Korean Journal of Remote Sensing, Vol.22, No.3, 2006, pp.199~209

An Adjustment for a Regional Incongruity in Global Land Cover Map: case of Korea

Youn-Young Park*, Kyung-Soo Han*†, Jong-Min Yeom, and Yong-Cheol Suh***

Dept. of Satellite Information Science, Pukyong National University*

Dept. of Atmospheric Science, Pukyong National University**

Abstract : The Global Land Cover 2000 (GLC 200) project, as a most recent issue, is to provide for the year 2000 a harmonized land cover database over the whole globe. The classifications were performed according to continental or regional scales by corresponding organization using the data of VEGETATION sensor onboard the SPOT4 Satellite. Even if the global land cover classification for Asia provided by Chiba University showed a good accuracy in whole Asian area, some problems were detected in Korean region. Therefore, the construction of new land cover database over Korea is strongly required using more recent data set. The present study focuses on the development of a new upgraded land cover map at 1 km resolution over Korea considering the widely used K-means clustering, which is one of unsupervised classification technique using distance function for land surface pattern classification, and the principal components transformation. It is based on data sets from the Earth observing system SPOT4/VEGETATION. Newly classified land cover was compared with GLC 2000 for Korean peninsula to access how well classification performed using confusion matrix.

Key Words : land cover, K-means clustering, the principal components transformation, SPOT4/VEGETATION.

1. Introduction

Use of remotely sensed data from satellite makes susceptible to various interpretations and extraction of thematic information, such as land cover and use, about Earth's surface at multiple spatial scales. In particular, the land cover maps are needed for global climate and ecosystem process models, as well as to characterize the distribution and status of major land surface types for environmental and ecological applications (Liang, 2001). Richardson (1922) first noted the importance of the role of the vegetation representing in numerical weather prediction. Recent climate modeling has also focused on the representation of CO_2 fluxes between the land surface and the atmosphere (Sellers *et al.*, 1996) in order to represent explicitly the biomass evolution. This corroborates the fact for having a thorough and realistic description of the land surface characteristics in meteorological models (Masson *et al.*, 2003).

Received 2 April 2006; Accepted 21 June 2006.

† Corresponding Author: K. – S. Han (kyung-soo.han@pknu.ac.kr)

During the last two decades, numerous scientific issues have addressed for the improvement and update of land cover database at the regional, continental and global scales. The availability of Landsat and other satellite images of the earth's surface greatly increased our capability to obtain up-to-date land cover information around the globe. So far, however, the use of these and similar high resolution data has suffered from the infrequent coverage and high costs, in addition to the high data volume (Cihlar, 1996). As an interim step, effort has in recent years been directed toward the use of medium resolution optical data such as obtained by the NOAA Advanced Very High Resolution Radiometer (AVHRR), Moderate Resolution Imaging Spectroradiometer (MODIS), and VEGETATION (VGT). The spatial resolution yields crucial matter of discussion since quite different results could be obtained according to the resolution of the final product. The relative importance of this issue depends on the geographic location on the study area. Considering meso-scale research, user's requirements indicate that 1 km resolution is highly desired. Despite the fact that meteorological models consider larger grid resolution for meso-scale applications, the sub-grid information is mandatory for an appropriate aggregation of the fluxes.

A multitemporal image classification, as one of the most often used techniques for information extraction, has been applied for the national scale land cover mapping. The actual mulitemporal image classification may be performed by supervised classification and unsupervised classification approaches (Richards, 1993; Jensen, 1995; Lillesand and Kiefer, 2000). In the supervised classification, sufficient training pixels should be available for each class to progress the signature representativies of those classes (Huang, 2002). However, the advent of satellite-borne sensors for acquiring remotely sensed data has indicated that for many applications purely supervised classification is no longer feasible, because small, irregular, or sparsely distributed features in a relatively large area being analysed make it difficult to locate these features precisely and collect a sufficient number of representative training samples from them (Swain, 1978). Unlike supervised classification, neither of prior knowledge nor training sets is required to produce a classification map in the unsupervised methods (or clustering methods). Therefore, the image can be automatically segmented into spectrally distinct unknown classes.

There are currently two global 1 km resolution land cover products available, derived from the Advanced Very High Resolution Radiometer (AVHRR) data. The first was produced by the U.S. Geological Survey for the International Geosphere-Biosphere Programme (IGBP) and the second by the University of Maryland (UMD). These databases show a good distribution of land types with good accuracy in Korea. However it should be updated for new surface condition because these were established in 1993. The Global Land Cover 2000 project (GLC 2000), as a most recent issue, is to provide for the year 2000 a harmonized land cover database over the whole globe (http://www.gvm.sai.jrc.it/glc2000/defaultGLC2000.htm). The classifications were performed according to continental or regional scales by corresponding organization using the data of VGT sensor onboard Satellite Probatoire de l'Observation de la Terre (SPOT). Even if GLC 2000 land cover classification for Asia provided by Chiba University (Tateishi *et al.*, 2003) showed an acceptable accuracy in whole Asian area, some problems were detected in Korean region. The distribution of the classes in this map is excessively rough (96% of Korean territory is covered by only two land types), in particular, urban areas were too underestimated. In addition, the 86 regions selected as ground truth data, which are used

for training and validation in their supervised classification do not include any Korean site. Therefore, the construction of new land cover database over Korea is strongly required using more recent data set. The aim of this study is to obtain a newly upgraded land cover map at 1 km resolution over Korea considering the widely used K-means clustering, which is one of unsupervised classification technique using distance function for land surface pattern classification, and the principal components transformation.

2. Data

The VGT instrument was launched on-board SPOT-4 in March 1998. Unlike scanner sensors (e.g. AVHRR), the VGT instrument uses the linear-array technology and thus produces high-quality imagery at coarse resolution with greatly reduced distortion. Three standard VGT products are available: VGT-P (physical product), VGT-S1 (daily synthesis product) and VGT-S10 (10-day synthesis product). VGT-S1 products provide the surface reflectance obtained after pre-processing, such as, geometric and atmospheric corrections. The geometric accuracy is less than 0.3 pixel for local distortion. Pixels are sampled using uniform grid spacing, allowing to correct distortion for inter-band registration, satellite orbit, attitude and elevation. The atmospheric correction has been applied to the VGT-P images for Reyleigh scattering, ozone light absorption, aerosol extinction, and water vapor using the SMAC code (Rahman and Dedieu, 1994). S10 is computed from all the passes on each location acquired during 10 day periods. The periods are defined according to the legal calendar: from 1st to 10th, from 11th to 20th, from 21st to the end of each month. A 10-day synthesis (S10) based on the selection of the "best" measurement of VGT-S1 pixels on the entire period. The selection is based on the Maximum Value Composite (MVC) approach for NDVI, as it is commonly accepted today, even if many problems associated to that selection are identified. This technique helps minimize the effect of variability in atmospheric optical depth and eliminate most cloudy pixels.

72 synthetic images on 10-day basis (S10 product) acquired fore 2 years (from April 2001 to March 2003) was used in the analysis. The data are being received by the Kiruna station (Sweden), processed and archived by the VITO production center located in Belgium and distributed by Spot-Image. All S10 images with spatial resolution of 1km in Platte Carrée projection were collected for South-East Asia category, and the Korean peninsula, as area of interest, was selected for geographical domain [124°E, 43°] - [131°E, 33°N]. Pixels on sea or great lakes are set to zero; the sensor is not programmed above the oceans. Unfortunately, above water the MVC algorithm gives bad results when clouds are present, as it chooses clouds that have a higher NDVI value instead of water. For this reason, the presence of water is determined from a sea/land indicator, derived from the Digital Chart of the World.

We used IGBP and UMD for the accuracy assessment as reference dataset. Both use data obtained from AVHRR sensor embarked on the National Oceanic and Atmospheric Administration (NOAA) satellite. Table 1 summarizes the key similarities and differences between the IGBP and UMD products. The data derived from the AVHRR that were used in the two classification sequences were collected based on monthly maximum Normalized Difference Vegetation Index (NDVI) composites collected from April 1992 to March 1993 inclusive. The IGBP used the 12 monthly maximum NDVI data while the UMD used all five AVHRR

Table 1. Characteristics of the IGBP DISCover and University of Maryland global land cover products.

Characteristics	IGBP	UMD
Sensor	- AVHRR	- AVHRR
Time series of data	- April 1992 ~ March 1993	- April 1992 ~ March 1993
Input	- 12 Monthly NDVI composites	- 41 metrics derived from NDVI - All available channel data
Classification method	- Unsupervised classification	- Supervised classification
Classification scale	- Global	- Global
Number of Classes	- 17	- 14

Table 2. Aggregated IGBP and UMD classes corresponding to reference classes and number of agreed pixels as reference data set.

Reference Class	Corresponding IGBP classes	Corresponding UMD classes	Number of Agreed pixels
Forests	- Evergreen Needleleaf Forest - Evergreen Broadleaf Forest - Deciduous Needleleaf Forest - Deciduous Broadleaf Forest - Mixed Forests	- Evergreen Needleleaf Forest - Evergreen Broadleaf Forest - Deciduous Needleleaf Forest - Deciduous Broadleaf Forest - Mixed Forests - Woodland	147998
Openlands	- Open Shrublands - Savannas - Grasslands	- Open Shrublands - Grasslands	739
Cropland	- Cropland - Cropland/Natural Vegetation	- Cropland	28437
SVA	- Permanent Wetlands - Urban and Built-Up - Barren or Sparsely Vegetated	- Bare grounds - Urban and Built-Up	790
Total pixels of reference set			177964

SVA: Sparsely vegetated areas

channels as well as the NDVI in deriving 41 multi-temporal metrics from the 12 monthly composites. We set four classes as reference class using the well-known land cover classes (forest, openlands, cropland, and sparsely vegetated areas). The UMD and IGBP classes allocated to reference classes are listed up in Table 2. The reference data set was determined from the selection of agreed pixels in UMD and IGBP classes corresponding to each reference class (Table 2).

3. Principal Components Transformation

Principal components analysis (PCA) was used before clustering step. In data mining we often encounter situations where there are a large number of variables in the database. In such situations it is very likely that subsets of variables are highly correlated with each other. The accuracy and reliability of a classification or prediction model will suffer if you include highly correlated variables or variables that are unrelated to the outcome of interest. Superfluous variables can increase the data-collection and data-processing costs of deploying a model on a large database. The dimensionality of a model is the number of independent or input variables used by the model. One of the key steps in data mining is finding ways to reduce dimensionality without sacrificing accuracy. PCA is a mathematical procedure that

Table 3. The result of principal component analysis.

PC #.	$\lambda \times 10^3$	σ, (%)	PC #.	$\lambda \times 10^3$	σ, (%)	PC #	$\lambda \times 10^3$	σ, (%)
1	3714.3	93.49	25	1.05	99.56	49	0.32	99.89
2	103.0	96.09	26	1.02	99.59	50	0.30	99.89
3	32.10	96.90	27	0.89	99.61	51	0.29	99.90
4	24.80	97.52	28	0.85	99.63	52	0.27	99.91
5	12.30	97.83	29	0.81	99.65	53	0.28	99.91
6	10.90	98.11	30	0.74	99.67	54	0.25	99.92
7	8.83	98.33	31	0.65	99.69	55	0.24	99.93
8	6.94	98.50	32	0.60	99.70	56	0.25	99.93
9	5.74	98.65	33	0.59	99.72	57	0.23	99.94
10	4.31	98.76	34	0.56	99.73	58	0.22	99.94
11	4.09	98.86	35	0.55	99.74	59	0.21	99.95
12	3.57	98.95	36	0.49	99.76	60	0.21	99.95
13	3.04	99.03	37	0.48	99.77	61	0.20	99.96
14	2.90	99.10	38	0.46	99.78	62	0.19	99.96
15	2.88	99.17	39	0.43	99.79	63	0.17	99.97
16	2.10	99.22	40	0.42	99.80	64	0.18	99.97
17	1.96	99.27	41	0.42	99.81	65	0.17	99.98
18	1.89	99.32	42	0.41	99.82	66	0.16	99.98
19	1.71	99.36	43	0.38	99.83	67	0.15	99.99
20	1.57	99.40	44	0.39	99.84	68	0.14	99.99
21	1.53	99.44	45	0.37	99.85	69	0.13	99.99
22	1.31	99.48	46	0.36	99.86	70	0.12	99.99
23	1.24	99.51	47	0.33	99.87	71	0.10	100.00
24	1.14	99.54	48	0.33	99.88	72	0.10	100.00

PC: Principal component, λ: Eigenvalue, σ: Accumulated total variance

transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. The objective of principal component analysis is to reduce the dimensionality (number of variables) of the dataset but retain most of the original variability in the data. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. In general, PCA is performed to reduce the data volume while retaining significant information. However, the goal of PCA in this study is 'how many principal components should be retained without discarding important information carried in the original data?' as well as the reduction of data volume. Therefore we do not apply any specific rule, such as Kaiser's rule and rule-N, for the component selection. To select the most informative components, we have used a threshold on the total explained variance by the PCA. For 72 initial components, 13 principal components were retained, representing 99% of variance in original data set. Table 3 shows the distribution of the Eigenvalues for each component and Accumulated total variances.

4. Unsupervised Clustering and Post-treatment

In order to generate classified a multi-temporal image products generated using each of these image fusion techniques, an unsupervised classification procedure using the K- means clustering algorithm was widely used. The techniques aforementioned in

the previous section are using an unsupervised procedure based on the K-means clustering algorithm (Algorithm AS 136, Hartigan and Wong, 1979). The examination carries on the clusters generated from the spectral characteristics of the normalized imagery (Jensen, 1995). In this way, each cluster is statistically separable. This method uses minimum distance criteria and resembles the k-nearest-neighbour rule method. The requirement for clusters classification by distance functions is that the pattern classes tend to have clustering properties. The K-means algorithm is based on the minimization of a performance index, which is defined as the sum of the squared distances from all points in a cluster domain to the cluster center, a Euclidian distance. K-means clustering allocates each pixel to one of k groups or clusters to minimize the within-cluster sum of square. In this case, the total sums of squares, E, within each cluster is computed as the sum of the centered sum of squares over all non-missing values of each of the variables. That is,

$$E = \sum_{i=1}^{K} \sum_{j=1}^{p} \sum_{m=1}^{n_i} f_{v_{im}} w_{v_{im}} \delta_{v_{imj}} (x_{v_{imj}} - \overline{x}_{ij})^2 \qquad (1)$$

where v_{im} denotes the row index of the mth observation in the ith cluster in the matrix x; n_i is the number of rows of x assigned to group i; f denotes the frequency of the observation; w denotes its weight; δ is zero if the jth variable on observation v_{im} is missing, otherwise δ is one; and $\overline{x}_{ij}$ is the average of the non-missing observations for variable j in group i. This method sequentially processes each observation and reassigns it to another cluster until the total within-cluster sums of squares is increased. The major issues that are inherent to an unsupervised classification relate to the arbitrary selection of the number of classes and their labeling. Fixing the initial number of allowed clusters is a critical stage. In some situations, an inappropriate number can lead to an ill-posed problem. Likelihood ratios, Bayesian techniques and Monte Carlo cross-validation are amongst the more popular probabilistic approaches for clustering. In non-probabilistic methods, a regularization approach is often adopted which is penalizing given a large number of clusters (Strehl, 2002). However, if the data is labeled, then mutual information between cluster and class labels can be used to determine the number of clusters. In this study, the initial number of clusters is arbitrarily fixed to 20. The K-means clustering analysis over Korea was performed using 13 selected principal components as input variables.

The post-treatment of the classification results was necessary and focused on the labeling and the agglomeration of the clusters. The initial 20 cluster image was agglomerated into an image with 12 clusters. Then, the labeling of clusters was carried out considering prior knowledge and the confusion matrix with IGBP and UMD databases. However, urban detection is not simple because this consists of very complex surface types. The study performed an urban mask using previous database, UMD land cover, which was established in 1993. The use of this urban mask means that new urban area due to the increase of the urban areas during 13 years should be identified. Figure 1 shows an example for the case of Seoul, Pyongyang, Busan, Hamhung-Hungnam, Deagu, Chungjin and Deajun. In our first run of classification, some part of sparsely vegetated areas were mainly found around the cities detected in 1993 (Figure 1a-1, 1b-1, 1c-1, 1d-1, 1e-1, lf-1, and 1g-1). The majority of these pixels classified in sparsely vegetated areas around the cities of South Korea were identified into the class "artificial surfaces" by a supervised classification from a priori knowledge (Figure 1a-2, 1b-2, 1c-2, and 1d-2). This parts of North Korea were also identified to "artificial surfaces" considering the fact that this phenomenon appeared over the cities of North Korea, although we

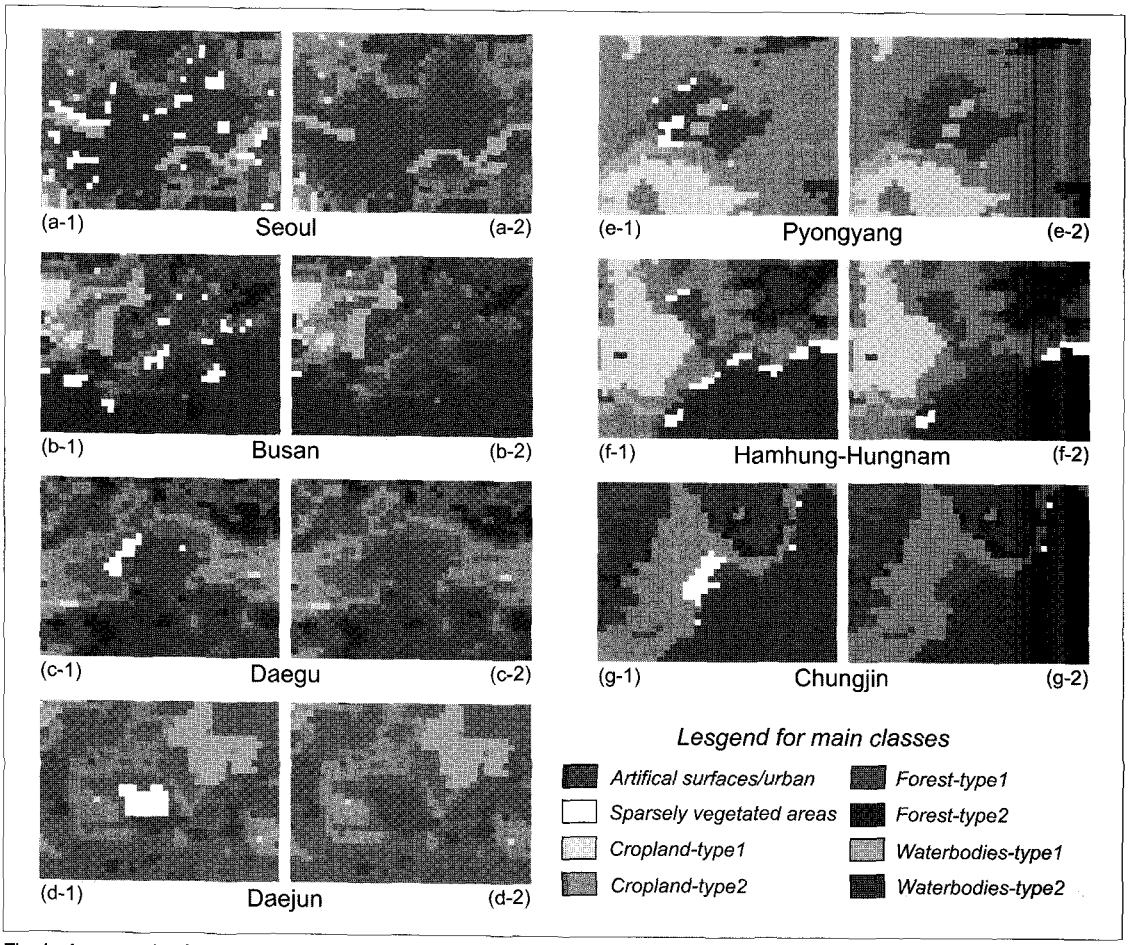

Fig. 1. An example of new urban area detection (before: a-1, b-1, c-1, d-1, e-1, f-1, g-1; after: a-2, b-2, c-2, d-2, e-2, f-2, g-2).

have not useful information for North Korea (Figure 1e-2, 1f-2, and 1g-2). The final result of the classification for the Korean Peninsula is shown in Figure 2 The classification map consists of 12 classes (Table 4). Forests (broad-leaved, coniferous, and mixed forests) and croplands (cropland and grassed cropland) dominate the study area. In particular, mixed forest has a high rate among other classes (Table 4).

5. Accuracy Assessment and Comparison

In this section, Accuracy assessment and comparison were performed to represent accuracy and reliability for new classified land cover. In regard to accuracy assessment, the most widely used method is the computation of confusion or error matrix (Foody, 2002). As a simple cross-tabulation of the mapped class label against that observed in the ground or reference data for a sample of cases at specified locations, it provides an obvious foundation for accuracy assessment (Campbell, 1996; Canters, 1997; Foody, 2002). The accuracy assessment through confusion matrix was performed, in order to access how well the new classified land cover map compared with those. The new land cover was used 72 synthetic images on 10-day basis (S10 product)

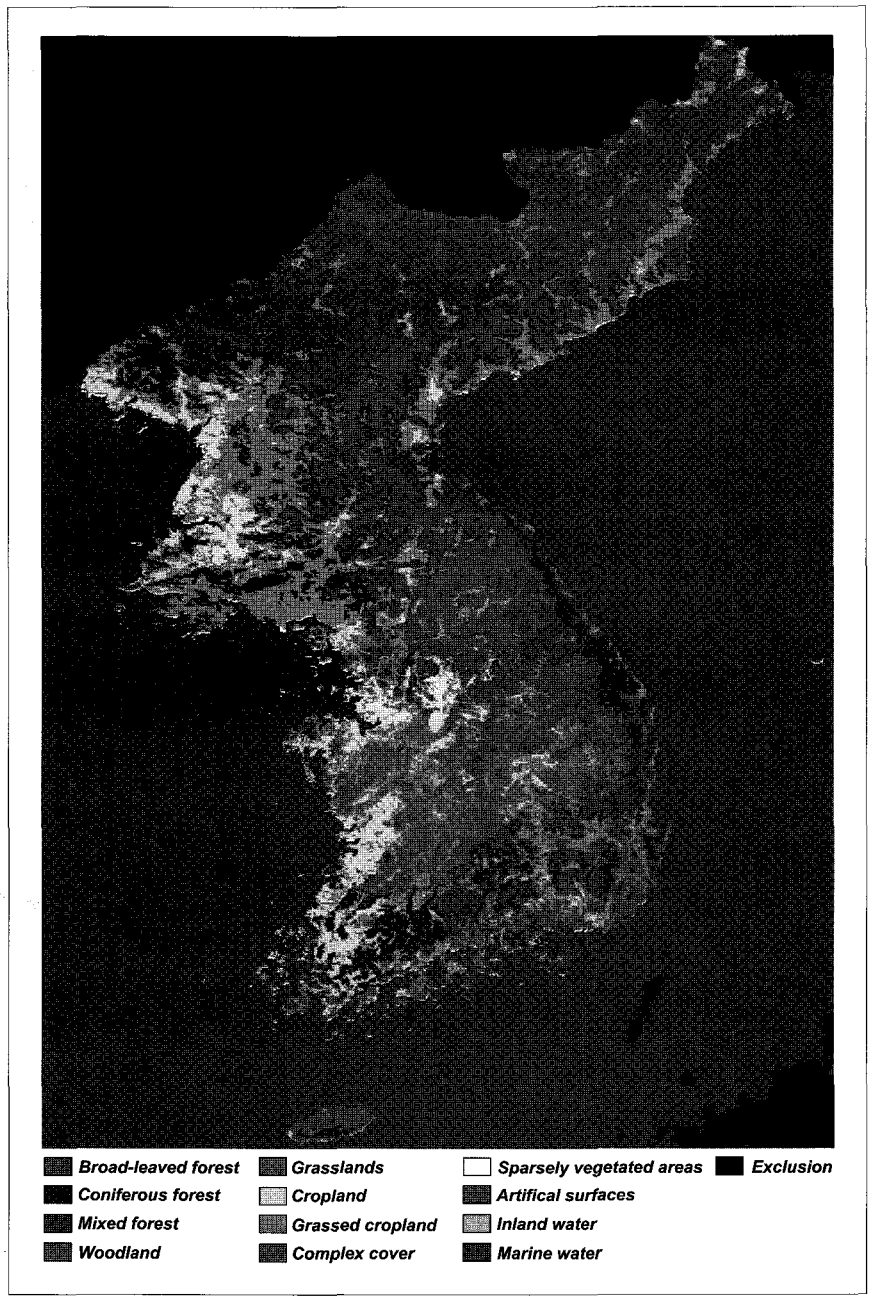

Fig. 2. The final result of new land cover at 1km resolution for the Korean Peninsula.

acquired fore 2 years (from April 2001 to March 2003) that were derived from VEGETATION-1 & -2. An unsupervised classification method using K-means clustering was performed to produce the new land cover. The other side, GLC for Korean peninsula used the 12 monthly maximum NDVI data from November 1999 to December 2000 and Supervised maximum likelihood method was performed.

Each column of the confusion matrix represents the new classified class types, while each row

Table 4. The legends of newly classified land cover and inland contribution of each class for the Korean Peninsula.

Code	Class name	Inland contribution (%)
1	Broad-leaved forest	18.36
2	Coniferous forest	10.80
3	Mixed forests	29.48
4	Woodland	5.32
5	Grasslands	0.82
6	Cropland	6.27
7	Grassed cropland	19.74
8	Complex cover	7.08
9	Sparsely vegetated areas	0.59
10	Artificial surfaces	0.66
11	Inland water	0.88
12	Marine water	-

represents the reference data set from UMD and IGBP agreement (Table 5). UMD and GLC2000 land cover were combined to select the reference data which was same category of the overlapping area of the four classes. The following statistical measures were calculated, based upon elements of the confusion matrix:

- User's accuracy (so-called reliability)
- Producer's accuracy (so-called accuracy)

The user's accuracy or reliability is defined as the probability that a sample from the classified image actually represents that category on the ground (Thunnissen and Noordman, 1996), and is expressed as:

$$\text{Reliability} = \frac{\text{number of correctly classified pixels in a class}}{\text{total number of pixels classified in that class}} \quad (5)$$

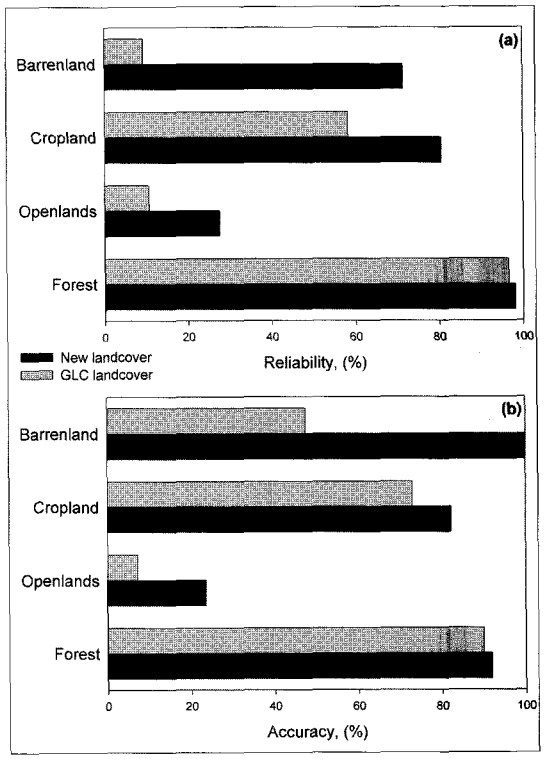

Fig. 3. The reliability and accuracy of new land cover classification GLC land cover for barrenland, cropland, openlands, forest.

The Producer's accuracy is the probability for a reference sample to be correctly classified (Thunnissen and Noordman, 1996). That is:

$$\text{Accuracy} = \frac{\text{number of correctly classified pixels in a class}}{\text{total number of pixels in that class as derived from the reference}} \quad (6)$$

The results of the accuracy assessment are represented in Figure 3 (a) and (b). Figure 3 (a) shows reliability value for new classified land cover and GLC2000 land cover for four classes as barrenland, cropland, openlands, forest. Figure 3 (b) is represents

Table 5. Confusion matrix of each class types (forest, openlands, cropland, sparsely vegetation) for new classified data and Reference set data from UMD/IGBP agreement.

Reference set data* -UMD/IGBPagreement	New classified data					
	Forests	Openlands	Croplands	Sparsely vegetation	Complex cover	Total
Forests	135923	63	5371	23	6618	147998
Openlands	18	174	311	211	25	739
Cropland	2724	397	23368	84	1864	28437
Sparsely vegetation	0	0	0	790	0	790
Total	138665	634	29050	1108	8507	177964

the results of the accuracy value for each classified land cover. The reliability value of new landcover is generally better for all land cover types, especially cropland (80.44%) and forest (98.02%), than these of GLC land cover in Figure 3 (a). The producer' accuracy of each classification land cover for barrenland, cropland, openlands, and forest was represented in Figure 3 (b). Producer's accuracy of individual classes for new land cover were consistently high, while accuracy of each classes for GLC2000 was lower than newly classified land cover. In conclusion, the user's and producer's accuracy for the new classified land cover are better than those of GLC2000 for Korea study area.

6. Conclusion

A new regional land cover product of the Korean Peninsula for the year 2000 has been produced. This product served as an input to GLC-2000 for Asia. The land cover classes were, however, aggregated to a higher level following a global legend. Several conclusions may be drawn from the current work. So, this study focuses on the development of a new upgraded land cover map at 1 km resolution over Korea considering the widely used K-means clustering, which is one of unsupervised classification technique using distance function for land surface pattern classification, and the principal components transformation. It is based on data sets from the Earth observing system SPOT4/VEGETATION. Most of Korean territory is covered by only two land types in GLC-2000, while new land cover consists of 12 classes. Accuracy assessment was performed for new classified land cover and GLC2000 over Korean study area to represent reliability and accuracy value for each Classification land cover using confusion matrix. The new classified land cover was generally better accuracy value than GLC2000 for all of class types. An improved classification land cover was presented over Korean study area using K-mean clustering and principal components transformation methods. It is used for input data for numerical weather forecast model to improve the predict accuracy and climate model to simulate the severely changing climate systems for Korean peninsula.

Acknowledgments

This work funded by the Korean Meteorological Administration Research and Development Program under Grant CATER 2006-4160.

References

Cihlar, J., H. Ly, and Q. Xiao, 1996. Land cover classification with AVHRR multichennel composites in northern environments, *Remote Sensing of Environment*, 58: 36-51.

Hartigan, J. A. and M. A. Wong, 1979. Algorithm AS 136: A K-means clustering algorithm, *Applied Statistics*, 28: 100-108.

Huang, K. Y., 2002. The use of a newly developed algorithm of divisive hierarchical clustering for remote sensing image analysis, *International Journal of Remote Sensing*, 23(16): 3149-3168.

Jensen, J. R., 1995. Introductory digital image processing: A remote sensing perspective. Englewood Cliffs, NJ, Prentice Hall, pp. 316.

Liang, S., 2001. Land-cover classification methods for multi-year AVHRR data, *International Journal of Remote Sensing*, 22(8): 1479-1493.

Lillesand, T. M. and R. W. Kiefer, 2000. Remote sensing and image interpretation. New York,

Wiley & sons, pp. 720.

Masson, V., J. L. Champeaux, F. Chauvin, C. Meriguet, and R. Lacaze, 2003. A global database of land surface parameters at 1-km resolution in meteorological and climate models, *Journal of Climate*, 16(9): 1261-1282.

Rahman, H. and G. Dedieu, 1994. SMAC: a simplified method for the atmospheric correction of satellite measurements in the solar spectrum, *International Journal of Remote Sensing*, 15(1): 123-143.

Richard, J. A., 1993. Remote sensing digital image analysis: an introduction. Berlin, Springer-Verlag, pp. 281.

Richardson, L. F., 1922. *Weather prediction by numerical process*, Cambridge University press, London.

Sellers, P. J., L. Bounoua, G. J. Collatz, D. A. Randall, D. A. Dazlich, S. O. Los, J. A . Berry, I. Fung, C. J. Tucker, C. B. Field, and T. G. Jensen, 1996. Comparison of radiative and physiological effects of doubled atmospheric CO_2 on climate, *Science*, 271: 1402-1406.

Strehl, A., 2002. Relationship-Based Clustering and Cluster Ensembles for High-Dimensional Data Mining, PhD thesis, The University of Texas at Austin, May 2002.

Swain, P. H., 1978. Fundamentals of pattern recognition in remote sensing. In Remote Sensing: the Quantitative Approach, New-York, McGraw-Hill, pp. 136-187.

Tateishi, R., H. Sato, and Z. Lin, 2003. GLC 2000 regional products, Asia, Chiba University, Japan.