ISP initial result.

## II. STUDY AREA AND DATA

Study area as shown in Fig. 1, is an area of typical plain water systems, where rivers crisscross the region, scattered with lakes. With the development of urbanization, the urban area expands to the suburbs constantly. Original natural stream systems were destroyed to varying degrees by human activities, breaking the regulation capacity of original natural water system.

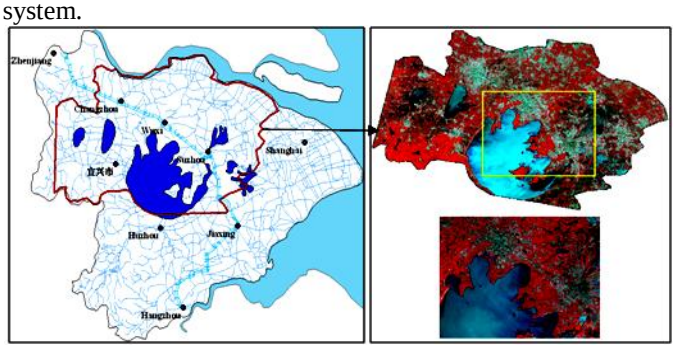



Fig. 1. Map of study area

In this research, a Landsat TM image acquired on May 24, 2010 under clear weather conditions was used. In addition, a part of ALOS image was selected for ISP training and validation purpose. The ALOS data was a pan-sharpened image by fusing the multispectral 10m resolution image and 2.5m resolution image pan. The resulting data training and validation data is a 3m resolution image. Because there is no atmospheric effect on the ISP's estimate, the DN values were only transformed into at-sensor radiance values.

## III. RESEARCH METHOD

### *A. Training data and validation data development*

The fused ALOS image was classified into impervious surface and pervious surface. Impervious surface includes building and road. Pervious includes farmland, vegetation and water. The percentage of impervious surface was calculated within a 10*10 window based on the classification results. Then the percentage image was resampled to a 30 m resolution image (Fig. 2), cloud and water was excluded. For sample selection, 3600 center points of the 60 rows * 60 columns of regular square grid was generated from the 30m resolution image (Fig. 3). Then the sample points were checked and the points with large errors were modified. Finally, the samples were divided into two independent subsets, 2600 samples for ISP estimation and 1000 samples for validation.

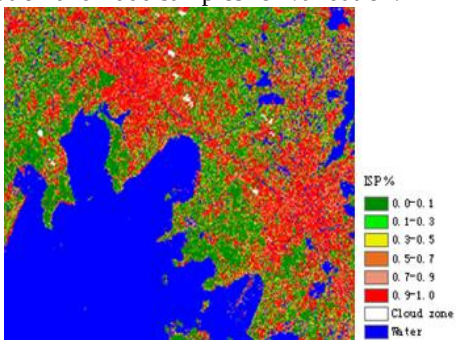



Fig. 2. ISP classification result of sample area

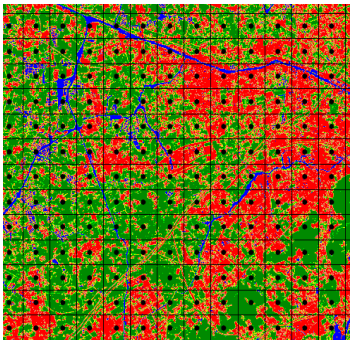


Fig. 3. Regular square grid of sample area

### *B. Attributes for CART model*

According to vegetation–impervious surface–soil (V-I-S) model, each cell is regarded as the linear combination of impervious surface, vegetation and water. Dozens of different index models about these three land cover types have been researched by scholars and experts. The most widely used index models are chose in this paper. Normalized Difference Built-up Index (NDBI), Urban Index (UI) and Index based Built-up Index(IBI) were used for impervious; Normalized Difference Water Index(NDWI), Modified Normalized Difference Water Index (MNDWI) and Combined Index of Water Index(CIWI) were used for water; Ratio Vegetation Index (RVI), Normalized Difference Vegetation Index (NDVI) and Soil Adjusted Vegetation Index (SAVI) were selected for vegetation.

Texture is also the important information for extraction of object information. In this paper, eight texture metrics of GLCM of the first principal component of TM image were used, including Mean, Variance, Homogeneity, Contrast, Dissimilarity, Entropy, Second Moment and Correlation. The size of window is defined as 3*3, moving step length is one pixel and moving direction is zero degree.

In addition, the first four bands of MNF (Minimum Noise Fraction Rotation) rotation were chosen, and the first three components of K-T transformation were chosen: brightness (KT-b), greenness (KT-g) and wetness (KT-w) .

Totally, 31 attributes were used for the CART model (see TABLE I).

TABLE I. ATTRIBUTE VARIABLE OF CART MODEL ANALYSIS

| code | type | code | type |
|---|---|---|---|
| B1 | TM1 | B17 | MNF1 |
| B2 | TM2 | B18 | MNF2 |
| B3 | TM3 | B19 | MNF3 |
| B4 | TM4 | B20 | MNF4 |
| B5 | TM5 | B21 | KT-b |
| B6 | TM7 | B22 | KT-g |
| B7 | RVI | B23 | KT-w |
| B8 | PVI | B24 | Mean |
| B9 | NDVI | B25 | Variance |
| B10 | SAVI | B26 | Homogeneity |
| B11 | NDWI | B27 | Contrast |
| B12 | MNDWI | B28 | Dissimilarity |
| B13 | CIWI | B29 | Entropy |
| B14 | NDBI | B30 | Second Moment |
| B15 | UI | B31 | Correlation |
| B16 | IBI | | |

### *C. Attribute variables reduction*

The redundancy of feature set can lead to the complexity of the decision tree and the situation in which the useful rules be hidden. How to reduce the dimension of feature set is a hot topic in remote sensing community. The VPRS proposed by Ziarko[8] et al. is a powerful tool to reduce the redundancy of data. Compared with the Pawalak's rough set model, it loosens the strict constraint of approximate boundary: $0.5<\beta\leq 1$. Approximate boundary narrows as the threshold parameter β increases. Therefore it decreases the size of uncertain region in rough set and has a tolerance to the inconsistency of data, and the value of tolerance is decided by the threshold β. The dimension of variables is reduced by the measure of classification quality of β[9]. The specific process is as follows:

1. The remote sensing image was viewed as an information system, 2600 training samples as objects, the 31 features

as condition attributes, and the ISP of training data as decision attributes. Thus an information decision table was created;

2. Naive Scalar[10]algorithm embedded in Rosetta software was used to discrete the condition attribute;
3. Calculate the equivalent sets of condition attributes and decision attributes and lower approximation and upper approximation of threshold parameter β
4. Attribute reduction of decision table: initialize the root node R and the queue Q, take out the first note N of the queue; if N is not pure, then estimate the β measure of classification quality of each note, select the attribute with the biggest measure, split N into $N_{[1]}$, $N_{[2]}$,$\cdots N_{|N|;}$
5. Input the $N_{[1]}$, $N_{[2]}$,$\cdots N_{|N|}$ to the queue, then turn to step 4.
6. Check the decision table consistency: if inconsistent phenomenon does not appear in the decision table when one attribute is deleted, then this attribute can be removed from the decision table, otherwise it should be reserved. Repeat the steps 4 and 5, until no redundant attributes exist.

TABLE II. β REDUCTION INFORMATION

| β region | Attributes reduction set | Classification quality | Decision tree notes | $R^2$ |
|---|---|---|---|---|
| [0.5,0.667] | {B1,B7,B12,14, B15, B16,B18, B19,B22, B24,B25,B27,B31} | 0.84 | 41 | 0.89 |
| [0.5,0.75] | {B1,B4,B7,B11, B12, B14,B17,B19,B20, B21,B22,B24,B25, B28,B30,B31} | 0.64 | 65 | 0.68 |

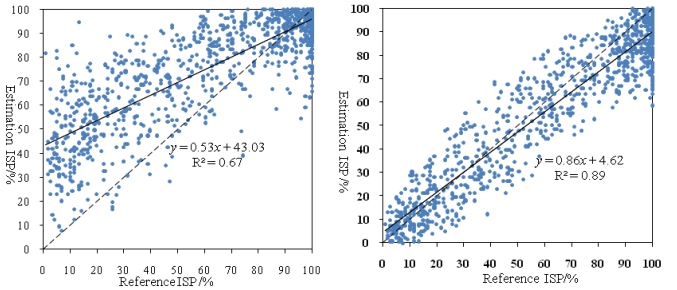
When β is 0.667[11][12](see TABLE II), the number of attributes was reduced from 31 to 13, and decision tree had only 41 nodes, and the sample classification quality and ISP estimation accuracy were higher. When β was 0.75, classification quality and estimation accuracy decreased, and the nodes of decision tree increased. Without reduction by VPRS, the number of nodes of decision tree built by 31 attributes was 426. In a word, when β is 0.667, the decision tree built by 13 attributes was simple. This greatly reduced the complexity of decision tree and improved precision. So the feature set determined by the β threshold was the optimal set.

## IV. RESULTS ANALYSIS

In the paper, 2600 ISP training samples were taken from the ALOS image as the target variable of the CART model, and 13 attribute features chosen by VPRS as attribute variables for the estimation of ISP. In order to verify the validity of attribute reduction of variable precision rough set, the same validation samples were used for the estimation of ISP by using the single CART algorithm and 31 features.

TABLE III. PERFORMANCE ASSESSMENT OF ISP ESTIMATION

| Method | MAE(%) | RMSE(%) | $R^2$ |
|---|---|---|---|
| Single CART algorithm | 18.9 | 25.1 | 0.67 |
| Ensembling CART algorithm | 9.5 | 11.9 | 0.89 |




a. Single CART algorithm b. Ensembling CART algorithm

Fig. 4. Scatter plot of estimating and reference ISP

1000 samples were used for comparison of the validity of the results of the two methods respectively. MAE, RMSE and $R^2$ were selected as verifiable indicators. The ensemble learning of CART is superior to the single CART algorithm (see TABLE III). MAE and RMSE were reduced by 9.4 and 13.2 percent respectively, and $R^2$ increased by 0.22. Fig. 4 shows the relationship between the estimated and reference values. As shown in Fig. 4a, ISP's high value area was underestimated. In Fig. 4b, the slope reaches 0.86, intercept is only 4.62. the overestimation has been improved.

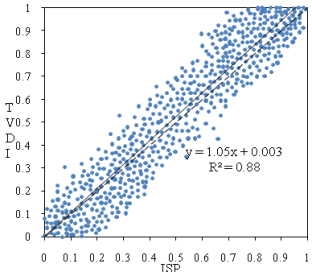



Fig. 5. Scatter plot of TVDI and ISP

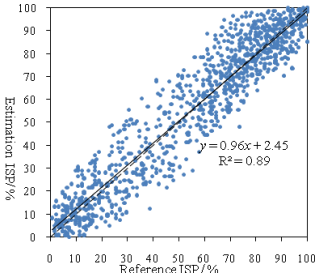



Fig. 6. Scatter plot of estimating and reference

There is underestimation in the high value region of ISP, which was indicated by the fitted line deviating far from the line y = x as the ISP increases. It was found that there is a good correlation between Temperature Vegetation Dryness Index (TVDI) [13]and ISP. Then the rules extracted by using the correlation were used to eliminate the overestimation. Because the high values of ISP usually occur in dense downtown areas, a 10000m buffer was made for the central part of Suzhou, Wuxi and Changzhou. 600 pixels were taken randomly and their ISP values were normalized. Then the correlation between the ISP and TVDI values of the 600 pixels were carried out. As shown in Fig. 5, $R^2$ is 0.88, slop is 1.05, intercept is only 0.003. And the fitted line gradually shift away from the line of y = x when ISP value increases, indicating that the value of TVDI is higher than that of ISP with the increase of ISP. Consequently, the rules can be extracted as follows: in the buffer zone, if the value of one pixel's ISP is no more than 0.7 and TVDI greater than 0.7, the value of ISP of this pixel is the mean of all pixels in a 5*5 window with the value greater than 0.7. Then the results are checked by using 1000 validation samples. As shown in the Fig. 6, though the value of $R^2$ has no change, the RMSE decreased to 10.01%, the fitted line and the line of y = x nearly overlap. The high value zone of ISP has been improved.

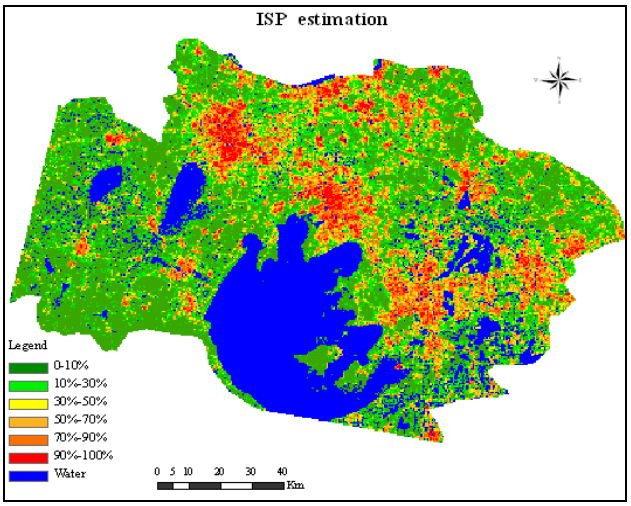



Fig. 7. ISP estimating of study area

The final estimates of ISP as shown in Fig. 7, the high value area of ISP is mainly distributed in city centre of Suzhou, Wuxi and Changzhou and the street center of township. Due to factors such as high density population, crowded buildings and heavy traffics and other factors, the value of ISP is basically greater than 90%.The value gradually declines from the city center to the suburb due to the decline of population and the widely distributed ponds, rivers and forest land. Although ISP has fallen, it was greater than 50%. Though the ISP of the green land in the city and the transition zone from city to rural areas has dropped due to better vegetation coverage, these green land areas are separated by the surrounding impervious surface. Moreover, the size of patch is small and the landscape is fragmented. So the ISP is still greater than those in the other farmland and vegetation areas, mostly between 30% and 50%. The value of ISP less than 30% usually occurs in the area with good vegetation coverage and without large area of surrounding of impervious surface.

## V. Conclusions

In this paper, VPRS was introduced into the CART analysis to reduce the redundancy of feature set. Finally, attribute variables with less correlation and satisfactory classification accuracy were chosen, and then the initial classification results were improved by using TVDI threshold value.

1. The verification shows that the method is valid and reliable. ISP estimation accuracy is higher than that of single CART algorithm: root mean square error has been reduced by 13.2% and $R^2$ has been improved by 0.22. The overestimation of low ISP values in single CART algorithm was improved effectively.
2. Post-processing rules generated by using the correlation between ISP and TVDI were used to improve the overestimation of low values in the results of CART algorithm. The fitting line slope of ISP estimated value and reference was improved from 0.86 to 0.96 and intercept decreased from 4.62 to 2.45. This indicates that the accuracy of ISP estimates has been improved dramatically.
3. However, there are certain spectral confusion between classes such as bare soil in farmland and unused land in city, both of which are difficult to be distinguished, and this may cause the overestimation of low ISP values. In addition, the sensitivity of CART algorithm to the size of samples is still to be resolved in the future study.

## Acknowledgment

This research is supported by the Fundamental Research Funds for the Central Universities (No. 2009B1020127) and special research funding for public industry welfare from the Ministry of Water Resources (No.201101024).

## References

[1] Wan H, Wu B F, Li X S, etc. Extraction of impervious surface in Hai Basin using remote sensing[J]. Journal of Remote Sensing, 2011(02): 388-400.

[2] Herold N D, Koeln G, Cunnigham D. Mapping impervious surfaces and forest canopy using classification and regression tree (CART) analysis[C], 2003.

[3] Ma X M, Li X l, Li X F, etc. Extracting valley impervious surfaces and it's change information based on the technology of data mining[J]. Bulletin of Surveying and Mapping, 2008(12): 34-37.

[4] Liao M SH, Jiang L M, Lin H, etc. Estimating urban impervious surface percent using Boosting as a refinement of CART analysis[J]. Geomatics and Information Science of Wuhan University, 2007(12): 1099-1102.

[5] Yang L, Huang C, Homer C G, et al. An approach for mapping large-area impervious surfaces: synergistic use of Landsat-7 ETM+ and high spatial resolution imagery[J]. Canadian Journal of Remote Sensing. 2003, 29(2): 230-240.

[6] Lu D, Weng Q. Spectral mixture analysis of the urban landscape in Indianapolis with Landsat ETM+ imagery[J]. Photogrammetric Engineering and Remote Sensing. 2004, 70(9): 1053-1062.

[7] Ridd M K. Exploring a VIS (vegetation-impervious surface-soil) model for urban ecosystem analysis through remote sensing: comparative anatomy for cities†[J]. International Journal of Remote Sensing. 1995, 16(12): 2165-2185.

[8] Ziarko W. Variable precision rough set model[J]. Journal of computer and system sciences. 1993, 46(1): 39-59.

[9] Wan L. Research of rules extraction method based on VPRSM in incomplete information systems[D]. China University of Petroleum (East China), 2009.

[10] Shao W. Water spatial element recognition research based on feature selection[D]. Hohai university,2011

[11] Chang Z l, Zhou Q M. Method based on variable precision rough set to build decision tree [J]. Computer Engineering and Design. 2006(17): 3175-3177.

[12] Zhao Y l, Wan J W, Gu SH SH, Choice of threshold value based on variable precision rough sets[J]. Control and Decision, 2007(01): 78-80.

[13] Wang W, Zhang Y J, Drought Monitoring Index and its Application in Semiarid Area. Geo-Information Science, 2008(02): 273-278.