A Cost Effective Reference Data Sampling Algorithm Using Fractal Analysis

Byoung-Kil Lee, Yang-Dam Eo, Jae-Joon Jeong, and Yong-Il Kim

A random sampling or systematic sampling method is commonly used to assess the accuracy of classification results. In remote sensing, with these sampling methods, much time and tedious work are required to acquire sufficient ground truth data. So, a more effective sampling method that can represent the characteristics of the population is required. In this study, fractal analysis is adopted as an index for reference sampling. The fractal dimensions of the whole study area and the sub-regions are calculated to select sub-regions that have the most similar dimensionality to that of the whole area. Then the whole area's classification accuracy is compared with those of sub-regions, and it is verified that the accuracies of selected sub-regions are similar to that of whole area. A new kind of reference sampling method using the above procedure is proposed. The results show that it is possible to reduce sampling area and sample size, while keeping the same level of accuracy as the existing methods.

Manuscript received December 26, 2000; revised July 26, 2001.

Byoung-Kil Lee (phone: +82 2 880 6286, e-mail: Basil@snu.ac.kr) is with the Department of Urban Engineering, Seoul National University, Seoul, 151-742, Korea.

Yang-Dam Eo(e-mail: Eo@purdue.edu) is with the Purdue University, West Lafayette, IN, 47907, USA.

Jae-Joon Jeong (e-mail: Hayoon@chollian.net) and Yong-Il Kim (e-mail: Yik@snu.ac.kr) are with the Department of Urban Engineering, Seoul National University, Seoul, 151-742, Korea.

I. INTRODUCTION

1. Background

The classification of remotely sensed imagery may introduce errors for various reasons. These errors are propagated to the consequent applications of the classification results. Therefore, the quantification of errors is required for applying the classification results to the related fields such as geographic information systems (GIS). In the accuracy evaluation process, the reference data are required for comparison, but obtaining a sufficient amount of reference data requires much time and tedious work for data sampling. The purpose of this study is to propose an alternate sampling method that can mitigate the workload of the reference data acquisition.

In Congalton's paper [1], [2] several well-known and plain reference data sampling methods are explained. The simple random sampling method is widely used in research projects, but the random sampling is inefficient in terms of time and cost. Therefore, the stratified random sampling can be preferred to the stratified land-cover for reducing the sample size, and the stratified sampling can be applied to geometric strata as well as land-cover strata. Some researches recommended the use of the systematic unaligned sampling method, but some other researchers mentioned that the cluster sampling is cost-effective. However, Janssen and van der Wel [3] suggested that poor accessibility and limited budgets may result in a cluster sampling method, and Congalton [1] pointed out that location of clusters is another drawback of cluster sampling method.

Arkansas Forestry Commission (AFC) developed Rapid Assessment Track (RAT) based on clustering. In RAT sampling, a total of 100 USGS 7.5 minute quadrangles were randomly selected in proportion to the relative area of Arkansas' 10 major

landforms [4]. Here randomly selected sites had limitations in representativeness. Stehman [5] used poststratification to improve the precision of accuracy assessment; *i.e.,* poststratification is an estimation technique, not a sampling design.

On the other hand, De Cola [6] improved the classification accuracy of each class by fractal analysis of classified Landsat images. Jaggi *et al.* [7] compared and analyzed the characteristics of the three fractal measurements (*i.e.,* isarithm method, triangular-prism method and variogram method) for remotely sensed images. Qiu *et al.* [8] studied the fractal characteristics of classes in hyperspectral imageries.

Recently, Kim *et al.* [9] explained the spatial characteristics of classification errors. There, it is noted that the distribution of error pixels is affected by the complexity of topography and land-cover.

2. Purpose

So far, there are large demands on the accuracy assessment method in nationwide project for minimizing direct data collection time, costs, and land access limitations, and maximizing representativeness. The objective of this study is to develop the sampling method with statistical representativeness and efficiency in terms of cost and time.

Therefore, this study proposes a sampling method using sub-regions that imply the spatial distribution of the entire image to obtain the representative and aggregated samples to complement the stratified random sampling. Here fractal dimension is used as an indicator to find the representative sub-regions of the whole area.

II. FRACTALS

Since Mandelbrot introduced the concept of a fractal, a lot of calculation methods of fractal dimensions have been developed for various kinds of spatial issues. The fractal has significant potential in measuring and analyzing the spatial and radiometric aspects of remotely sensed data [7], [10].

1. Basic Concepts of Fractals

In Euclidean geometry, lines have a dimension value of 1, but in fractal geometry the dimension of a curve varies from 1 to 2, depending on its complexity. Similarly, for a surface, the fractal dimension value lies between 2 and 3, and it reflects the overall characteristics of the surface [8]. Thus, fractal dimension can be a good descriptor of spatial complexity and variability.

Self-similarity is another important feature of fractals. In fractal space, there will always be several "sub-regions" that have the same characteristics as the "whole area" at every scale. Therefore, the fractal dimension is independent of scale, and is a variable of spatial complexities only. However, real topographic features have limitation in representations. So, the spatial characteristics of topography are not self-similar at all scales, and may have constant fractal dimensions within certain ranges of scale.

2. Isarithm Method

In this study, the isarithm method was used for calculating the fractal dimension of classified results; the algorithm has been slightly tailored to suit raster images. In the general isarithm algorithm, the length of the isarithm line (a certain pixel value) is determined first, and a binary image is generated using the isarithm line as a threshold value. Then the binary pixel values are compared with each other in both the row and column directions. If the pixel value differs from that of a pixel located at the step size distance, then count is increased. After the comparison is completed, the regression line of count on the step size in log scale is determined. The slope of the regression line is the fractal dimension [11]. However, since the algorithm is applied to classified results, the isarithm line must be the nominal pixel value, *i.e.,* class number in this study. Consequently, the fractal dimension is extracted from each binary image of a class (see Fig. 1). The step size of sub-regions should be equal to that of the entire image for each iteration, and should not be greater than half the image size [12].

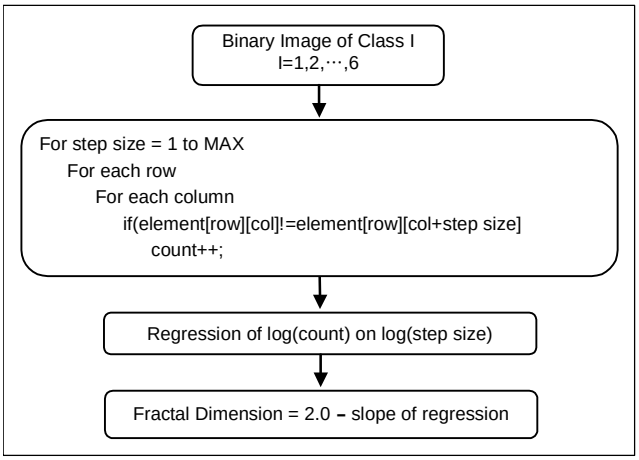

Fig. 1. Isarithm method.

III. TEST

1. Study Area

Landsat TM images for 12 map sheets of 1/50,000 scale at Seoul, Daejeon and Pohang regions in Korea were chosen for study areas, and the size of each image is 840×760. Linear polynomial transformation and the nearest neighborhood interpolation were used for

geometric correction. Considering terrain features of the study area, 6 classes were selected such as forest, water body, field, paddy, dry brook, and urban area. The Maximum Likelihood Classifier was used for classification. The reference data were edited from 1:50,000 scale digital maps and aerial photos.

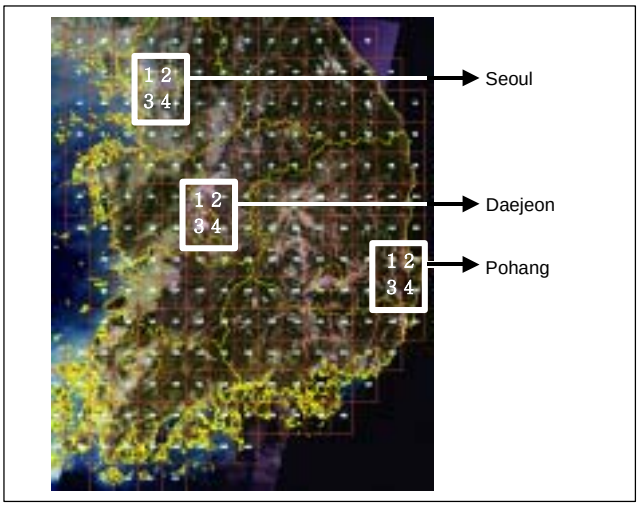

Fig. 2. Study area.

2. Reference Data Sampling Using Fractal Dimension and Accuracy Evaluation

First, binary images of each class were generated to select the sub-region for reference data sampling, and fractal dimension and area ratio of whole image were calculated for each class. Then they were calculated again for sub-regions of a given size. When the differences between the calculated values of sub-regions and those of the whole area were smaller than a pre-defined threshold, that sub-region was selected for sampling site (see Fig. 3). The window sizes of the sub-regions were 125×125, 100×100, 75×75, and 50×50 pixels, and sample sizes are 1000, 700, 500, and 200, respectively. The reference data for each class was randomly extracted from selected sub-regions for the proposed sampling method (Hereafter the proposed method is referred to as the fractal sampling method). The simple random sampling method and systematic sampling method were applied to compare with the fractal sampling method. Here, sample sizes were equal to that of the fractal method. Accuracy evaluations were performed with the confusion matrix.

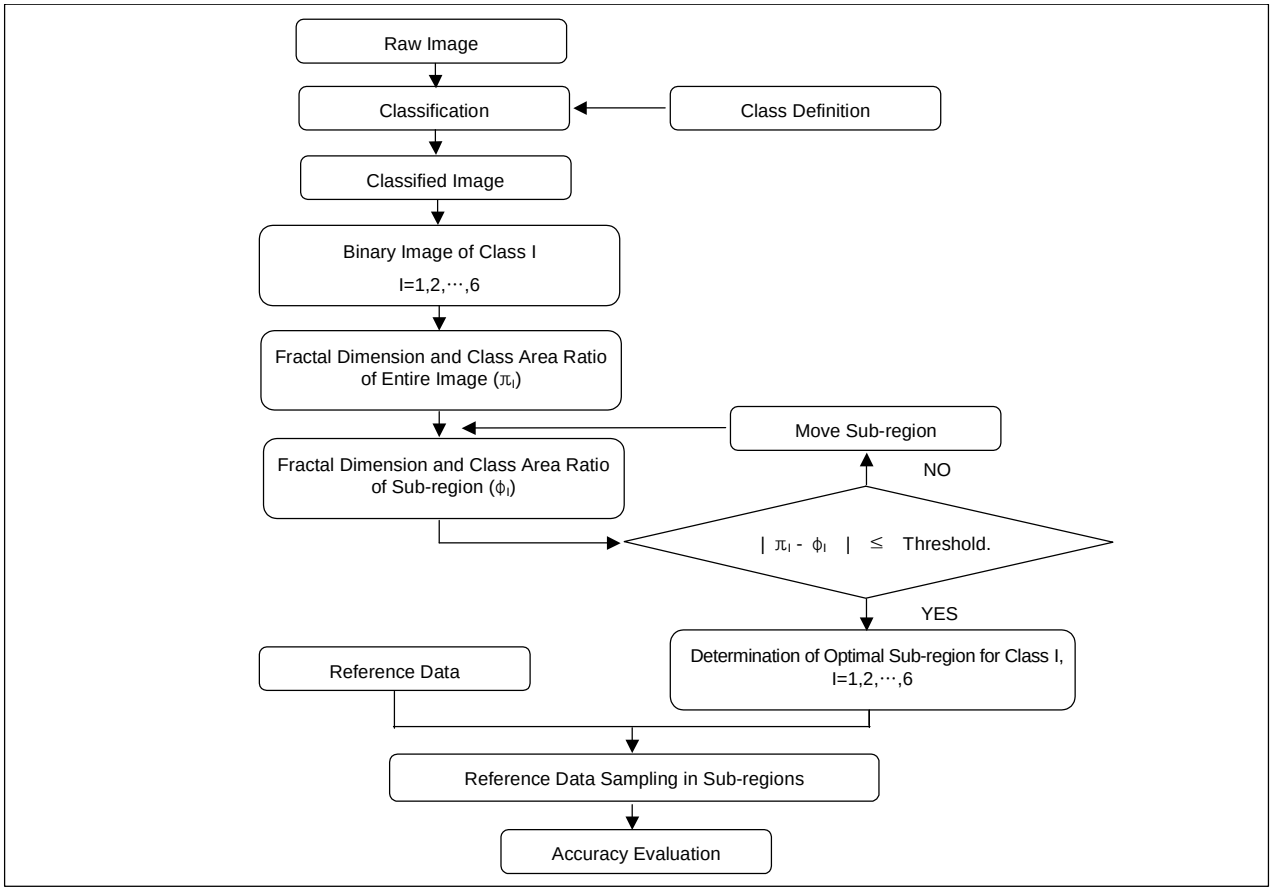

Fig. 3. Reference data sampling algorithm using fractal dimension.

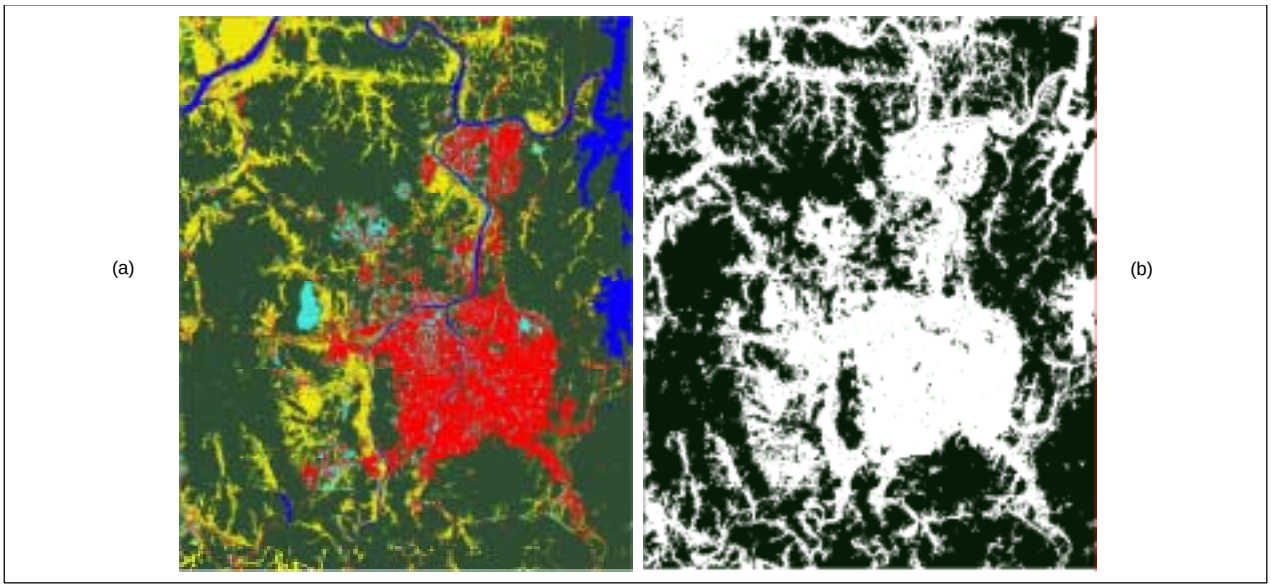

Fig. 4. (a) Classified image and (b) binary image of forest class for Daejeon-3 area.

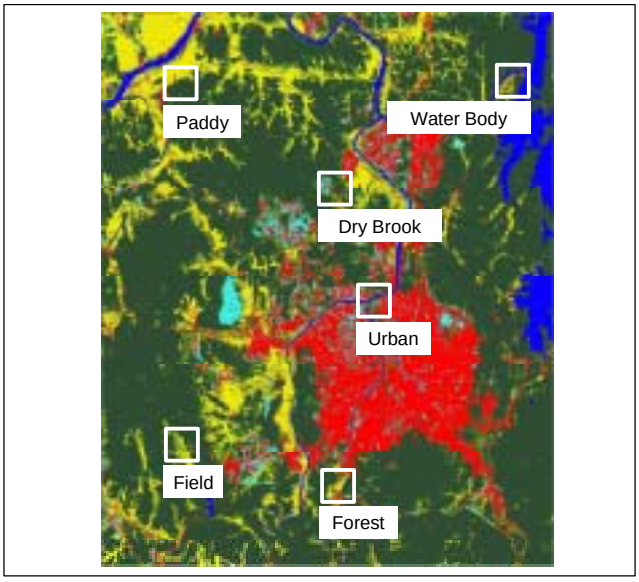

Fig. 5. Selected sub-regions for six classes of Daejeon-3 area at size of 50×50.

Calculating the area ratio as well as fractal dimension gives statistical meaning of the random sampling to the fractal sampling. Using the area ratio, the fractions of each class in the sub-regions are equal to that of the whole area.

In this study, six different binary images were created for six classes of each image. Zero was assigned to the pixels of the corresponding class, and one was assigned to the pixels of other classes (see Fig. 4). The size of sub-regions were 125×125, 100×100, 75×75 and 50×50 for the fractal sampling. Here, the step size was 2^4(=16) for calculating the fractal dimension at every size of the sub-regions. Figure 5 shows the final results of the selected sub-regions for six classes at the size of 50×50.

IV. RESULTS

1. Relationship between Fractal Dimension and Accuracy Estimation

Accuracies of population were evaluated as the reference accuracy values of whole area. In this study, the differences of area ratio, fractal dimension, and accuracy of sub-regions from those of whole area were computed. Figure 6 shows the relationship among the area ratio differences, fractal dimension differences, and accuracy differences. As shown in the figure, the accuracy discrepancy increases as the fractal dimension difference becomes larger.

2. Evaluation of Accuracy Estimation

In this study, four accuracy values were calculated using various sampling methods. Accuracy of population was described as reference. Accuracy estimation of existing methods, simple random sampling and systematic sampling, was extracted with the same sample size as the fractal sampling method. Accuracy estimation of the fractal sampling method was obtained using best or optimal sub-regions that have the minimum fractal dimension difference and the minimum area ratio difference. These accuracies are compared with each other and analyzed.

Table 1 shows the overall accuracies of 12 study areas. Tables 2 to 4 show the absolute values of accuracy differences between sampling methods.

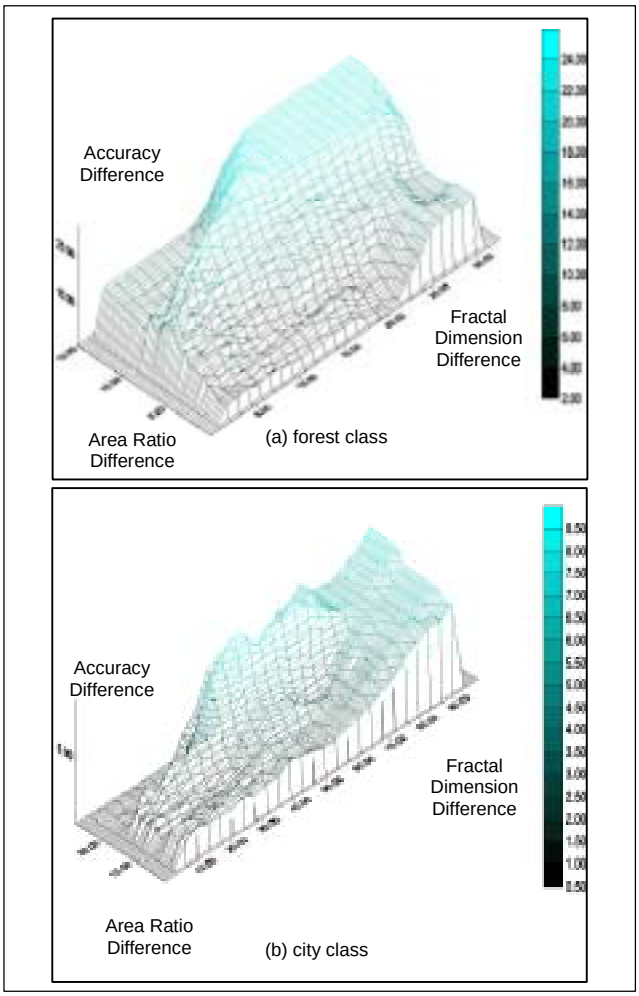

Fig. 6. Relationships among fractal dimension, area ratio, and accuracy differences.

The averages of accuracy differences by the size of the sub-regions are listed in Table 5, where the accuracy differences between fractal sample and population vary from 0.33% to 0.92%. Table 6 shows that the standard deviations of differences between fractal and population range from 0.15% to 0.60%, which means that the fractal sampling method has quite a good possibility for accuracy evaluation, and is considerably stable for practical use.

3. Analysis of Sampling Cost

The sampling area for simple random sampling or systematic sampling is 650㎢ per one 1/50,000 map. However, if the proposed method is applied to sampling, the area is decreased to 14㎢, 9㎢, 5㎢ and 2.25㎢ as the size of the sub-region is decreased to 125×125, 100×100, 75×75, and 50×50, respectively. This sub-region size is very small compared to the whole area.

Traveling distances, with sample size of 200, are about 228㎞ for systematic sampling, about 325㎞ for random sampling and about 48㎞ for fractal sampling per one 1/50,000 map.

So, it is possible to reduce the cost of reference data. In addition, the fractal sampling can estimate accuracy equivalently to existing sampling methods.

Table 1. Overall accuracies of population for each study area.

Area	Seoul-1	Seoul-2	Seoul-3	Seoul-4
Overall Accuracy (%)	73.60	78.05	82.97	78.12
Area	Daejeon-1	Daejeon-2	Daejeon-3	Daejeon-4
Overall Accuracy (%)	77.32	77.16	74.87	76.61
Area	Pohang-1	Pohang-2	Pohang-3	Pohang-4
Overall Accuracy (%)	74.93	78.25	70.85	72.28

4. Statistical Tests

Based on the results of our analysis, it can be assumed that the fractal samples are quite similar to the population itself as well as random samples and systematic samples. So, statistical tests have been performed to check whether the fractal samples have statistical significance or not. Statistical tests were the T-test of equality of means and the F-test of equality of variances. The test results are listed in Table 7.

The results from T-test of equality of means confirm the proposed method. The null-hypothesis is that the accuracy of population equals the accuracy of fractal sample. And, the F-test of equality of variances is another evidence. The null-hypothesis is that the variance of accuracy of population equals the variance of accuracy of fractal sample. The regions of rejection are obtained at a significance level of 5% for each test [13].

The formulas of test statistics are as follows:

Test statistics of T-test: $$T = \frac{\overline{X} - \overline{Y}}{S_p \sqrt{1/n_X + 1/n_Y}},$$

Test statistics of F-test: $$F = \frac{S_Y^2}{S_X^2}.$$

These two tests show that the null-hypothesis should be accepted, and that the fractal sampling method can extract a statistically representative sample.

Table 2. Accuracy differences among sampling methods for Seoul area (%).

Area/Sample Size \ Methods		Population vs. Random	Population vs. Systematic	Population vs. Fractal	Random vs. Fractal	Systematic vs. Fractal
1	1000	0.25	0.23	0.32	0.07	0.55
	700	0.12	0.18	0.38	0.26	0.55
	500	0.19	0.32	1.18	1.37	1.51
	200	2.19	1.96	0.63	1.56	1.33
2	1000	0.19	0.34	0.19	0.38	0.15
	700	0.42	0.05	0.38	0.79	0.33
	500	1.17	0.37	0.22	1.39	0.58
	200	0.68	0.14	1.80	1.12	1.94
3	1000	0.49	0.16	0.16	0.64	0.32
	700	0.40	0.40	0.46	0.06	0.85
	500	1.60	0.88	0.26	1.34	0.62
	200	0.40	0.52	0.83	1.23	1.35
4	1000	0.31	0.35	0.46	0.77	0.11
	700	0.43	0.38	0.43	0.00	0.81
	500	0.63	0.83	0.10	0.73	0.93
	200	1.30	1.28	1.79	3.10	3.08

Table 3. Accuracy differences among sampling methods for Daejeon area (%).

Area/Sample Size \ Methods		Population vs. Random	Population vs. Systematic	Population vs. Fractal	Random vs. Fractal	Systematic vs. Fractal
1	1000	0.45	0.33	0.39	0.06	0.07
	700	0.37	0.44	0.50	0.13	0.06
	500	1.95	0.37	1.16	3.11	1.53
	200	0.75	0.05	0.43	0.32	0.38
2	1000	0.24	0.30	0.37	0.61	0.67
	700	0.08	0.11	0.30	0.23	0.42
	500	0.31	1.50	0.38	0.07	1.88
	200	1.81	0.87	0.57	1.24	1.45
3	1000	0.36	0.07	0.50	0.14	0.43
	700	0.14	0.40	0.32	0.18	0.08
	500	0.84	1.23	0.77	1.61	0.46
	200	2.35	0.32	0.24	2.10	0.56
4	1000	0.27	0.38	0.22	0.05	0.16
	700	0.08	0.37	0.05	0.03	0.32
	500	1.57	0.81	0.55	1.03	1.36
	200	1.78	1.07	1.61	0.17	2.67

Table 4. Accuracy differences among sampling methods for Pohang area (%).

Area/Sample Size \ Methods		Population vs. Random	Population vs. Systematic	Population vs. Fractal	Random vs. Fractal	Systematic vs. Fractal
1	1000	0.45	0.32	0.01	0.44	0.31
	700	0.34	0.13	0.04	0.31	0.09
	500	0.75	0.06	0.61	0.14	0.55
	200	0.98	1.52	0.24	0.75	1.29
2	1000	0.16	0.10	0.47	0.32	0.38
	700	0.26	0.32	0.15	0.42	0.47
	500	1.55	0.31	1.26	2.81	0.95
	200	2.88	2.19	1.21	1.66	3.40
3	1000	0.24	0.04	0.47	0.24	0.44
	700	0.38	0.39	0.24	0.62	0.15
	500	1.69	0.75	1.28	0.41	2.04
	200	2.10	1.26	1.32	0.78	2.58
4	1000	0.08	0.21	0.38	0.29	0.17
	700	0.24	0.29	0.31	0.07	0.60
	500	1.21	0.44	1.41	2.62	0.97
	200	2.09	2.81	0.40	1.69	2.42

Table 5. Averages of accuracy differences by size of sub-region (%).

Sample Size \ Methods	Population vs. Random	Population vs. Systematic	Population vs. Fractal	Random vs. Fractal	Systematic vs. Fractal
1000	0.29	0.23	0.33	0.33	0.31
700	0.27	0.29	0.30	0.26	0.39
500	1.12	0.66	0.76	1.39	1.11
200	1.61	1.17	0.92	1.31	1.87
Average	0.82	0.59	0.58	0.82	0.92

Table 6. Standard deviations of accuracy differences by size of sub-region (%).

Sample Size \ Methods	Population vs. Random	Population vs. Systematic	Population vs. Fractal	Random vs. Fractal	Systematic vs. Fractal
1000	0.13	0.12	0.15	0.24	0.19
700	0.14	0.13	0.15	0.25	0.28
500	0.57	0.42	0.47	1.02	0.54
200	0.77	0.86	0.60	0.80	0.97
Average	0.75	0.60	0.47	0.84	0.85

Table 7. Statistical test results of population vs. fractal sampling.

Sample Size	Test Statistics of T-test	Test Statistics of F-test	Region of Rejection of T-test	Region of Rejection of F-test
1000	0.193	0.972	2.074	2.82
700	0.248	0.977		
500	0.318	0.866		
200	0.577	0.944		
For All Size	0.842	0.893	1.980	1.600

V. CONCLUSION

The results of this study are as follows.

First, it shows that the sub-regions, which can represent the whole image, can be extracted using fractal analysis.

Second, the fractal sampling method can make statistically representative and reasonable samples of the population, so it can be used in practice as an alternative sampling method.

Third, the fractal sampling method can select several sites for alternative, therefore, inaccessibility problems in simple stratified sampling can be reduced.

Forth, sampling sites can statistically represent the whole area, because sub-regions are selected based on fractal theory, Compared to this sampling sites are randomly selected with cluster sampling in general.

Last, this study shows that the sampling cost can be reduced using the proposed sampling method, namely fractal sampling.

In this study, only the isarithm method is applied for measuring the fractal dimension; however, other algorithms for fractal analysis shall be exploited in consequent studies.

While the locations of the samples for one class are close-by, the sub-regions for each class are located further apart. Since this may cause extra overhead in sampling, it is required to develop a selection algorithm that can make the sampling areas for all classes in one sub-region.

ACKNOWLEDGEMENT

The authors acknowledge, with appreciation, the advice for this paper by Henry Theiss who is an Assistant Professor at Purdue University, and support of the "BK21."

REFERENCES

[1] R.G. Congalton, "A Comparison of Sampling Schemes Used in Generating Error Matrices for Assessing the Accuracy of Maps Generated from Remotely Sensed Data," *Photogrammetric Engineering and Remote Sensing*, vol. 54, no. 5, 1988, pp. 593-600.

[2] R.G. Congalton, "A Review of Assessing the Accuracy of Classifications of Remotely Sensed Data," *Remote Sensing of Environment*, vol. 37, 1991, pp. 35-46.

[3] L.L.F. Janssen and F.J.M. van der Wel, "Accuracy Assessment of Satellite Derived Land-cover Data," *Photogrammetric Engineering and Remote Sensing*, vol. 60, no. 4, 1994. pp. 419-426.

[4] R.S. Dzur, M.E. Garner, K.G. Smith, W.F. Limp, D.G. Catanzaro, and R.L. Thompson, "Cooperative Accuracy Assessment Strategies for Sampling a Natural Landcover Map of Arkansas," *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium*, 1996, pp. 517-526.

[5] S.V. Stehman, "Cost-Effective Practical Sampling Strategies for Accuracy Assessment of Large-area Thematic Maps," *Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium*, 1996, pp. 485-492.

[6] Lee De Cola, "Fractal Analysis of a Classified Landsat Scene," *Photogrammetric Engineering and Remote Sensing*, vol. 55, no. 5, 1989, pp. 601-610.

[7] S. Jaggi, Dale A. Quattrochi, and Nina Siu-Ngan Lam, "Implementation and Operation of Three Fractal Measurement Algorithms for Analysis of Remote-Sensing Data," *Computers & Geosciences*, vol. 19, no. 6, 1993, pp. 745-767.

[8] Hong-lie Qiu, Nina Siu-Ngan Lam, Dale A. Quattrochi, and John A. Gamon, "Fractal Characterization of Hyperspectral Imagery," *Photogrammetric Engineering and Remote Sensing*, vol. 65, no. 1, 1999, pp. 63-71.

[9] Y.I. Kim, Y.D. Eo, and B.K. Lee, "Analyzing the Spatial Distribution Pattern of Image Classification Error," *Journal of the Japan Society of Photogrammetry and Remote Sensing*, vol.38, no. 2, 1999.

[10] Nina Siu-Ngan Lam and Lee De Cola, *Fractals in Geography*, PTR Prentice-Hall, 1993.

[11] Michael Batty and Paul Longley, *Fractal Cities*, Academic Press, 1994.

[12] Keith C. Clarke and Diane M. Schweizer, "Measuring the Fractal Dimension of Natural Surfaces Using a Robust Fractal Estimator," *Cartography and Geographic Information Systems*, vol. 18, no. 1, 1991, pp. 37-47.

[13] Y.W. Nam, *Quantitative Geography*, Peob Mun Sa, 1995.

Byoung-Kil Lee received the B.S. and M.S. degrees in urban engineering from Seoul National University, Korea in 1990 and 1992, respectively. And he got the Ph.D. degree in the field of spatial informatics at Seoul National University in 2001. He worked in GIS company from 1992 to 1997. Currently, he is a researcher in the Research Institute of Engineering Science at Seoul National University. His main research interests are digital photogrammetry, remote sensing, and GIS data construction and analysis.

Yang-Dam Eo received the B.S. and M.S. degrees in urban engineering from Seoul National University, Korea in 1989 and 1991, respectively. And he got the Ph.D. degree in the field of spatial informatics at Seoul National University in 1999. Currently, he is a visiting scholar in the civil engineering department at Purdue University. His main research interests are the satellite image classification and the analysis in the remote sensing community.

Jae-Joon Jeong received the B.S., M.S. and Ph.D. degrees in urban engineering from Seoul National University, Korea in 1996, 1998, and 2001, respectively. Currently, he is a researcher in the Research Institute of Spatial Data Korea, Inc. His main research interests are the remote sensing and GIS analysis, especially cellular automata and fractal.

Yong-Il Kim received B.S. degree in urban engineering from seoul National University, Korea in 1986. And he got the M.S. and Ph.D. degrees in the field of remote sensing in 1988 and 1991, respectively. He joined the faculty of Seoul National University in 1993, where he is currently working as an associate professor at the school of civil, urban, and geosystem engineering. He stayed at Cornell University for one year as a visiting researcher in 1997. His major research interests include remote sensing, global positioning system (GPS), geographic information systems, etc. And during past 10 years, he has been involved as project leader in several large projects such as standardization of digital road map databases, development of feature extraction algorithms for remote sensing, etc. At present, he is a member of Surveying Committee of National Geographic Institute, and also a director and editor of Journal of the Korean Society for Geo-Spatial Information Systems, Journal of the Korean Society of Geodesy, Photogrammetry, and Cartography, and Journal of the Korean Society of Remote Sensing.