THE USE OF SPOT DATA FOR ESTABLISHING A CORRELATION OF SATELLITE IMAGERY AND INCOME LEVELS IN MEXICO CITY

Elizabeth A. Dudley-Murphy, Project Scientist,
Merrill K. Ridd, Director, Don H. Card, Project Scientist

University of Utah
ISPRS #1119

ABSTRACT: Developing nations are growing at an unprecedented rate, with the increasing concentration of the population in major cities. Given the rapid changes in urban environments, the need for urban data economically collected on a repetitive, timely basis is essential.

The ability of the French HRV (high-resolution visible) SPOT (System Probatoire pour l'Observation de la Terre) satellite multispectral data for differentiating socioeconomic patterns across Mexico City is evaluated. Binary discriminant analysis (BDA), a statistical technique introduced by Alan Strahler is used to show the association between classified SPOT digital data and Mexico City income data.

The analysis consists of the following steps: (1) Image processing of SPOT data; (2) Digitizing the income data and rectifying it to the same map coordinate system as the image; (3) analyzing both data sets using Binary discriminant Analysis (BDA).

A chi Square testing shows a strong correlation between spectral clusters and income levels. A relationship exists between the spectral reflectance of different cover surfaces and socioeconomic variables.

The analysis possesses certain limitations due to the general nature of the income data available. These are pointed out in the conclusion.

KEY WORDS: Mexico City, SPOT, Socioeconomic, Urban, Income

INTRODUCTION

Developing countries today face an unprecedented rate of urban expansion. Statistics published by the United Nations estimate that by the year 2025, 65% of the worlds population will live in urban areas, compared to approximately 41% at the present. "Some of the large cities in developing nations are so crowded and polluted that it appears they have reached the limit of the carrying capacity of their environment" (Vining, 1985). Conventional tools such as aerial photography, are no longer adequate for monitoring the rate and direction of urban growth. Tools that systematically and objectively monitor growth can improve planning for urban expansion. Satellite imagery provides continuous, current, regional data that can be used for mapping urban patterns and urban growth. Further, the digital, multispectral format provides for a uniform and objective foundation on which to base the urban area analysis.

The hypothesis of this research is that satellite multispectral data can be used as a surrogate for detecting socioeconomic patterns within an urban environment. Statistical correlations between Spot spectral reflectance data and income levels are explored. It is anticipated that analysis will show a relationship between amount of vegetation, soil, and impervious surfaces on the one hand and socioeconomic variables on the other. To the extent this technique can be shown to be effective, then this procedure may prove to be a valuable tool for urban analysis and planning in both developing and developed countries.

AREA OF STUDY

Mexico City, Mexico was selected as the study site for the research. Due to its rapid growth and the heterogeneous composition of land cover, this city manifests itself as an ideal test site for evaluating the characteristics of Spot data in an urban setting.

A metropolis covering more than 1000 square kilometers, Mexico City is located in the Basin of Mexico. In the west and southwest there are evergreen-covered mountain ranges; to the north, industrial areas give way to rich agricultural land; to the east, marshland and the salty Lake Texcoco, and to the south, Xochimilco and its "Floating Gardens".

The population of the metropolitan zone of Mexico City has doubled in the last 14 years. It had in 1984 approximately 17 million people or about 20% of the national population. The growth has been extensive and uncontrolled to the point that it can no longer be monitored by conventional means based on aerial photography. Although aerial photography has been obtained for the perimeter of the metropolitan area each year, urban growth is so rapid and widespread, interpreters analyzing the data are unable to complete the entire data set before it is time to gather information once again (Personal contact with planners in Mexico City, 1987).

The problem of urban expansion in Mexico City is complex and extensive. It is occurring without official land use development or planning, exerting substantial pressure on the urban environment. The results include substandard housing, unimproved water and sewage facilities, and degradation of surrounding landscapes.

In order to monitor this expansion, planners and policy makers must coordinate efforts and obtain accurate monitoring and planning capabilities. Tools are needed to provide systematic and current information to measure the changing conditions and provide a continuous image of the city. Satellite data in combination with collateral data may be the most reasonable method for obtaining information on a regional scale.

DATA SOURCES

The data used for this investigation include a Spot image of Mexico City collected March 22, 1986, an income map of the metropolitan area created by BIMSA (a Spanish acronym for Bureau of Marketing Investigations), orthophotos at a scale of 1:20,000; and a map of the city at a scale of 1:50,000.

The Spot imagery was purchased from Spot in Reston Va., with funding provided by the U.S. MAB (Man and the Biosphere) Urban Ecosystem Directorate. The maps and orthophotos were made available through the Center for Remote Sensing and Cartography at the University of Utah Research Institute.

The income data was obtained from the marketing agency BIMSA in Mexico City. Bimsa conducted a study to provide current, objective data of the socioeconomic make-up of the city for estimating population growth and for analysis leading to a better understanding the citys' dynamics (Bimsa, 1986). To obtain the most systematic evaluation of the entire zone, it was divided into two areas: the Federal District and surrounding municipalities. The measure by which income levels were determined include the following:

1. Population data from the ninth and tenth census (Bimsa, 1986)
2. Average family income based on a study conducted by Bimsa in 1984 and 1985 that employed door to door polls obtaining information regarding types of products used per household, and residential land-use classification (Bimsa, 1986).
3. Visual observation of each area included:
   a. number of cars per household
   b. T.V. antennae
   c. public transportation into the area
   d. type of commercial buildings in the vicinity (shops, banks, etc.)
   e. landscaping; individual household and neighborhood in general

The desire of the investigators to represent their findings graphically led to the creation of a color coded map of the city divided by income levels. This is useful for giving a general idea of the socioeconomic pattern of the city and provided a source for visual comparison. The index used to make the divisions was minimum wage; the levels are defined as the following:

Income class 1	30 times minimum wage
Income class 2	20-30 times minimum wage
Income class 3	10-20 times minimum wage
Income class 4	5-10 times minimum wage
Income class 5	2- 5 times minimum wage
Income class 6	1- 2 times minimum wage
Income class 7	minimum
Income class 8	1- 3 times minimum wage (Rural)

ANALYTICAL PROCEDURE

The methodology for the study consisted of three steps: (1) image processing of SPOT data, (2) digitizing the BIMSA map and rectifying it to the same reference coordinate system as the image, (3) analyze both data sets using Binary Discriminant Analysis (BDA).

Image Processing of Spot Data

The digital data collected by Spot in March 1986, were reformatted and down loaded from computer compatible tapes onto the multispectral image-processing system ERDAS (Earth Resource Digital Analysis Systems). A study area was defined from the Spot scene that encompassed the Mexico City metropolitan area with, adjoining cities in the east. The extraction of thematic information from the digital data was performed using the unsupervised clustering algorithm CLUSTR, which uses the statistical properties of the digital data and produces spectral signatures that are then grouped into clusters or "classes". This algorithm allows the user to set the following parameters: " (1) maximum number of clusters, (2) minimum distance between clusters, (3) maximum allowable cluster radius, (4) number of points until merger" (ERDAS manual). Operating in a two pass sequence; the program reads the entire data set grouping the clusters by their vector means; the second pass reads the entire data set and applies a minimum distance classifier identifying the means computed in the first pass. The output is a GIS (Geographic Information System) file containing clusters or "classes" suitable for further analysis. Based on a priori knowledge of the study area, the interpreter can identify these classes by land cover composition. Eighty spectral classes were extracted from the three bands of Spot data. CLUSTR (ERDAS' name for the algorithm) also allows the user to specify a skip factor, which determines how finely the data set will be sampled in the first pass. In this case a skip factor of 1 was specified, so that every pixel was sampled.

The classified Spot data were rectified to a reference map, in this case, one oriented to the Universal Transfer Mercator (UTM) grid coordinate system. Using a 1:50,000 scale map of the city and orthophotos at a 1:20,000 scale, ground control points were identified and then located on the image. Rectify, a geometric correction program was used to resample the data, allowing for image to map registration.

Income Data Classification

In order to register the income map to the Spot image, the data had to be input into the Erdas system. Due to the large size of the map, the data were transferred to mylar, then digitized and entered into the processing system. To make the maps comparable, it was then rectified to the same UTM coordinate system and scale as the Spot data, and color-coded to approximate the Bimsa map.

BINARY DISCRIMINANT ANALYSIS

A method was sought that would measure the level of statistical correlation between spectral data (a continuous variable) and income levels (discrete variable). Binary discriminant analysis (BDA) was used to compare the two data sets. This method identifies binary variables and their common trends, which are most important for discriminating between groups, (Strahler, 1978). Strahler names this technique binary discriminant analysis in that structurally it resembles discriminant analysis and it operates on presence or absence (binary data). He applies this technique to the presence or absence of vegetation on environmental sites. In this study BDA is used to compare two data sets, the SPOT spectrally derived data and the digitized income map data.

To illustrate this method, Strahler cites a vegetation study where BDA is used on lists of species to "reveal similar patterns of preference or avoidance among species responding significantly to a multistate environmental parameters such as soil type, rock type, or aspect" (Strahler, 1978). The procedural nature of BDA is easily understood and seems to be a natural for the study presented here. The intent of this investigation is to assess which classified spectral clusters associate with the different income levels, if any, and to identify discriminating variables. From the satellite is created an image based on reflectance from landcover composition. It is these environmental characteristics that will be the discriminating factors between the income classes.

The procedure is made up of two steps:
1. Construction of a contingency table for each combination of the two variables. In Strahler's study the two variables were vegetation species and site factor, in which species were recorded as present or absent within plots. In the case of this research the contingency table recorded the presence or absence of clusters within income levels.
2. Application of Principal Components Analysis (PCA) which defines orthogonal axes that best separate groups, and places variables on a continuum according to their principal component scores on each of the group separating components. By applying PCA on the income data it shows if there are significant groupings among the classes and if these groupings are associated with a principal component. (Strahler presents two approaches for this step, principal components analysis and factor analysis. Factor analysis is not used in this study.)

Binary Discriminant Analysis Procedure

A contingency table was created using the mean for all three bands for each of the 80 classes. The means were input into CO-OCCUR, a program written by Dr. David Wilkie, now at Harvard University, that compares two GIS files, a base (ground truth) map and a test (spectrally classified) map. In this case the base map was the income GIS file, corresponding to the columns in the matrix and the test map was the 80 class spectral file, corresponding to the rows of the matrix. The numbers in the table correspond to the number of pixels per cluster and their distribution among the income levels (Table 1). On a one-to-one basis, one can infer an association between income and cluster based on the frequency of pixels per income level. In cluster one the highest number of pixels falls within income level five (51,532 pixels), indicating that cluster one is more strongly associated with income level five than any other. Underlined numbers represent modal income level for each spectral cluster. Note that more cluster modes fall within levels five and six than any others. This is comparable to the large areas represented by these two income levels on the BIMSA income map. There are eight clusters in the table that register no pixels, that is they have zeros in all of the cells. These correspond to clusters that lie outside the study area and are deleted from further analysis.

CONTAB a program written by Don Card, was employed to compute Chi Square and Haberman's scores. Chi Square tests independence between cluster and income level for all combination of clusters and income levels. That is, it tests the statistical independence of each cell of the contingency table. At the .05 significance level, there appears to be a substantial relationship between spectral classes and income levels.

Having established the relationship between spectral clusters and income levels from the analysis of the contingency table and the chi square table, the cells of the

Income Levels

Cluster	1	2	3	4	5	6	7	8
1	2206	2198	17350	11539	51532	23071	1683	32
2	1067	794	4146	2474	7926	4797	552	7
3	156	65	219	135	899	1720	604	11
4	4145	2505	10260	6917	21228	12544	944	36
5	1749	908	4487	2133	9235	12628	2273	51
6	290	26	187	67	202	137	14	0
7	3368	1716	8539	6389	22408	21423	2347	107
8	82	29	103	47	168	110	4	0
9	25	18	91	37	765	1319	401	5
10	131	74	479	349	5814	7509	3264	20
11	93	7	35	11	49	31	1	0
12	262	171	992	644	7684	12373	8918	28
13	1	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0
15	2829	556	2103	908	2617	1602	198	16
16	4707	1490	4885	1965	4856	2776	276	22
17	2126	637	2663	1305	4756	3897	508	25
18	309	134	447	257	1095	813	102	6
19	14	5	54	31	550	834	181	6
20	37	3	22	14	42	32	1	0
21	219	83	264	159	882	822	170	7
22	0	0	0	0	0	0	0	0
23	13	6	83	52	445	526	137	3
24	698	140	496	287	877	590	71	4
25	46	29	124	78	338	178	15	0
26	0	0	0	0	0	0	0	0
27	5024	980	3352	1117	2998	1565	207	23
28	7	1	18	7	42	17	3	0
29	61	1	21	9	36	102	46	0
30	0	0	0	0	0	0	0	0
31	0	0	0	0	2	0	0	1
32	34	29	463	350	1593	554	40	1
33	773	477	3675	2995	21485	33647	21063	63
34	1	1	0	0	7	7	10	0
35	0	0	0	0	6	1	0	1
36	0	0	0	0	0	0	0	0
37	0	0	0	0	0	0	0	0
38	1934	1483	12006	14433	75739	86748	18224	130
39	0	0	0	0	0	0	0	0
40	0	0	0	0	0	1	3	0
41	0	0	0	0	3	0	0	0
42	0	0	0	0	0	0	0	0
43	9	6	103	26	161	151	13	2
44	0	0	0	0	0	0	1	0
45	1330	279	872	507	653	435	64	5
46	23	2	75	30	390	317	15	8
47	5	5	56	31	335	371	46	2
48	2020	3285	24147	27397	133007	75307	6171	130
49	0	0	1	4	6	8	0	0
50	23	6	29	19	94	162	39	1
51	4	1	2	0	13	28	5	0
52	0	0	17	8	13	0	13	0
53	21	5	35	26	35	18	6	0
54	5	0	5	9	4	1	8	0
55	17	2	14	14	9	1	0	0
56	40	12	218	167	1227	1385	289	6
57	4	1	37	24	113	23	0	6
58	0	0	7	4	0	0	1	0
59	0	0	0	13	0	0	9	0
60	0	0	0	0	3	0	0	0
61	185	101	838	695	6058	4461	2858	40
62	4	13	72	36	248	142	10	0
63	183	376	1888	934	6121	2309	146	1
64	44	2	7	10	10	3	0	0
65	262	33	81	16	48	10	4	1
66	2381	467	1243	373	874	466	39	10
67	212	175	748	479	3742	5764	1127	23
68	252	26	154	33	76	22	6	0
69	439	95	375	212	251	252	35	1
70	185	96	1268	1193	6519	2903	1066	40
71	26	4	132	92	711	857	316	6
72	102	45	437	302	2717	2523	861	42
73	0	0	14	1	49	9	3	0
74	203	22	150	62	99	38	8	0
75	24	13	145	110	758	536	127	4
76	3	1	2	0	7	1	1	0
77	2683	392	1465	373	874	425	50	3
78	5	1	5	22	49	78	23	0
79	2266	294	1229	401	1046	588	70	5
80	3949	1608	5563	3262	9049	5630	547	55

Table 1. Contingency Table Based on Co-Occur Program

matrix are converted from frequency counts to standardized residuals. The values must be standardized because the values in the contingency table are discrete and must be converted to continuous variables. Using Habermans method of conversion for each cell, the standardized residual di,j is:

$$di,j = \frac{a_{ij} - \frac{r_i c_j}{n}}{\sqrt{[\frac{r_i c_j}{n}\,(1 - \frac{r_i}{n})\,(1 - \frac{c_j}{n})]}}$$

where i and j refer to the row and column of the table, and ai,j is the observed frequency in the i,j cell; ri and cj refer to row and column totals respectively and n refers to the total number of counts in the entire table (Strahler, 1978).The matrix of Haberman Scores is displayed in Table 2. The residual corresponds to the excess of deficit from what would be expected by chance. For example, the observed frequency for row 5 and column 2 of Table 1 is 908. Row and column sums (35610 x 21934) are multiplied and divided by the total count in the entire table (1,184,062). Using these figures to compute the denominator, the square root of the result is divided into the numerator and the result is the residual, in this case 9.910. A positive number indicates a tendency to occur more often than would be expected by chance, conversely, a negative number indicates a tendency to occur less often than expected by chance (Table 2). The higher the positive value in the cell, the greater the probability of a link between the two variables.

HABERMANS SCORES

Cluster	1	2	3	4	5	6	7	8
1	-41.28	0.55	56.83	28.33	62.56	-70.77	-73.59	-7.08
2	3.75	18.27	40.87	17.50	-3.30	-25.68	-25.11	-2.83
3	-1.48	-1.48	-10.46	-11.00	-19.31	17.83	20.98	3.98
4	32.78	41.91	56.11	33.42	-5.95	-45.88	-50.92	-2.25
5	7.16	9.91	16.25	-12.52	-38.72	29.50	-0.43	3.91
6	39.99	1.89	9.49	-1.01	-9.72	-10.01	-6.37	-0.91
7	9.23	12.41	20.16	14.78	-19.21	13.83	-33.92	6.54
8	11.23	5.18	5.37	-0.19	-4.53	-6.02	-5.90	-0.73
9	-8.90	-4.92	-12.38	-13.03	-10.35	20.33	16.26	1.65
10	-24.24	-15.12	-34.91	-30.78	-14.90	33.49	60.56	1.00
11	26.96	1.25	2.39	-1.80	-4.84	-5.29	-3.78	-0.45
12	-31.10	-18.24	-43.19	-39.91	-47.49	36.57	154.46	0.04
13	1.18	-0.39	-0.95	-0.82	-2.10	-1.79	-0.74	-0.08
15	111.81	24.38	30.56	1.29	-27.58	-33.77	-20.34	2.13
16	130.86	55.50	61.99	7.56	-40.67	-51.45	-31.19	0.91
17	55.37	18.72	25.08	0.08	-20.08	-15.52	-18.21	2.90
18	14.24	8.95	5.84	-0.51	-4.47	-6.21	-8.27	1.84
19	-7.51	-5.16	-10.51	-10.01	-6.22	14.68	5.41	3.47
20	11.34	-0.12	1.17	0.18	-3.11	-2.92	-3.14	-0.38
21	9.19	4.07	-1.64	-4.51	-5.74	-0.09	-1.23	2.93
23	-6.64	-4.33	-6.31	-6.59	-6.10	4.30	3.66	1.46
24	48.02	9.88	8.94	1.64	-11.77	-14.28	-10.30	0.69
25	1.20	2.87	3.34	0.75	0.49	-6.32	-5.99	-0.89
27	176.04	41.17	47.58	-2.84	-43.31	-51.67	-26.26	2.73
28	1.03	-0.78	2.00	-0.66	0.20	-3.26	-1.67	-0.31
29	13.61	-2.01	-1.95	-3.21	-8.88	1.64	5.98	-0.51
31	-0.36	-0.24	-0.58	-0.50	1.12	-1.09	-0.45	19.87
32	-9.49	-4.51	6.21	4.97	12.13	-16.72	-12.81	-1.14
33	-51.28	-30.36	-61.27	-50.90	-74.77	63.33	217.32	-1.45
34	-0.50	0.33	-2.09	-1.81	-2.30	-1.46	4.89	-0.18
35	-0.66	-0.43	-1.06	-0.92	1.61	-1.30	-0.83	10.82
38	-85.81	-45.66	-80.01	-23.47	-14.46	124.12	38.44	-4.54
40	-0.42	-0.28	-0.67	-0.58	-1.49	-0.16	5.59	-0.06
41	-0.36	-0.24	-0.58	-0.50	2.33	-1.09	-0.45	-0.05
43	-3.39	-1.64	5.42	-3.31	-5.04	-2.32	-4.41	2.05
44	-0.21	-0.14	-0.33	-0.29	-0.74	-0.63	3.81	-0.03
45	89.29	23.04	23.08	10.50	-27.25	-26.15	-13.02	0.78
46	-4.27	-4.52	-5.03	-7.26	-3.96	-3.07	-7.71	6.66
47	-6.12	-3.41	-5.37	-6.03	-3.30	4.11	-2.97	1.14
48	-105.52	-31.42	-31.15	44.04	144.08	-26.40	-106.03	-8.06
49	-0.98	-0.64	-0.86	1.83	-0.82	0.82	-1.23	-0.14
50	0.61	-1.02	-3.02	-3.18	-7.54	2.29	1.40	0.92
51	1.23	0.02	-1.52	-2.11	-1.69	3.92	0.89	-0.21
52	-1.49	-0.98	5.53	2.13	-1.51	-4.51	5.54	-0.21
53	5.81	1.25	5.15	4.18	-3.44	-4.70	-1.32	-0.36
54	3.25	-0.78	1.05	4.32	-2.73	-3.18	4.28	-0.16
55	9.70	0.93	3.64	4.76	-3.13	-4.48	-1.98	-0.22
56	-9.72	-7.10	-9.07	-7.90	-4.84	10.28	2.71	1.54
57	-2.06	-1.72	2.42	1.04	3.01	-6.84	-4.17	12.55
58	-0.72	-0.48	5.56	3.32	-2.58	-2.19	0.27	-0.10
59	-1.00	-0.66	-1.60	8.76	-3.57	-3.03	6.39	-0.14
60	-0.36	-0.24	-0.58	-0.50	2.33	-1.09	-0.45	-0.05
61	-21.15	-12.89	-24.11	-19.47	-5.07	-10.80	52.27	6.47
62	-4.11	0.79	2.13	-1.21	4.11	-1.78	-4.52	-0.69
63	-15.92	8.72	16.78	-2.55	26.90	-27.64	-24.93	-3.03
64	23.45	0.50	-0.24	1.77	-4.09	-4.74	-2.29	-0.25
65	56.97	8.53	5.48	-3.38	-11.19	-12.45	-4.84	1.00
66	139.55	34.64	28.25	-4.06	-33.42	-35.14	-18.12	2.27
67	-14.80	-4.52	-16.85	-17.77	-17.54	38.76	9.87	3.59
68	47.22	4.67	13.15	-1.88	-11.39	-13.23	-5.32	-0.70
69	45.36	11.66	16.93	7.63	-17.52	-12.11	-7.22	-0.34
70	-18.84	-11.60	-8.24	-0.61	15.68	-27.93	1.84	7.43
71	-7.99	-6.38	-8.25	-7.86	-8.32	5.54	12.17	2.62
72	-13.99	-9.25	-15.59	-14.89	-8.38	1.18	13.28	12.90
73	-2.15	-1.42	1.08	-2.62	2.28	-4.57	-1.51	-0.30
74	36.31	3.28	12.14	2.36	-9.82	-12.07	-5.11	-0.71
75	-6.39	-3.82	-3.62	-3.28	3.60	-0.54	0.33	1.88
76	3.07	1.36	0.42	-1.12	0.89	-1.87	0.04	-0.11
77	152.16	25.56	34.53	-5.60	-36.53	-38.67	-18.42	-1.02
78	-1.07	-1.36	-3.42	1.96	-2.87	3.77	3.16	-0.40
79	129.42	17.27	26.37	-3.37	-30.11	-32.67	-16.86	-0.05
80	76.81	44.28	47.12	18.81	-23.52	-40.60	-33.88	5.77

Table 2. Haberman Scores; Table of Standardized Residuals

The final stage is principal components analysis, which involves inputting the matrix of Haberman scores into the program PRINCOM written by Card. The program computes means, standard deviations, correlation coefficients, eigenvalues, and eigenvectors (Table 3). " Cases" represents the 72 spectral clusters in the study area and "variables" represents the income levels. The result is three principal components with their respective eigenvalues, explaining a cumulative percentage of 87.6 of the total variation. The eigenvectors identify the loadings on the three principal components for each of the income levels. It should be stressed at this time that this study is not the typical principal components analysis performed on satellite digital data. For use with spectrally derived data, typical PCA algorithms transform an image from original spectral axes to transformed principal component axes of the spectral data. The method presented here applies PCA to the values in the contingency table, i.e., those values resulting from the crosstabulation of the two GIS files.

The upper part of Table 4 shows the correlation coefficient matrix for the three principal components with each of the income levels. By squaring those coefficients, the percentage of explained variation is expressed (lower part of the table). The first variable (income class 1) has 47.7% of its total variance explained by principal component (PC) 1, PC1 and PC2 (.477 + .421) together explain 89.8% of the total variance.

CASES 72
VARIABLES 8

	1	2	3	4	5	6	7	8
MEANS	14.0680	02.8270	03.4584	-01.5335	-06.3879	-04.2754	00.3271	04.5783
STANDARD DEVIATIONS	48.2432	15.8764	23.2492	13.3058	25.1135	26.2219	38.8973	15.8440

CORRELATION COEFFICIENT

ROW 1	01.0000	00.8499	00.7026	00.1373	-00.1571	-00.5747	-00.2662	00.0893
ROW 2	00.8499	01.0000	00.9083	00.4283	-00.3388	-00.7147	-00.4648	00.0732
ROW 3	00.7026	00.9083	01.0000	00.6381	-00.0614	-00.8527	-00.6510	-00.0304
ROW 4	00.1373	00.4283	00.6381	01.0000	00.6050	-00.6731	-00.8713	-00.2169
ROW 5	-00.5171	-00.3388	-00.0614	00.6050	01.0000	-00.2357	-00.5632	-00.2519
ROW 6	-00.5747	-00.7147	-00.8527	-00.6731	-00.2357	01.0000	00.6442	00.0432
ROW 7	-00.2666	-00.4648	-00.6510	-00.8713	-00.5632	00.6442	01.0000	00.1000
ROW 8	00.0893	00.0732	-00.0304	-00.2169	-00.2519	00.0432	00.1000	01.0000

EIGENVALUES

4.31355 2.26372 1.31196

CUMULATIVE PERCENTAGE OF EIGENVECTORS

0.47928 0.73081 0.87658

EIGENVECTORS

VECTOR 1	0.3324	0.4157	0.4549	0.3758	0.0623	-0.4241	-0.3807	-0.0549
VECTOR 2	-0.4315	-0.3045	-0.1163	0.3695	0.6378	-0.3534	-0.3114	-0.2565
VECTOR 3	-0.8298	0.0038	0.0832	0.0153	0.1199	-0.2231	-0.1295	0.6521

Table 3. Principal component scores based on PRINCOM program

	Correlation Coefficients		
Income Class	PC1	PC2	PC3
1	.690	-.649	-.095
2	.863	-.458	.004
3	.945	-.175	.095
4	.781	.556	.018
5	.129	.960	.137
6	-.881	-.053	-.255
7	-.791	-.469	-.148
8	-.114	-.386	.747

	Correlation Coefficients Squared			
Income Class	PC1	PC2	PC3	Σ
1	.477	.421	.009	.907
2	.745	.210	.000	.955
3	.839	.031	.009	.933
4	.610	.309	.000	.919
5	.017	.922	.019	.958
6	.776	.003	.065	.844
7	.626	.220	.022	.868
8	.013	.149	.558	.720

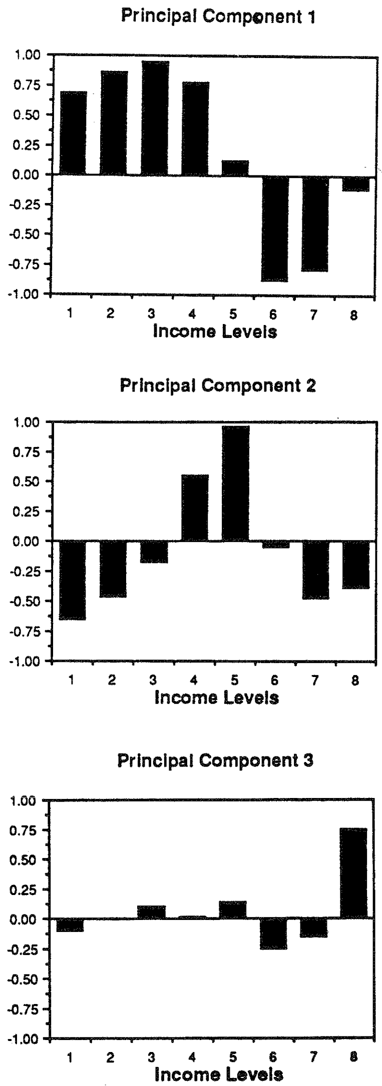
Table 4. Correlation Coefficient Matrix

RESULTS

The values in the correlation matrix in Table 4 are presented in Figure 1 in the form of bar graphs. It is clear that each of the components is characterizing a different aspect of the data. The first principal component (PC1) seems to be highly associated with the higher income levels, and negatively associated with the lower incomes. PC2 seems to characterize best the middle incomes, and PC3 seems to represent the lower incomes.

The principal component scores for each cluster were ordered per component from minimum to maximum with the corresponding spectral cluster number in each case. This was done in order to assign a gray level value to each cluster with a value of 0 representing black, and 255 representing white. This was accomplished by accessing the statistical trailer file for the classified spectral image and recoding the values three times, once for each component. Here again the images seem to be represent different characteristics across the city. In all of the images, the brighter values corresponded with the higher and middle incomes as related to the BIMSA map. It would seem then, that when comparing the bar graphs and the gray level images, that the principle components could be defining three major income divisions--low, middle and high.

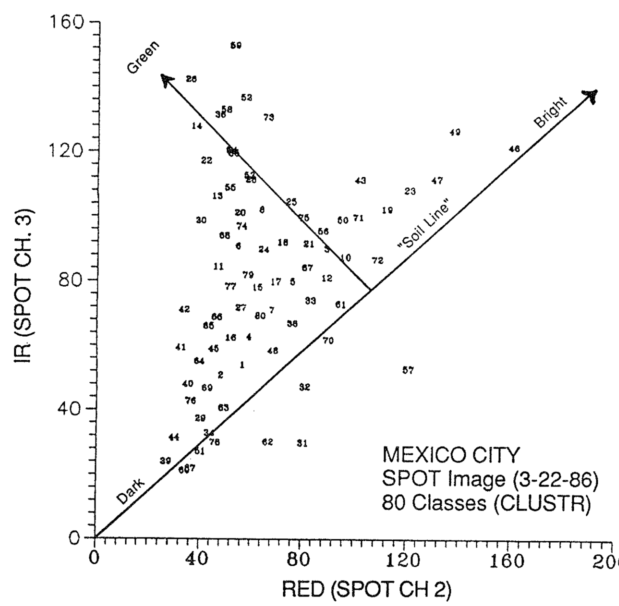
Figure 1. Representation of principal components by bar graphs



Another method of representing the data visually is to use the scatterplot, a two dimensional graph whose points represent values of two variables, in this case the variables are bands two and three. The calibrations on the graph refer to the mean spectral value for the 80 classes (Figure 2). Using a concept developed by Richardson and Weigand, the scatterplot may be interpreted as a method of distinguishing surface cover. This method shows the data in a spectral context in relationship to a perpendicular line extending from the bottom left corner to the top right, which according to the authors, represents the "soil line". The upper right corner corresponds to green signatures, and the lower left corner depicts dark signatures. All signatures between these extremes depict some combination of all three. Actual identification of the signatures is derived from gathering field data. Chung (1989), investigating the use of SPOT data in Salt Lake City, chooses to refer to the "soil line" as the "nonvegetation line" stating that this definition is more applicable to urban area studies.

On the plot there is a vegetation transition from the green point to the dark point, grading from green grass through shrubs and trees to less dense trees approaching the dark point, owing to shadowing in dense tree canopies. For the present purpose, this transition zone is referred to as the vegetation axis.

Figure 2. Scatter plot of 80 signatures derived from CLUSTR



To enhance the visual interpretation of the scatterplot, the clusters were color-coded according to the income levels, using the class modes from the contingency table with colors approximating those of the Bimsa map. Using the concept introduced by Richardson and Weigand that identifies surface cover, a correlation can be established between amount of vegetation and income level. On the scatter plot, the clusters representing the highest income levels lie along the left side of the plot consistent with the vegetation axis. This area represents a mixture of green grass, trees, and shrubs, typical of the well-established high income areas of Mexico City (verified in the field). This area is typified by large estates on large plots of land, surrounded by mature trees, shrubbery and large well-maintained lawns. Clusters representing income level three found near the green point on the scatter plot represent areas within the city of moderately high income with smaller homes on smaller plots of land. The neighborhoods have more grass than trees. Clusters corresponding more closely to the lower classes, lie along the soil line. In comparing this with the principal component bar charts, it would seem that the principal discriminating variable is amount of vegetation. Higher incomes tend to associate with more vegetated areas and lower income with lesser vegetated areas. The middle income levels are more difficult to distinguish; this is likely due to more heterogeneous surface cover, and the fact that this corresponds to a great extent with the central city.

CONCLUSION

The positive results in associating SPOT spectral data with income levels suggests the potential use of satellite data as a tool for analysis of income level distributions. Binary discriminant analysis was found to be an effective method for testing the association between Spot land-cover classes derived from unsupervised classification and income data.

The groupings of the classes based on the values of the contingency table and the principal components analysis are consistent and show a strong relationship between spectral class and family income levels.

Based on the results of this study it is anticipated that this type of analysis will be applicable to other cities, particularly in developing countries. The results from this investigation indicate that the primary predictor of socioeconomic status in an urbanized area seems to be amount of vegetation. the three principal components characterize different aspects of urban spectral reflectance. Principal component one associates highly with high income levels, and based on field observations, vegetation is the major environmental discriminating factor. Principal component two associates highly with middle income levels, possibly relating to mixed composition. Principal component three is less predictive of the income levels.

For further research using satellite data to detect socioeconomic variables in an urban setting, the use of more detailed spatial income data is recommended. Within the large spatial units of the income levels of the BIMSA map, there is a significant amount of spectral variation due to heterogeneous environmental characteristics. Income data on a block-by-block basis would very likely establish a stronger and more realistic association.

REFERENCES

Chung, J.J., 1989. MS. Spot Pixel Analysis for Urban Ecosystem Study in Salt Lake City, Utah, Department of Geography, University of Utah, Salt Lake City, Utah.

Richardson, A.J., and C.L. Weigand, 1977. Distinguishing Vegetation from Soil Background Information. Remote Sensing of the Environment. 8:307-312.

Strahler, A.H., 1978a. Binary Discriminant Analysis: A New Method for Investigating Species-Environment Relationships. Ecology. 59:108-116.

Strahler, A. H., 1978a. Response of Woody Species to Site Factors of Slope Angle, Rock Type, and Topographic Maryland as Evaluated by Binary Discriminant Analysis. Journal of Biogeography. 5:403-423.