A Geological Example of Improving Classification of Remotely Sensed Data Using Additional Variables and a Hierarchical Structure

Knut Conradsen and *Jan Gunulf*
The Institute of Mathematical Statistics and Operations Research, The Technical University of Denmark, 2800 Lyngby, Denmark

ABSTRACT: A geological example of two useful improvements of ordinary (Bayesian) discriminant analysis of Landsat MSS data is reported upon. The first improvement considers the value of augmenting the number of variables with linear (factor scores) and non-linear transformations (ratios). Second, a hierarchical procedure that utilizes *a prior* knowledge on a hierarchical structure in the populations considered is introduced. Finally, it is demonstrated that the use of the augmented set of variables with the hierarchical scheme reduced the rate of misclassification of hydrothermally altered rocks from 8.98 percent to 0.55 percent. Based on these results, it was decided to use schemes like the one described in regional mappings of color anomalous zones in central east Greenland.

INTRODUCTION

THE PRESENT WORK is a part of a major study on the applicability of remote sensing methods to mineral exploration in arctic areas. A full description is given in Conradsen *et al.* (1982) and most of the geologic results are given in Conradsen and Harpøth (1984).

The data analyzed in the present paper are Landsat MSS data. The area investigated is the environments of Malmbjerget situated south of Mestersvig in the central part of East Greenland. The test area is of size 9.3 km by 11.4 km giving a total of 23,400 pixels. The data were taken from a Landsat-2 scene, path 248 and row 9. The center of the scene is at 25.4° west longitude and 71.7° north latitude. The date of data collection was 18 September 1978.

The area is dominated by mountainous terrain with glacier-dissected alpine regions and peaks nearing 3000-m above sea level. Generally, the land surface is very well exposed. The climate is arctic, with an average yearly precipitation of 400 mm, and with average temperatures for July and January of 5°C and −20°C, respectively.

Geologically the area is a part of the Werner Bjerge Alkaline complex. At Malmbjerget there is a stockwork (porphyry) molybdenite deposit containing 150×10^6 t ore grading 0.23 percent MoS_2. The ore body is associated with a multiple intrusive alkali-granite stock. Widespread alteration is connected with the mineralizing events. The most conspicuous hydrothermal alteration is a distinct, color-anomalous zone in the quartz sericite pyrite zone. Nearest to the ore body this zone is intensely, homogeneously altered with red and yellow iron-oxide staining colors (predominantly stemming from goethite, limonite, and jarosite). In the sequel this hydrothermally altered zone is called the rust zone.

It is of obvious interest to investigate the possibility of an automatic recognition of areas similar to the rust zone. Due to the climatic conditions, the application of remote sensing techniques is/will be of great importance in mineral exploration in the arctic. A main objective is, of course, to delineate areas with a higher potential for mineralizations. The procedures described in the sequel were developed as a tool in a regional mapping of rust zones and other color anomalous areas in Central East Greenland.

THE DISTRIBUTION OF THE DATA

It is a well established fact that hydrothermal alteration zones may be enhanced on a color composite plot based on ratios between the MSS channels as well as on factor score plots (see, e.g., Rowan *et al.* (1977) or Conradsen and Harpøth (1984)). Therefore, it is quite natural to base a classification on these variables as well as on the original MSS bands.

In this presentation we use a factor analysis based on the correlation matrix. First, we give the basic statistics in Table 1.

PHOTOGRAMMETRIC ENGINEERING AND REMOTE SENSING,
Vol. 52, No. 8, August 1986, pp. 1181–1187.

0099-1112/86/5208-1181$02.25/0
©1986 American Society for Photogrammetry and Remote Sensing

TABLE 1. MEANS, STANDARD DEVIATIONS, AND CORRELATIONS FOR THE FOUR MSS BANDS BASED ON 23,400 PIXELS AROUND MALMBJERGET.

Variable	Mean	Standard deviation	Correlation B4	B5	B6	B7
B4	77.1	40.5	1.000			
B5	101.0	62.6	0.991	1.000		
B6	89.1	59.4	0.985	0.997	1.000	
B7	63.6	43.3	0.974	0.990	0.995	1.000

An unrotated principal factor analysis based on the correlation matrix gave the following formula for computing the factor scores (F1, F2, F3, F4) from the channel values (B4, B5, B6, B7):

$$\begin{bmatrix} F1 \\ F2 \\ F3 \\ F4 \end{bmatrix} = \begin{bmatrix} 0.250 & 0.252 & 0.252 & 0.251 \\ 4.482 & 0.593 & -1.345 & -3.717 \\ -6.432 & 8.844 & 6.032 & -8.528 \\ 2.655 & -14.105 & 16.474 & -4.998 \end{bmatrix} \begin{bmatrix} B4 \\ B5 \\ B6 \\ B7 \end{bmatrix}$$

The variance explained by each factor is 99.15 percent, 0.70 percent, 0.10 percent, and 0.05 percent, respectively. Thus, almost all variation is explained by the first factor, which is simply the average of the four MSS bands. Factor 2 measures the difference between the longer and shorter wave-lengths, and factor 3 measures the difference between the extreme channels B4 and B7 on one side and B5 and B6 on the other side. Finally, factor 4 gives the difference between B5 and B7 on one side and B4 and B6 on the other.

The aim is to distinguish between the rust zone and other types of pixels. In order to investigate the possibility of this, six groups of training areas were chosen. They are given in Table 2.

In Figures 1 to 3 the cumulative distributions for the six training sets of the values from channel 5 (B5), of the ratios between the values in channel 4 and channel 5 (B4/B5 = Q4/5 for short), and of the factors scores from factor 3 (F3) are given. It is immediately seen that the variables are very different with respect to distinguishing between the groups (training sets). Band 5 shows rather low values for shadow, rust, and rock and very high values for glacier and snow. The distribution of the ratio Q4/5 shows a very little difference between rock, rust, glacier, and snow, but these are very different from the shadows. Finally, the distribution of factor 3 is very similar for shadow, rock, and glacier, but these distributions deviate a lot from the distributions for snow and rust. The distribution for snow has an enormous range, whereas rust is characterized by high values of factor 3. The distributions of the remaining variables show a similar pattern.

TABLE 2. THE GROUPS IN THE TRAINING SET AND THEIR SIZES.

Type	Number of pixels
Rock	260
Rust	37
Shadow Rock	84
Shadow Ice	61
Snow	188
Glacier	246
Total, abs. & rel.	876 3.7%

THE DISCRIMINATION PROCEDURES

We now present four different procedures. The main idea in the hierarchical procedure is to utilize *a priori* information on a hierarchical structure in the groups. The augmentation of the number of variables consists of adding non-linear functions of the MSS bands (ratios) as well as linear functions (factor or principal component scores). The basic discrimination procedure is a stepwise linear discriminant procedure, where the variables are entered or deleted from the discriminant function according to the partial F-value for the variable. It is assumed that the groups considered only differ with respect to their means. Thus, the covariance matrices (dispersion matrices) are assumed to be equal from a computational point of view.

In order to investigate whether the addition of the ratios and the factors really improves a subsequent classification, a set of classification functions were computed on the basis of the four original variables B4, B5, B6, and B7 and on the basis of all variables. In the last situation the variables are linearly dependent wherefore the dispersion matrix is singular. Therefore, one must choose a subset of the variables, and then perform a stepwise discriminant analysis with the program BMDP7M (Dixon and Brown, 1979). In the stepwise analysis the variables entered were (in order) B5, Q4/6, F2, Q6/7, Q4/7, Q4/5, F3, Q5/7, F4, and Q5/6.

Based on the two sets of variables, the appropriate classification functions were determined. In Table 3 an evaluation of the classification function is given. Here the 876 pixels from the calibration areas are classified by means of the classification functions found. 1-Band corresponds to discrimination based on B4, B5, B6, and B7 whereas 1-All corresponds to a procedure based on the ten variables mentioned earlier. The results for 2-Band and 2-All correspond to procedures that will be described later. The so-called jackknifed classifications are giving classifications where the pixel classified is not used in the computation of the classification functions. If there is a great discrepancy between the ordinary and the jackknifed classifications, the results from the classifications are not reliable. This can be due to inhomogeneity within the groups (in relation to the difference between groups), or to overfitting, i.e., to the inclusion of too many variables relative to the number of observations. In the literature the term cross-validation is sometimes used instead of the present use of the term jackknifing.

From Table 3 we see that, for the classification

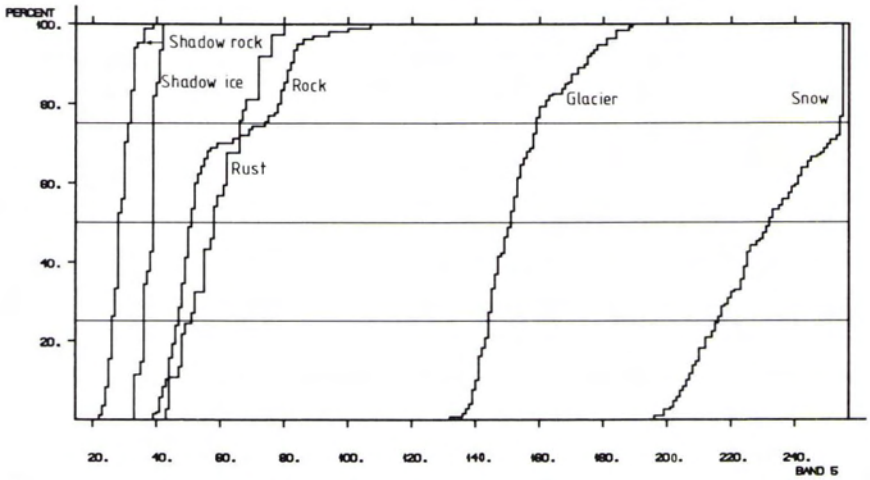

FIG. 1. The cumulative distribution functions for MSS band 5 (B5) for the six groups in the training set.

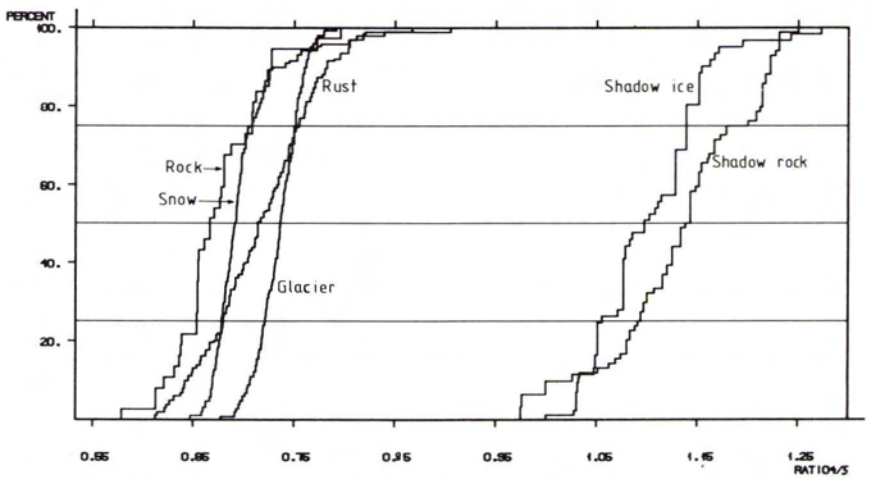

FIG. 2. The cumulative distribution functions for the ratio Q4/5 = B4/B5 for the six groups in the training set.

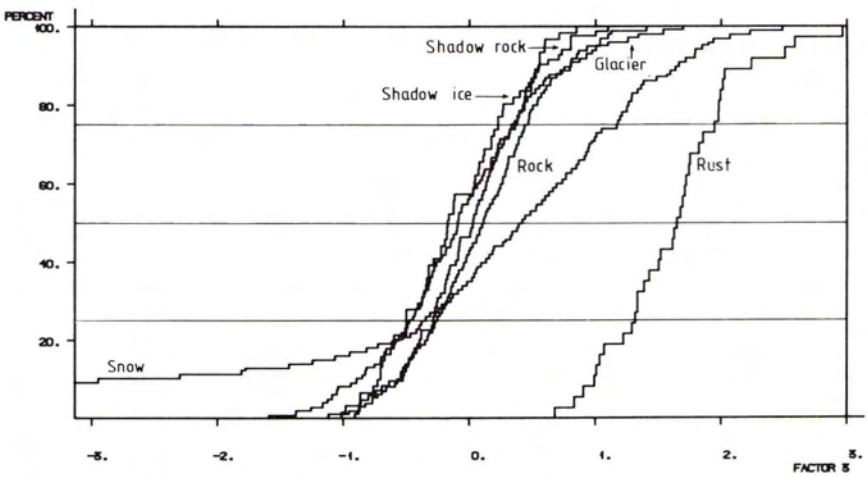

FIG. 3. The cumulative distribution functions for factor 3 (F3) for the six groups in the training set.

TABLE 3. EVALUATION OF ONE-STEP AND TWO-STEP CLASSIFICATIONS INVOLVING THE ORIGINAL VARIABLES (BAND) AND THE AUGMENTED SET OF VARIABLES (ALL, I.E., INCLUDING RATIOS AND FACTOR SCORES). THE ORDINARY AS WELL AS THE JACKKNIFED CLASSIFICATION RESULTS ARE SHOWN.

Population, method and number of observations		Percent correct		Number classified into											
				Rock		Rust		Shadow Rock		Shadow Ice		Snow		Glacier	
		Ord.	Jack.	Ord.	Jack.	Ord.	Jack.	Ord.	Jack.	Ord.	Jack.	Ord.	Jack.	Ord.	Jack.
	1-Band	86.9	86.9	226	226	34	34	0	0	0	0	0	0	0	0
Rock	1-All	96.2	95.8	250	249	10	11	0	0	0	0	0	0	0	0
	2-Band	94.2	93.8	245	244	15	16	0	0	0	0	0	0	0	0
260	2-All	98.8	98.5	257	256	3	4	0	0	0	0	0	0	0	0
	1-Band	89.2	89.2	4	4	33	33	0	0	0	0	0	0	0	0
Rust	1-All	89.2	89.2	4	4	33	33	0	0	0	0	0	0	0	0
	2-Band	97.3	94.6	1	2	36	35	0	0	0	0	0	0	0	0
37	2-All	83.8	83.8	6	6	31	31	0	0	0	0	0	0	0	0
	1-Band	89.3	89.3	0	0	0	0	75	75	9	9	0	0	0	0
Shadow	1-All	84.5	83.3	0	0	0	0	71	70	13	14	0	0	0	0
Rock	2-Band	90.5	90.5	0	0	0	0	76	76	8	8	0	0	0	0
84	2-All	92.9	91.7	0	0	0	0	78	77	6	7	0	0	0	0
	1-Band	93.4	93.4	0	0	0	0	4	4	57	57	0	0	0	0
Shadow	1-All	93.4	93.4	0	0	0	0	4	4	57	57	0	0	0	0
Ice	2-Band	100	100	0	0	0	0	0	0	61	61	0	0	0	0
61	2-All	96.7	93.4	0	0	0	0	2	4	59	57	0	0	0	0
	1-Band	100	100	0	0	0	0	0	0	0	0	188	188	0	0
Snow	1-All	99.5	99.5	0	0	0	0	0	0	0	0	187	187	1	1
	2-Band	99.5	99.5	0	0	0	0	0	0	0	0	187	187	1	1
188	2-All	99.5	99.5	0	0	0	0	0	0	0	0	187	187	1	1
	1-Band	100	100	0	0	0	0	0	0	0	0	0	0	246	246
Glacier	1-All	100	100	0	0	0	0	0	0	0	0	0	0	246	246
	2-Band	100	100	0	0	0	0	0	0	0	0	0	0	246	246
246	2-All	100	100	0	0	0	0	0	0	0	0	0	0	246	246

based on the four original variables as well as the one based on the augmented set of variables (1-Band and 1-All), there are no misclassifications between the 'combined' groups: rock + rust, shadow rock + shadow ice, snow + glacier. On the other hand, there are misclassifications within the combined groups. The major difference between the 'original' and the 'augmented' analyses are that the classifications of the rock pixels are better for the augmented set of variables, whereas the results for the shadow-on-rock pixels are slightly inferior with the augmented set. The jackknifed classifications do not differ significantly from the ordinary classifications.

In Plates 1 and 2 we show the classifications of the total area, i.e., a classification of the 23,400 pixels by means of the two schemes. The major difference between the two images is that many more pixels are classified as rust instead of rock when only the four MSS bands are used. According to field geologists working in the area, the bulk of those are misclassifications. They should have been classified as rock pixels. Another difference is that the classifications based on the four bands show a more irregular pattern with less homogeneous areas. An adjustment of the prior probabilities for the rust group did, of course, remove many of the rust classified pixels. However, it was not only the misclassified pixels that were removed, but also many of the pixels that were known to be rusty.

In order to investigate this problem further, we have in Figure 4 given the projection of the 876 samples from the training sets on the plane determined by the first two canonical variates (canonical discriminant functions) based on B4, B5, B6, and B7. The canonical plot is a mapping of the points on the plane that maximizes between group variation as opposed to the within-group variation. In other words, the canonical plot gives the 'best' separation of the groups. For a more thorough discussion of this plot, see, e.g., Seber (1984). We see that there is a good distinction between shadow, rust + rock, and glacier + snow, whereas the differences between the subgroups are much smaller. This is in good accordance with the results obtained in the evaluation of the classification functions given above.

The distributions given in Figures 1 to 3, however, show that it should be possible to construct a discriminant function that could distinguish between the 'combined' groups. Therefore, a group of new classification functions was determined. First, functions were determined in order to discriminate between three merged groups: rust + rock, snow + glacier, and shadow ice + shadow rock. Second, functions were determined in order to discriminate between rock and rust, between snow and glacier,

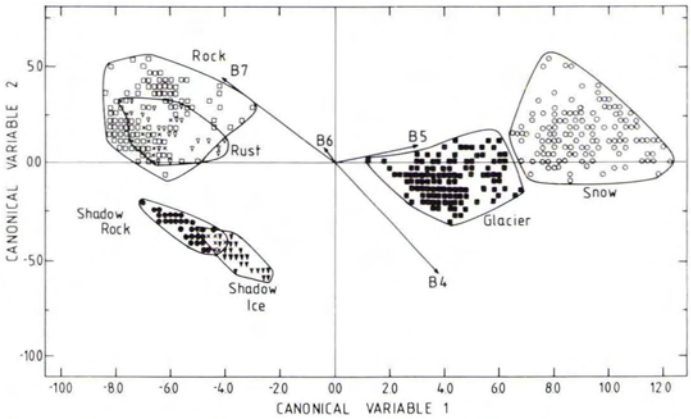

FIG. 4. The first two canonical variates for the six calibration groups, and the contribution of the standardized variables to the canonical variates shown as vectors. Overlaps of different groups are shown by x.

and between shadow on rock and shadow on ice. These functions have been determined by stepwise discriminant analyses in two situations, namely, one where only the original variables were allowed to enter the classification functions and one where all variables were allowed to enter.

The variables selected and their order of entering are given in Table 4. We see that there are substantial differences in the relevance of the variables in the different situations. When discriminating between snow and glacier for instance, the two most important variables are the same, namely band 5 and band 7. In the rock-rust discrimination the five most important variables are either ratios or factor scores.

After having determined these classification functions, a classification procedure was set up in the following way:

- Determine to which of the three merged groups—'rust + rock', 'snow + glacier', and shadow on ice + shadow on snow—a pixel belongs, by means of the first set of classification functions.
- After having determined the merged group, another classification is performed by means of the second set of classification functions in order to find the relevant subgroup.

The evaluation of these hierarchical classifications are shown in Table 3 under the headings 2-Band and 2-All. It is seen that the hierarchical procedures are giving results that generally are better than the one-step procedures. However, a better measure of the quality of the classification schemes is again found

TABLE 4. THE VARIABLES USED IN THE CLASSIFICATION FUNCTIONS IN THE ONE-STEP AND THE TWO-STEP PROCEDURES, AND THEIR ORDER OF ENTERING. THE VARIABLES ARE DETERMINED BY MEANS OF STEPWISE DISCRIMINANT ANALYSES WITH F-TO-ENTER = F-TO-REMOVE = 0.01.

	1 step class.		2 steps classification							
	Rock, Rust, Sh. Rock, Sh. Ice, Snow, Glacier		Rock + Rust, Sh. Rock + Sh. Ice, Snow + Glacier		Rock, Rust		Snow, Glacier		Sh. Rock, Sh. Ice	
Variables	Band	All	Band	All	Band	All	Band	All	Band	All
B4	2		1	2	4				1	1
B5	1	1	3		3		1	1	3	2
B6	4		4	5	2		3	4	4	9
B7	3		2		1	6	2	2	2	
Q4/5		6		8				6		8
Q4/6		2		1		5				7
Q4/7		5		9		4		7		3
Q5/6		10		10		8		5		5
Q5/7		8		6		7		3		4
Q6/7		4		4		1				10
F1										
F2		3		3						
F3		7				2		8		
F4		9		7		3				6

PLATE 1. Classification of the entire test area by means of the procedure based on bands 4-7 (1 Step-band). (Legend: white = snow, blue = glacier, grey = shadow ice, black = shadow rock, green = rock, red = rust).

PLATE 2. As in Plate 1, but based on ten variables (1 Step-All).

PLATE 3. As in Plate 1, but based on a hierarchical scheme with four bands (2 Steps-Band).

PLATE 4. As in Plate 1, but based on a hierarchical scheme with ten variables (2 Steps-All).

TABLE 5. THE NUMBER OF PIXELS FROM OUTSIDE THE TRAINING SET THAT ARE MISCLASSIFIED AS RUST.

	1-step band	1-step All	2-step Band	2-step All
Number of pixels misclassified as rust	2023	1204	573	126
As percentage of the total number of classified pixels	8.98%	5.35%	2.54%	0.55%

by classifying not only the 876 pixels in the training sets, but also the whole test area (i.e., 23,400 pixels). The results of these classifications are shown in Plates 3 and 4. We see that the two-step procedure, based on all variables, gives the smoothest segmentation of the image that is found to be in good accordance with ground truth knowledge. From an exploration point of view, it is obvious that a classification procedure should 'find' all alteration zones and have as few rock pixels as possible erroneously classified as rust. The four different pocedures 'found' all known color anomalies in the area. But, they differed very much with respect to the number of rock pixels that were classified as rust. These numbers are given in Table 5, and they show very clearly the very different performance of the four analyses. Better results are definitely obtained by augmenting the number of variables. In cases, as the one studied here, where there is hierarchical structure in the groups between which one classifies, as hierarchical procedure where different variables are used in different steps is superior to an ordinary discrimination procedure.

Classification schemes based on the principles developed here were subsequently used in a regional reconnaissance mapping of color anomalous zones in central east Greenland. Some more detailed geological conclusions are reported in Conradsen and Harpøth (1984).

CONCLUSION

In the present study we have shown that in geological applications it may be advantageous to introduce new variables as ratio and factor scores in classification functions, and that a hierarchical procedure utilizing an equivalent structure in the populations to be classified also can give very substantial improvements in the classifications. In the present case the hierarchical procedure based on all variables permitted in the classification functions gave by far the best result. Thus, it is concluded that such schemes represent a useful supplement to ordinary Bayesian classifications.

ACKNOWLEDGMENTS

Geologist Ole Harpøth, The Northern Mining Company, Copenhagen, is thanked for valuable assistance in providing the ground truth information. Comments from the referees helped improve the presentation of the results. The present work was sponsored by the Commission of the European Communities under contract no. 112-79-1 MPP (DK).

REFERENCES

Conradsen, K., J. Gunulf, O. Harpøth, and G. Nilsson, 1982. *The Application of Remote Sensing in Mineral Exploration*. IMSOR, Tech. Univ. Denmark, Lyngby, 122 p.

Conradsen, K., and O. Harpøth, 1984. Use of Landsat Multispectral Scanner Data for Detection and Reconnaisance Mapping of Iron Oxide Staining in Mineral Exploration, Central East Greenland. *Economic Geology*. Vol. 79, No. 6, pp. 1229–1244.

Dixon, W. J., and M. B. Brown, 1979. BMDP-79, *Biomedical Computer Programs*. Berkeley, Univ. California Press, 880 p.

Rowan, L. C., P. H. Wetlaufer, A. F. Goetz, F. C. Billingsley, and J. H. Stewart, 1977. *Discrimination of Rock Types and Detection of Hydrothermally Altered Areas in South-Central Nevada by the Use of Computer-Enhanced ERTS Images*. U.S. Geol. Surv., Prof Pap. No. 883, 35 p.

Seber, G.A.F., 1984. *Multivariate Observations*. New York John Wiley and Sons, 686 p.

(Received 23 May 1985; revised and accepted 14 February 1986)

Errata

The description of the photograph on the cover of the April 1986 issue of *PE&RS* should have stated that it was a combination of a Landsat TM image and a 1:25,000-scale topographic map and that it was reproduced on the cover at approximately 1:50,000 scale.