Correlation Studies between Landsat MSS Data and Population Density in Japan

Kenji Naito and Shin-ichi Hanaki

Nippon Electric Co., Ltd.
Central Research Laboratories
Kawasaki, Japan

Jun Yamamoto and Kageo Akizuki

Waseda University
Tokyo , Japan

ABSTRACT

A relation was analyzed between remote sensing-based predictors (Landsat multispectral data) and ground-based criterion (population density according to the census) by computer-aided analysis.

The data used were Landsats 1 and 2 multispectral scanner (MSS) digital tapes and grid square basis population density according to the census in Japan.

Correlation analysis showed that MSS band 5 had a positive correlation with population density, while band 7 had a negative one.

Assuming the relationship to be linear, multiple correlation analysis was applied, and significant correlation of approximately 0.75 was found.

Furthermore, assuming the relation to be a complex and nonlinear system, a heuristic self-organization approach, known as the GMDH (Group Method of Data Handling), was applied, and more precise analysis was made for subdivided population density classes. Accuracy of population density identification from Landsat data was approximately 60 per cent through 75 per cent, according to population density values of each subdivided class.

14th Congress of the International Society of Photogrammetry, Hamburg 1980
Comission No. VII , Working group 8
Presented Paper

Correlation Studies between Landsat MSS Data and Population Density in Japan

Kenji Naito and Shin-ichi Hanaki

Nippon Electric Co., Ltd.
Central Research Laboratories
Kawasaki, Japan

Jun Yamamoto and Kageo Akizuki

Waseda University
Tokyo , Japan

ABSTRACT

A relation was analyzed between remote sensing-based predictors (Landsat multispectral data) and ground-based criterion (population density according to the census) by computer-aided analysis.

The data used were Landsats 1 and 2 multispectral scanner (MSS) digital tapes and grid square basis population density according to the census in Japan.

Correlation analysis showed that MSS band 5 had a positive correlation with population density, while band 7 had a negative one.

Assuming the relationship to be linear, multiple correlation analysis was applied, and significant correlation of approximately 0.75 was found.

Furthermore, assuming the relation to be a complex and nonlinear system, a heuristic self-organization approach, known as the GMDH (Group Method of Data Handling), was applied, and more precise analysis was made for subdivided population density classes. Accuracy of population density identification from Landsat data was approximately 60 per cent through 75 per cent, according to population density values of each subdivided class.

I. INTRODUCTION

Collection and arrangement of current environmental information is necessary in each country, because environmental monitoring and earth resources conservation have become an important problem on worldwide basis. Remote sening from a high altitude is one useful method to solve this problem, since environmental information is acquired simultaneously and periodically over a large area. To utilize these remotely sensed data as effective information, the relation between ground truth data and the sensed data should be analyzed.

In this paper, population density data were taken as one ground truth data. A census is taken every five years in Japan. Population density data are available from computer magnetic tape memory within a few months after the census.

High estimation probability for determining population density from Landsat imagery data had been shown for rather roughly divided population density classes.(1) A more precise relationship was analyzed. Identification models are discussed using several different methods. Experimental results show how a population distribution pattern is reflected on Landsat imagery.

Preprocessing of data used will be described in Section 2, and analytical results will be discussed in Sections 3, 4 and 5.

II. DATA PREPROCESSING

The study was concentrated on a test site in the middle part of Japan, namely Tokai and Kinki districts, as shown in Fig. 1. A description of data used here is shown in Table 1. Data from Landsats 1 and 2 computer compatible tapes (CCTs) were displayed on a color image display. Pertinent ground control points (GCPs) were selected on the display using a computer control tablet pen, referring to topographical maps of a scale of 1:50,000. GCP image coordinates were calculated and appeared on the display. GCP map coordinates were read out using a computer controlled digitizer. Warp function coefficients between Landsat and map coordinates were calculated by the least square method. Significant correlation of greater than 0.9 between bands 4 and 5, and also bands 6 and 7, is generally obtained. Thus two bands were handled for data reduction. Then, only bands 5 and 7 were geometrically corrected using nearest neighbor resampling method to register with the maps. They were averaged over two or four hundred pixels to form a pixel representing a 500 meter square or 1 kilometer square area.

Population density data were extracted from "statistics on grid square basis", based on data from the 1970 and 1975 Japanese census results, which are available in the form of data stored on CCTs in the computer memory(2) . The earth's surface is divided into the grid squares at every constant latitude and longitude. Four grid sizes are used. For example, corresponding latitude, longitude and actual distance on the earth's surface between grid lines in the largest grid size are 60', 40' and 80 km by 80 km, respectively.

As population density data are recorded according to the ascending order of longitude and latitude values, they were rearranged in the same order as Landsat data, augmented into a rectangular image with dummy data padding. They were classified into 64 or 128 classes at constant intervals, and into 16 classes at logarithmic intervals, as shown in Table 2. Preprocessed data on the image display are shown in Figs. 2 and 3.

III. CORRELATION ANALYSIS

A matrix of correlation coefficients relating logarithms of population density with MSS bands 5, 7 and their combination was computed. Two dimensional histogram maps are shown in Fig. 4 for the Tokai district.

Band 5 reveals a positive correlation, while band 7 gives a negative one. Considering only band 5, the band data reflects volume of artificial structures, such as houses and roads, which are concentrated in urbanized areas, and is usually proportional to population density. Band 7 usually is sensitive to vegetation green and data reflecting band 7 data represents vegetation amounts. Healthy plants tend to be few in urbanized areas. Bands 5, 7 and their combination image has a fairly high correlation value of approximately 0.6 through 0.75 with greater than 100 persons, 3,000 persons and 100 persons population per 1 km square, respectively. In later analysis, population density of over 100 persons will particularly be discussed in a logarithmic case.

IV. IDENTIFICATION BY MULTIPLE REGRESSIVE ANALYSIS

Three multiple regressive models were shown in Table 3, where y is population data per 1 km^2 as object variable, x_1 is MSS band 5 data per 1 km^2 as exposition variable, x_2 is MSS band 7 data per 1 km^2 as exposition variable, and a_i is the coefficient. The coefficient a_i values were calculated with the least square method. The results are shown in Table 4. The multiple correlation coefficient is one of the indexes which shows an exact multiple regressive model, the correlation coefficient between observation y and calculation y^* by multiple regressive model.

Population density classes 7 through 15 were estimated with the MSS bands 5 and 7 on the test scene using multiple regressive model.

For each class, the population figures logarithmically increase. However, it seems sufficient to estimate to the data in thousands, so five classes were classified. Multiple regressive models matching ratio, estimated population density to actual one, is shown in Table 5. It shows that classes 7 through 10 more closely matched experimental and estimated data because those were classified into one class on estimating. So it seems that, if the population density class is rougher than the population density class in thousands, data will be estimated more correctly.

V. IDENTIFICATION BY GMDH

GMDH is one of the methods with which the essentially complex system is treated, whose trait is that: (3)

1) A complex, multivariable and nonlinear system with a few input and output data can be identified and predicted.
2) Calculation quantity is less, and the algorithm is stabler for multivariable than usual stochastic prediction.

3) It has the ability of mathematical description with optimal complex in numerical correction sense. Let the relation between input and output be nonlinear as

$$y = F(x_1, x_2, ..., x_N) \quad (1)$$

where x_i (i=1, 2, ..., N) is input data, y is output data. The identification of the relation between input and output, F, was executed by following algorithm.

GMDH Algorithm

1) The correlation coefficients were calculated between the output variable and each input variable. The larger ones were left as the "good" variables, and the smaller ones were dumped as "bad" variables.
2) Original data were divided into training and checking fields.
3) Concerning the combination of two variables on the input variables x_k (k = 1, 2, ..., N), the middle variables z_m(m= 1, 2, ..., N(N-1)/2) by equation (2)

$$Z_m = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j \quad (2)$$

where the coefficient a_n (n =0, 1, 2, 3) on this function were calculated with the least square method on the training data.

4) With the coefficients calculated on the training data, the checking data were translated on the equation (1) and the correlation coefficients were calculated concerning the checking data. So, the middle variables, the number of M (less than N), were taken in order of large size. The rest were left as is.
5) Going to step 3) as $x_i = z_i$, $x_j = z_j$, the next middle variables were taken. Steps 3) through 5) were repeated until this correlation coefficient was less than the previous correlation coefficient. Otherwise go to step 6).
6) The calculation was stopped, so the complete description was taken.

The complete description using this algorithm is shown in Table 6.

GMDH matching ratio is shown in Table 7. It showed that population density classes 7 through 10 were matched well, while the classes 14 and 15 were few matched for both districts, since the number of the classes 7 through 10 were more than that of the classes 14 and 15. The population density under 3,000 persons/km^2 by GMDH were matched as well as by multiple regressive model, however that over 3,000 persons/km^2 by GMDH were more matched than by multiple regressive model.

Population density estimation results from MSS bands 5 and 7 using the complete description are shown in Fig. 5.

The results on both districts were that;
1) the matching places were along the railroad line and principal cities. Because there are many artificial structures in these places.
2) the non-matching places were in the east of Kyoto and a factory area. In the east of Kyoto it is not known why this is so. The factory area only has artificial structures, however it actually has very low population density, as does the reclaimed land in Osaka Bay.

VI. Conclusion

The population density identification ability was investigated using the correlation between Landsat data and artificial structures.

Population density was classified logarithmically and linearly. The former is generally better matched with population density than the latter. Correlation analysis showed that MSS band 5 had a positive correlation with population density, while band 7 had a negative correlation. The combination imagery of bands 5 and 7 had a significant matching ratio value of approximately 0.75 with population of over 100 persons per 1 kilometer square. However, population density per 0.5 kilometer square had a value of approximately 0.5. So it seems that, the larger is the population density grid size, the more correctly is population density identified and predicted.

The identification by GMDH was more correctly than the identification by multiple regression analysis. In both methods, locations along the railroad line and principal cities were matched. However, there were non-matching places in the east of Kyoto, in the factory area etc.. Accuracy of population density identification from Landsat data was approximately 60 per cent through 75 per cent, according to population density values of each subdivided class.

ACKNOWLEDGMENT

The authors wish to thank Mr. Ogiwara. for his continuous encouragement all through the study. They also thank the members of the research laboratory for their helpful discussions.

REFERENCES

1. S. Murai, "Estimation of Population Density in Tokyo District from ERTS-1 Data", Proc. of 9th International Symp. on Remote Sensing of Environ., pp.13-22, April 1974.

2. "Statistics on grid square basis guide", (in Japanese) Japan Statistic Association, Feb. 1978.

3. A.G. Ivakhnenko, "Polynomial Theory of Complex System", IEEE trans. Systems, Man and Cybernetics, vol. SMC-6, no.4, pp.364-378, Oct. 1971.

Table 1. Used data description.

District	Landsat		Population density		
	Scene ID.	Date	Approx. grid size	Date	Number of pixels
Kinki	1093-01060	Oct. 24'72	$0.5_{km} \times 0.5_{km}$	Oct.'70	130x168
Tokai	2232-00473	Sept.11'75	$1_{km} \times 1_{km}$	Oct.'75	80x80

Table 2. Logarithmic population density classification.

Class	Persons/grid square	Number of pixels	
		Kinki	Tokai
1	0 - 5	76	30
2	6 - 9	4	0
3	10 - 19	11	134
4	20 - 29	10	196
5	30 - 59	23	546
6	60 - 99	48	584
7	100 - 199	324	1017
8	200 - 299	265	723
9	300 - 599	356	1232
10	600 - 999	298	902
11	1000 - 1999	476	1057
12	2000 - 2999	262	722
13	3000 - 5999	333	1073
14	6000 - 9999	143	298
15	over 10000	86	27
16	dummy	3685	13299
Total		6400	21840

Table 3. Multiple regressive models.

Order	Model	Coefficient number
1	$y = a_8x_1 + a_9x_2 + a_0$	3
2	$y = a_5x_1^2 + a_6x_2^2 + a_7x_1x_2 + a_8x_1 + a_9x_2 + a_0$	6
3	$y = a_1x_1^3 + a_2x_2^3 + a_3x_1^2x_2 + a_4x_1x_2^2 + a_5x_1^2 + a_6x_2^2 + a_7x_1x_2 + a_8x_1 + a_9x_2 + a_0$	10

where
y ; Population data per 1 km^2 as object variable
x_1; MSS band 5 per 1 km^2 as exposition variable
x_2; MSS band 7 per 1 km^2 as exposition variable
a_i; Regressive coefficient

Table 4. Multiple regressive model results.

Order	1	2	3
Correlation coefficient	0.7318	0.7580	0.7665
Coefficient			
a_1			0.000087
a_2			-0.000055
a_3			-0.001086
a_4			-0.000283
a_5		-0.00591	0.007892
a_6		-0.00055	0.004290
a_7		-0.00914	0.044402
a_8	0.22053	0.63985	-0.109422
a_9	-0.15217	0.04625	-0.454401
a_0	7.27739	1.44030	8.368433

Table 5. Multiple regressive models matching ratio.

	Population (persons/km^2)	Class	Number	Matching number			Matching ratio		
				1	2	3	1	2	3
1	100- 999	7-10	1243	951	933	960	76.5	75.6	77.2
2	1000-2999	11-12	738	399	450	429	54.1	61.0	58.1
3	3000-5999	13	333	67	58	60	20.1	17.4	18.0
4	6000-9999	14	143	13	31	31	9.1	21.7	21.7
5	over 10000	15	86	0	0	3	0	0	3.5
Total			2543	1430	1472	1483	56.2	57.8	58.3

Table 6. GMDH complete description. (a) Kinki. (b) Tokai.

(a)

Step	Model	Correlation Coefficient
1	$a = 0.1808980\ x_1 - 0.1567604\ x_2 - 0.002281376\ x_1 x_2 + 1.860199$	0.5431219
	$b = 0.004125116\ x_1^2 - 0.1804208\ x_2 - 0.00006214188\ x_1^2\ x_2 + 9.793019$	0.5237048
2	$Y = 2.278139\ a - 2.762887\ b + 0.07595255\ ab + 7.150084$	0.6119052

(b)

Step	Model	Correlation Coefficient
1	$a = 0.3691307\ x_1 + 0.009606659\ x_2^2 - 0.00058822490\ x_1 x_2^2 + 2.735891$	0.7569802
	$b = 0.01140737\ x_1^2 + 0.1250603\ x_2 - 0.0004580773\ x_1^2 x_2 + 5.910581$	0.7268763
2	$y = 1.562449\ a - 6.753752\ b + 0.2829801\ ab + 33.28827$	0.8296094

Table 7. GMDH matching ratio. (a) Kinki. (b) Tokai.

(a)

	Population (persons/km^2)	class	Number	Matching number	Matching ratio(%)
1	100- 999	7-10	3874	2817	72.7
2	1000-2999	11-12	1779	971	54.9
3	3000-5999	13	1073	2	0.1
4	6000-9999	14	298	0	0
5	over 10000	15	27	0	0
Total			7051	3790	53.6

(b)

	Population (persons/km^2)	class	Number	Matching number	Matching ratio(%)
1	100- 999	7-10	1243	882	71.0
2	1000-2999	11-12	738	442	60.1
3	3000-5999	13	333	77	23.1
4	6000-9999	14	143	33	23.1
5	over 10000	15	86	3	3.5
Total			2543	1437	56.5

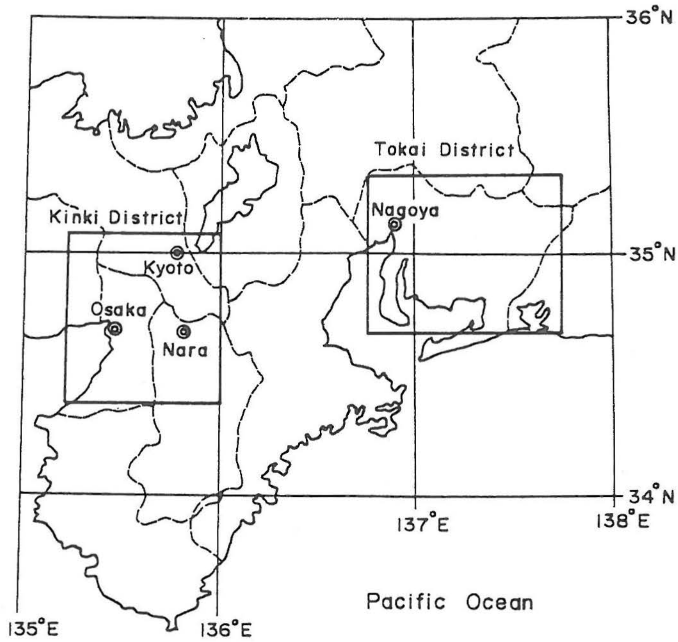
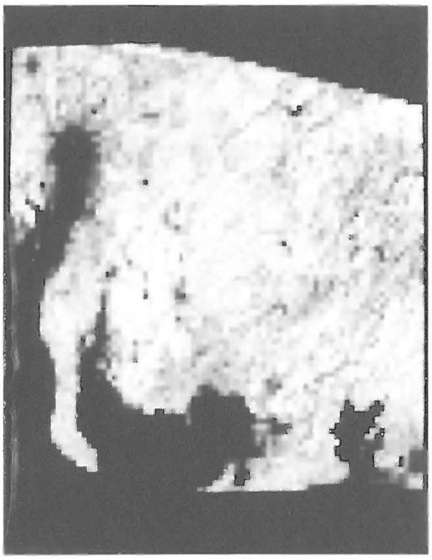


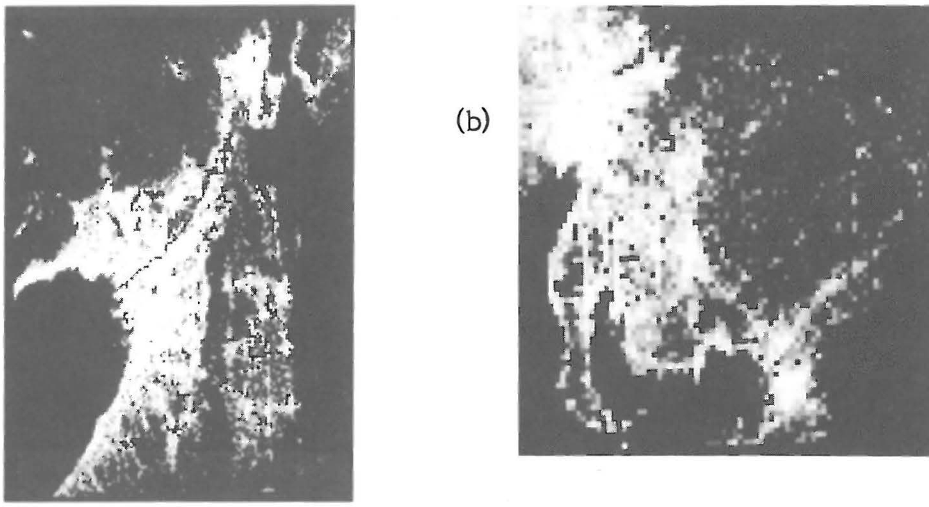
Fig. 1. Test sites covering Kinki and Tokai districts in Japan.

(a)



(b)

Fig. 2. Test site preprocessed MSS imagery. (a) Kinki band 7. (b) Tokai band 7.



(a) (b)

Fig. 3. Population density imagery. (a) Kinki. (b) Tokai.

Band 5

```
             1111111111222222222233333333334444444444555
    1234567890123456789012345678901234567890123456789012
   +---------+---------+---------+---------+---------+--
  1|        244422211111111 11 1 11 1  11        1 1
  2|             1   1       1               1
  3|         1 1 1 1    11        1  1
  4|         11111      1 1
  5|         1111  11 1 12111  1     1
  6|        124321 1 1   12 1 111      1
  7|1       28FKI86354222332232221211 1  1          1
  8|         2589765564442553441211111 11
  9|         12577543954479CA67242312111  1
 10+         1 12222343269BBA89844312 1 1 1  1
 11|         1 1 111234467DIJLIBB75441 2  11    1       1
 12|              1  1112368AABB95523 22 11       1
 13|           1       131257DACCE9964222111   1
 14|                    1  123768585221 11 1    1    1
 15|                          11498512     1
 16|Z2122212ZZZZZZZTPLIKEICFACIBA764323111111 11    11
```

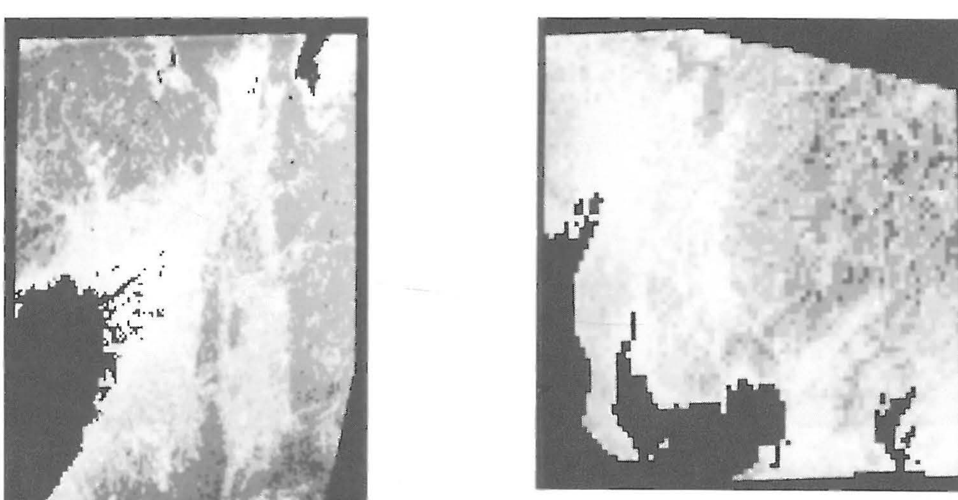
(a)

Band 7

```
             111111111122222222223
    123456789012345678901234567890
   +---------+---------+---------+
  1|11111 1 11 1111213434231
  2|                 1  11
  3|     1      11    112   1
  4|1              1  11111
  5|  1   1  1   1 1 11122 111
  6|        1   1   23331212 1
  7|1  1  111 11 1136BFJIHA5221
  8|11   1  11 1 11227ADDHC62211
  9|1 11111 11211121569EGJGB6531
 10+1    111  1112211237BGDGC8431
 11|1  1121 111  21337FGMXMGA431
 12| 1 11     11 12239DCGF9521
 13|1    1  11211238BIIJE9511
 14|        1155245677621  1
 15|        1388343111
 16|Z9C94A7649485CFOZZZZZZZWE8231
```

(b)

Fig. 4. Two dimensional histrogram for Tokai district.
(a) band 5. (b) band 7.



(a) (b)

Fig. 5. Population density estimation results.
(a) Kinki. (b) Tokai.