CLASSIFICATION OF LAND-COVER BASED ON STATISTICAL VERIFICATION

Iwao Yokoyama, Engineer
Taisei Aerial Survey Co.Ltd.
1174-10 Nishimachi, Kurume, Fukuoka, 830
Chikashi Degichi, Research Associate
Minoru Numata, A Professor Emeritus
Dept. of Civil Eng., Kyushu University
6-10-1 Hakozaki, Higashiku, Fukuoka, 812
Kazumi Matsuo, Engineer
Chuubu Regional Construction Bureaus
Ministry of Construction
Japan
Commission WG Ⅶ/4

1. INTRODUCTION

In conventional supervised approaches, land-cover categories are decided a priori taking into account the usage of classification results. Usually, their decisions are assisted by experimental judgments of experts. The accuracy of classification largely depends on the selections of training data. In order to classify the developed areas where various uses of the land are intricately mixed, we face difficulty in obtaining good training data.

In unsupervised approaches using clustering algorithms, clusters are formed on the basis of similarity, or distance calculated from the digital values, of multispectral reflections. Generally, we can derive more detailed information on the usage of developed areas from MSS or TM data. However, difficulties remain in determining an effective number of categories and interpreting the clusters according to these categories.

This paper describes a land-cover classification method that determines the categories and interprets the clusters. The method is based on F- and t-tests in multiple regression analyses. We present the application of this method to Landsat MSS, TM, and to airborne MSS data.

2. CLASSIFICATION METHOD

2.1 DETERMINING LAND-COVER CATEGORIES

In our classification method, the initial categories and their contents must be prepared in detail. We decided upon the categories through visual interpretation of the aerial color photographs. This was done by meshing the training area (we named mesh-areas), as shown in Fig.1. Then, using grids, we visually interpreted each of the cells in the aerial photographs into our initial categories corresponding to the ground resolution of MSS. The visually-interpretated photo-data serves the role of "ground-truth" in this method.

Let Y_{ik} (i=1-m) represent the number of the cells belonging to initial land-cover category i in mesh-area k (k=1-l), and let X_{jk} (j=1-n) represent the number of the pixels belonging to the cluster C_j in the same mesh-area k.

Now, assuming that Y_i is expressed by a linear function of X_j (equation (1) below), we can relate clusters with any one of the categories based on t-values calculated from partial regression coefficients. While, the partial regression coefficients largely depend on the variance and covariance for the independent and dependent variables selected. Therefore, assuming that X_j is expressed by a linear function of Y_i, any other correspondences may be obtained between the categories and clusters.

In this study, we used multiple regression analyses in a stepwise manner, in which Y_i and X_j were reciprocally used as the dependent and independent variables.

The regression equation is presented as follows:

$$Y = AX + E_1 \qquad (1)$$

When Y is replaced by X, the equation is presented as follows:

$$X = BY + E_2 \qquad (2)$$

where $Y=[Y_1 \cdot\cdot Y_i \cdot\cdot Y_m]^T$, $X=[X_1 \cdot\cdot X_j \cdot\cdot X_n]^T$

A,B=partial regression coefficient matrices ($m \times n$, $n \times m$)

E_1, E_2=error matrices ($m \times 1$, $n \times 1$)

The F-ratio for equation (1) is expressed as follows:

$$F_i = V_{Ri}/V_{Ei} \qquad (3)$$

where V_{Ri}=regression sum of squares, and
V_{Ei}=residual sum of squares

In this method, the validity of determining the categories is established by testing the significance of F-ratios.

In a stepwise manner, the significance levels of each t-value are varied with the numbers of independent variables selected by fixed F-to-enter or F-to-remove limits. Therefore, in order to interprete clusters accurately to be one of the categories, it is necessary to test for the significance of t-values at a fixed level. The t-value for partial regression coefficients is described as follows:

$$t_{ij}=(a_{ij}-\alpha_{ij})/(S^{jj}*V_{Ei})^{1/2} \qquad (4)$$

where a_{ij} =partial regression coefficient of land-cover category i on cluster j
α_{ij}=true parameter
S^{jj} =j-th diagonal element of the inverse matrix of the variance

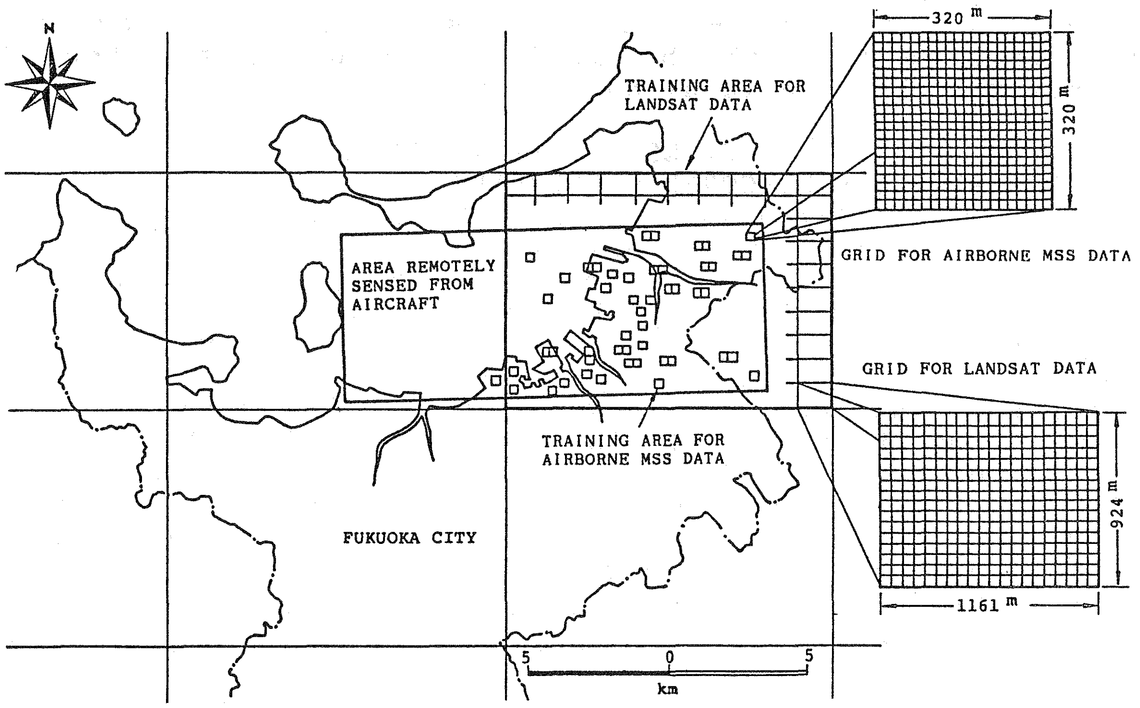
N
TRAINING AREA FOR LANDSAT DATA
AREA REMOTELY SENSED FROM AIRCRAFT
GRID FOR AIRBORNE MSS DATA
GRID FOR LANDSAT DATA
TRAINING AREA FOR AIRBORNE MSS DATA
FUKUOKA CITY
320 m
320 m
924 m
1161 m
5
0
5
km

Fig.1 Location of Training Areas and Grids

When a cluster is selected as the independent variable for different land-cover categories in equation (1), there is difficulty in interpreting its cluster to be in one of the categories. In this case, we interpret the cluster as being in the category having the largest value of t_{ji} calculated from the partial regression coefficients of cluster j on categories i from equation (2).

Based on the above statistical verification using F-ratio and t-value, initial land-cover categories are consolidated and clusters are combined step-by-step. Finally, it is possible to determine the land-cover categories and to interpret the clusters correctly.

We consolidate the categories having no corresponding clusters, into any one of the others. And we also combine the clusters having no corresponding categories, with any other clusters. When there are so-called multicollinearities among independent variables, they confuse our interpretation of clusters. Therefore, we remove those clusters showing negative correlation coefficients between the categories.

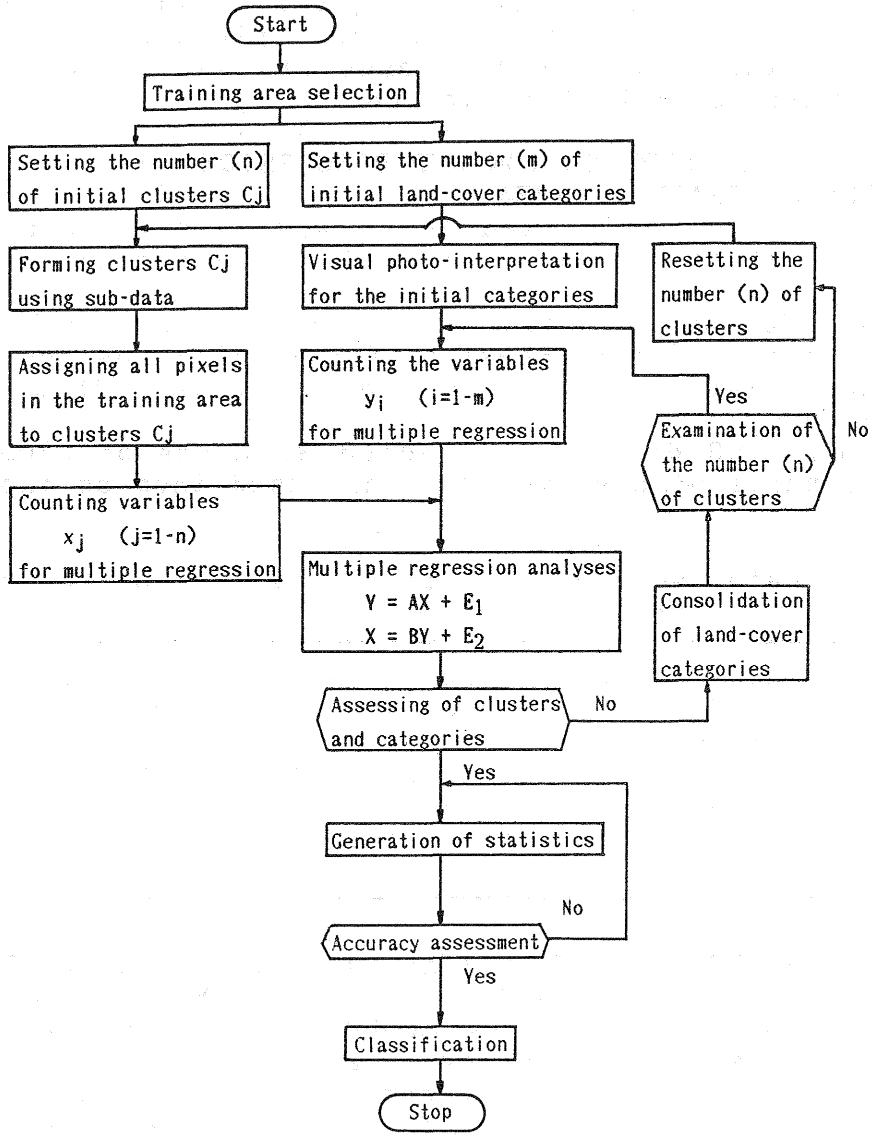
Start
Training area selection
Setting the number (n) of initial clusters Cj
Setting the number (m) of initial land-cover categories
Forming clusters Cj using sub-data
Visual photo-interpretation for the initial categories
Resetting the number (n) of clusters
Assigning all pixels in the training area to clusters Cj
Counting the variables y_i (i=1-m) for multiple regression
Yes
Examination of the number (n) of clusters
No
Counting variables x_j (j=1-n) for multiple regression
Multiple regression analyses
$Y = AX + E_1$
$X = BY + E_2$
Consolidation of land-cover categories
Assessing of clusters and categories
No
Yes
Generation of statistics
No
Accuracy assessment
Yes
Classification
Stop

Fig.2 Principal Flow of Classification

2.2 PROCEDURE OF CLASSIFICATION

The procedure of classification is shown in Fig.2 and explained as follows:

(1) Selecting training area and setting initial land-cover categories. Visually-interpreting the aerial color photographs into the initial categories. Then, counting the number of the cells photo-interpreted to be in each category i to provide the variables Y_i for the multiple regression analyses.

(2) Setting the number of initial clusters C_j. Forming the clusters by use of the sub-data consisting of only a few hundred pixels.

(3) Assigning all pixels in the training area to clusters C_j. Then, counting the number of the pixels assigned to each cluster C_j to provide the variables X_j.

(4) Performing the multiple regression analyses. Interpreting the clusters into categories according to the t-values.

(5) Consolidating those categories having no corresponding clusters to any one of the other categories in consideration of the values of t_{ji}. Re-forming those clusters which are unusually difficult to interpret.

(6) Generating the statistics for digital classification from the digital values of the pixels in the interpreted clusters. Then, finally, the overall data images are all classified by so-called a maximum likelihood method.

3. APPLICATION AND DISCUSSION

3.1 SELECTION OF TRAINING AREA

Table 1 shows the data used for classification and visual photo-interpretation. Location of training areas is shown in the above Fig.1.

For Landsat MSS and TM data, we selected an area of 9.2km× 11.6km located in the north-east of Fukuoka City as the training area. We divided it into 100 areas and used them as the above mesh-areas.

For airborne MSS data, we selected 80 areas of 320m× 320m from various parts of observed area and used them as the mesh-areas.

Table 1 Data Used for Digital Classification and Visual Interpretation

Landsat		Airborne	Aerial Color	Land-Use
MSS	TM	MSS	Photograph	Map
Dec.18 1981 Landsat-2	Oct.6 1984 Landsat-4	Aug.12 1979 Aircraft h=3200m	Nov.14,18 1981 1:10000	1975 1:25000

3.2 PHOTO-INTERPRETATION OF INITIAL LAND-COVER CATEGORIES

Considering the ground resolutions of the MSS and TM data, we initially defined twenty-six land-cover categories for the Landsat data and sixteen land-cover categories for airborn MSS data. For Landsat data, we divided each mesh-area into 320 cells of 58m× 58m using the grid shown in the above Fig.1. Then, we superimposed the grid onto the mesh-area of the aerial color photographs and visually-interpreted each cell as one of the initial land-cover categories which occupied the largest part of the cell. Table 2 shows the initial land-cover categories used for Landsat data.

In the airborne MSS data, we divided each mesh-area into 400 cells and visually-interpreting each cell as being in one of the 16 initial land-cover categories in the same way as for Landsat. Then, we obtained the variables Y_i for multiple regression analyses.

Table 2 Initial Land-Cover Categories Obtained by Visual-Photo-Interpreting (for Landsat Data)

Category		Contents
1	Water 1	sea,river,lake,pond
2	Water 2	shallow water,reef,sea bank,waterfront
3	Field 1	harvested rice field
4	Field 2	cultivated rice field
5	Orchard	orange orchard
6	Needle-leaved	coniferous trees
7	Broad-leaved	non-coniferous trees
8	Bamboo	bamboo
9	Open Land 1	play ground,sands(whitish)
10	Open Land 2	developing land,reclaimed ground(grayish)
11	Waste Land 1	waste land(brown)
12	Waste Land 2	waste land(with grass)
13	Pasture	farm field,grass
14	Wild Field	wild grass,bush
15	Pavement	covered by asphalt or concrete
16	Track	railway track,yard
17	Large Building	whitish and large building
18	Large Warehouse	grayish and large building
19	Industrial Area	industrial district
20	Urban Area 1	central business district
21	Urban Area 2	business district
22	Urban Area 3	apartments
23	Residential 1	densely-populated dist.in a city area
24	Residential 2	residential district in a city area
25	Suburban Area 1	newly developed district
26	Suburban Area 2	thinly populated dist.in a suburban area

3.3 CLUSTERING ALGORITHM

Because of computational restrictions, we selected sub-data consisting of 300 pixels from the training area and combined these into 40 clusters based on their Euclidean distances. Then, we assigned all the pixels within each training area into 40 clusters by a nearest neighbor method and obtained variables X_j by counting the number of pixels belonging to each cluster C_j in each mesh-area.

3.4 CONSOLIDATION OF LAND-COVER CATEGORIES

We consolidated the initial land-cover categories by the above procedures. Table 3 shows some of the t_{ij}-values calculated from the partial regression coefficients, which are derived from the equation (1) in the first regression step for Landsat MSS data. The t-values are statistically significant at the 0.5(%) level.

The clusters with a symbol(*) were interpreted to be in a single land-cover category. The clusters with a symbol(**) were interpreted to be in the land-cover category having the largest value of t_{ji} derived from the equation (2).

For example, see cluster X_1 in the table. Although it had significant t-values of 5.0, 4.7 and -6.4 for categories Y_6, Y_7 and Y_8 respectively, we interpreted it as the category Y_7 because its category had the largest value of t_{ji} among them. Accordingly, cluster X_8 was interpreted to be in category Y_8, leaving category Y_6 with no corresponding cluster. We consolidated this category Y_6 into category Y_8 at the next regression step, because Y_6 and Y_8 had the corresponding clusters of X_1 and X_8 in common. In this study, there was no category rejected by the F-test.

Table 3 t-values Calculated from Partial Regression Coefficients in Equation (1)

Category Number	Cluster Number X_1	X_2	X_3 ...	X_8 ...
:				
Y_6	5.0	—	—	5.9
Y_7	4.7**	4.2*	—	—
Y_8	-6.4	—	8.9*	5.2**
:				

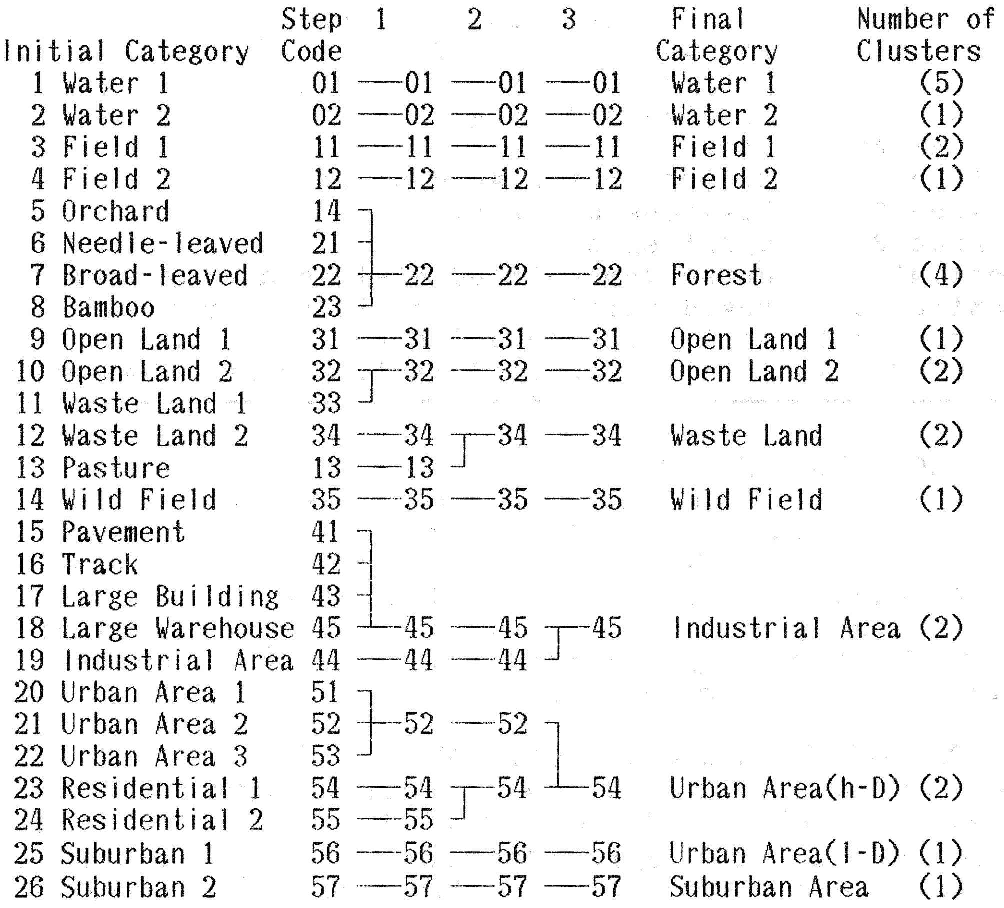
Initial Category | Step Code | 1 | 2 | 3 | Final Category | Number of Clusters
1 Water 1 01 —01 —01 —01 Water 1 (5)
2 Water 2 02 —02 —02 —02 Water 2 (1)
3 Field 1 11 —11 —11 —11 Field 1 (2)
4 Field 2 12 —12 —12 —12 Field 2 (1)
5 Orchard 14
6 Needle-leaved 21
7 Broad-leaved 22 —22 —22 —22 Forest (4)
8 Bamboo 23
9 Open Land 1 31 —31 —31 —31 Open Land 1 (1)
10 Open Land 2 32 —32 —32 —32 Open Land 2 (2)
11 Waste Land 1 33
12 Waste Land 2 34 —34 —34 —34 Waste Land (2)
13 Pasture 13 —13
14 Wild Field 35 —35 —35 —35 Wild Field (1)
15 Pavement 41
16 Track 42
17 Large Building 43
18 Large Warehouse 45 —45 —45 —45 Industrial Area (2)
19 Industrial Area 44 —44 —44
20 Urban Area 1 51
21 Urban Area 2 52 —52 —52
22 Urban Area 3 53
23 Residential 1 54 —54 —54 —54 Urban Area(h-D) (2)
24 Residential 2 55 —55
25 Suburban 1 56 —56 —56 —56 Urban Area(l-D) (1)
26 Suburban 2 57 —57 —57 —57 Suburban Area (1)

Fig.3 Consolidation of Land-Cover Categories for Landsat MSS

3.5 RESULTS AND DISCUSSION

Fig.3 shows the consolidations of land-cover categories for Landsat MSS data. Nine, two, and two land-cover categories were consolidated at the first, second, and third stages, respectively. Eventually, thirteen land-cover categories were established. In this figure, the number of the clusters interpreted to be in each of these final categories are written in parenthesis.

Table 4, Table 5, and Table 6 shows the proportions of land-cover categories within each training area, and the correlation coefficients derived from the calculations between the numbers of photo-interpreted cells and those of digital classification results within mesh-areas using Landsat MSS, TM, and airborne MSS data, respectively.

As can be seen in Table 4, there are 13 categories defined for the Landsat MSS data and their correlation coefficients are within reasonable accuracy.

As shown in Table 5, there are 13 categories defined for the Landsat TM data. Comparing these with those categories for the MSS data, "Needle-leaved trees", "Broad-leaved trees" and "Track" can be classified. However, urbanized areas are classified into two categories of "Urban areas (high density)" and "Suburban areas (low density)". Landsat MSS data gives information exclusively about spatial characteristics of ground such as high, middle or low density. While, due to higher resolution, TM data gives us more detailed information about the detailed physical characteristics of the ground. So that, the method shows some differences between the classification results by the TM data and by the visually-interpreted results of aerial photographs provided for the MSS data.

As shown in Table 6, there are 11 categories defined for the airborne MSS data and their correlation coefficients are within reasonable accuracy overall except in the case of "Open land".

Table 4 Proportions of Final Land-Covers and Correlation Coefficients (Landsat MSS)

	Category	Land-Cover(%) Photo	 MSS	Correlation Coefficient
1	Water 1	26.80	25.48	0.99
2	Water 2	3.35	1.59	0.80
3	Field 1	5.83	6.88	0.87
4	Field 2	1.92	4.74	0.82
5	Forest	8.71	8.46	0.99
6	Open Land 1	1.22	0.36	0.71
7	Open Land 2	7.87	6.03	0.61
8	Waste Land	5.51	7.21	0.87
9	Wild Field	2.42	2.52	0.72
10	Industrial Area	12.42	11.99	0.85
11	Urban Area(high-D)	13.31	11.91	0.91
12	Urban Area(middle-D)	5.96	7.87	0.80
13	Suburban Area(low-D)	4.67	4.96	0.74

Photo means a visual-interpretation of an aerial color photograph

Table 5 Proportions(%) of Final Land-Covers and Correlation Coefficients (Landsat TM)

Category		Land-Cover(%)		Correlation
		Photo	TM	Coefficient
1	Water 1	26.80	25.38	0.99
2	Water 2	3.35	0.64	0.84
3	Field	7.75	6.01	0.91
4	Needle-leaved trees	2.79	1.79	0.93
5	Broad-leaved trees	5.92	5.15	0.94
6	Open Land 1	1.22	2.80	0.65
7	Open Land 2	3.37	8.08	0.65
8	Waste Land	8.72	9.40	0.80
9	Wild Field	3.71	6.36	0.74
10	Track	0.57	0.57	0.88
11	Industrial Area	11.86	10.98	0.73
12	Urban Area(high-D)	9.44	7.50	0.93
13	Suburban Area(low-D)	14.50	14.33	0.93

Table 6 Proportions(%) of Final Land-Covers and Correlation Coefficients (Airborne MSS)

Category		Land-Cover(%)		Correlation
		Photo	MSS	Coefficient
1	Water 1	3.07	3.23	0.96
2	Water 2	2.05	2.05	0.74
3	Field	6.03	5.95	0.95
4	Forest	4.75	5.38	0.96
5	Open Land	1.43	0.49	0.68
6	Waste Land	13.52	14.14	0.72
7	Wild Field	4.76	9.10	0.84
8	Building(whitish)	2.56	1.60	0.78
9	Building(grayish)	15.40	15.37	0.79
10	Urban Area(high-D)	22.92	24.09	0.86
11	Suburban Area(low-D)	23.51	18.60	0.80

3.6 APPLICATION TO MULTI-TEMPORAL MSS DATA

We applied this classification method to Landsat MSS data observed in different years shown in Table 7. Table 7 shows 10 defined categories and their proportions within the training area. Comparing these with those categories shown in Table 4, two "Field" and two "Open land" categories are combined into one category respectively, and three "Urban areas" are re-formed into two categories, taking their stabilities into account. Table 7 shows that the "Field" and "Forests" decrease while the urbanized areas increase.

As can be seen in Table 8, they are within reasonable accuracy and stable overall. However, the correlation coefficients of "Open land" and "Waste land" are relatively low due to possible mis-classifications caused by observations of seasonal differences.

Table 7 Proportions(%) of Land-Covers Derived from Multi-Temporal MSS Data

Category	Photo	Landsat MSS Observation			
	Nov.1981	Sep.1979	Oct.1980	Nov.1981	Nov.1984
Water 1	21.81	24.48	24.65	25.21	24.73
Water 2	3.35	3.50	3.11	3.03	3.37
Field	9.04	10.93	10.91	10.41	9.77
Forest	8.71	10.07	9.45	9.30	8.87
Open Land	4.59	5.78	7.55	7.64	8.34
Waste Land	5.07	5.28	4.53	4.65	3.26
Wild Field	6.64	7.00	6.31	6.54	5.21
Industrial	11.86	8.69	8.72	8.54	9.51
Urban(high-D)	9.44	9.12	9.17	9.85	10.56
Suburban(low-D)	14.50	15.13	15.60	14.83	16.40

Table 8 Correlation Coefficients between Visual-Photo-Interpretation and MSS Digital Classification

Category	Landsat MSS observation			
	Sep.1979	Oct.1980	Nov.1981	Nov.1984
Water 1	0.96	0.99	0.99	0.99
Water 2	0.85	0.84	0.90	0.82
Paddy Field	0.95	0.97	0.97	0.96
Forest	0.98	0.98	0.99	0.99
Open Land	0.83	0.81	0.69	0.57
Waste Land	0.44	0.45	0.69	0.49
Wild Field	0.62	0.72	0.82	0.76
Industrial	0.80	0.85	0.87	0.88
Urban(high-D)	0.92	0.95	0.95	0.94
Suburban(low-D)	0.80	0.87	0.92	0.88

4. CONCLUSION

This paper is summarized as follows:

(1) The proposed method gives statistical verification of defining land-cover categories and assigning clusters into these categories.

(2) The method can be applied to multi-temporal MSS data, provided there is no notable difference in the observation times (season and year) between photo-interpretation data and MSS data.

(3) The method gives with a reasonable accuracy of r=0.99-0.61 overall, 11, 13, and 13 categories for airborne MSS, Landsat MSS, and TM data, respectively. In four sets of multi-temporal Landsat MSS data, the classification results to 8 categories did not fluctuate widely, with r=0.99-0.80, not including the results for open land and waste land.

REFERENCES

1)Townshent,J. and Justice,C: Information extracting from remotely sensed data, International journal of remote sensing, Vol.2,Num.4,pp.313-329,1981.

2)Urano,Y. et al.: Abstract of the annual covention of architectual institute of Japan,pp.421-422,1980.

3)Tsubaki,H., Okuno,T., and Yasuoka,Y.: Data Analyses in Remote Sensing with Special Reference to Applications of Regression Models, Journal of The Remote Sensing Society of Japan, Vol.3,No.4,1983