Mathematical Geology, Vol. 16, No. 2, 1984

An Algorithm for Clustering Profile Data and Its Application to Near-Surface Ice Content Data from Wet Coastal Tundra Soils near Barrow, Alaska[1]

G. F. Estabrook[2] and S. Outcalt[3]

An algorithm to cluster profile data into groups that minimize the sum of the intra-group variances was applied to near-surface soil ice content data collected near Barrow, Alaska, in wet tundra terrain. When the algorithm was requested to produce 2-5 groups and group mean profiles, the results were consistant with the modern theory of ice segregation. This process produces much of the variability of near surface soil ice stratigraphy in nature. These results strengthen the case for employing the algorithm on other profile data sets as an aid in hypothesis formulation.

KEY WORDS: clustering, ice, tundra, moisture content.

INTRODUCTION

A frequently encountered problem in the earth sciences is that of grouping, or "clustering," profile measurements of a continuous variable taken at different times or locations to yield a set of representative mean profiles. In the case of large data sets, it is frequently difficult to discuss the characteristics of "typical profiles." It would, therefore, be of great practical value if an algorithm were available to cluster profiles into representative groups for further analysis.

One reasonable criterion might be to assign profiles to groups so that the sum of the intra-group variances is minimized. Confidence in a heuristic procedure to do this would be increased if this minimum sum of intra-group variances was estimated from several random starting permutations of the data.

[1] Manuscript received 16 August 1982; revised 22 March 1983.

[2] The University of Michigan Herbarium, and Division of Biological Sciences, Ann Arbor, Michigan 48109, U.S.A.

[3] Department of Geological Sciences, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.

THE DATA

A series of moisture content determinations was made on 5-cm core sections of frozen soil collected from the surface downward to a depth of 80 cm at a site near Barrow, Alaska. In this data set, 51 profiles were complete from the surface to 80 cm. This yielded a matrix of 51 X 16 observations in which the initial core section and oven dry weights were obtained to estimate ice content by volume. The data matrix is included as Table 1. It is now necessary to briefly discuss the process of ice segregation, which is responsible for the variability observed in nature.

THE PROCESS OF ICE SEGREGATION IN THE ACTIVE LAYER

During the annual thermal regime, the uppermost soil layers of the Arctic Tundra thaw during the brief summer and refreeze again in the autumn. This shallow region of annual frost and thaw is called the "active layer." Beneath the active layer is the region of "permafrost," or perennially frozen ground. The base of the permafrost zone is reached at the level where the local geothermal gradient crosses the pressure-melting point near the 0°C isotherm. In the vicinity of Barrow, Alaska, where the ice content profile data were collected, the active layer varies between 40 and 60 cm in depth (Brown, 1969). The base of the permafrost is encountered at a depth of approximately 400 cm (Lachenbruch and Marshall, 1969; Brewer, 1958).

The movement of soil water within the active layer is controlled by the soil water potential gradient, which is almost a pure function of the temperature gradient at subfreezing temperatures. At above-freezing temperatures, the soil water potential is controlled primarily by the water content and thus maintains a strong feedback system coupled to water flux. Below the freezing point, the potential/temperature derivative is approximately 12,400 cm of water/°C, and as the temperatures are negative, the relationship leads to extremely high negative soil water potentials (positive tensions) at temperatures that are only slightly depressed below the freezing point. The hydraulic conductivity of a soil decreases abruptly with dessication and, therefore, decreases rapidly with increasing soil water tension. Therefore, there is a narrow thermal region between approximately 0 and −0.05°C in medium-textured soils where soil water potential gradients produce flow from warm to colder regions in an environment where the hydraulic conductivity is still sufficiently large to support flow in a diminished water cross-section fraction. Water is mobile at slightly subfreezing temperatures, and in the presence of weak thermal gradients the volume of this region of "subfreezing water movement" may be significant. Where the hydraulic conductivity of the soil becomes *limiting* due to low temperature, ice lenses will form if the flux of water into the region is greater than the rate of heat extraction divided by the heat of fusion (Outcalt, 1980). Thus, the process of "ice segregation" which results in lens growth depends upon a delicate balance between the flux of

Table 1

Core (group)	Ice[a]															
15-14 (2)	77	57	53	48	39	42	32	40	46	46	56	55	55	80	37	74
15-26 (2)	61	53	48	47	44	41	50	65	58	54	48	51	65	62	63	68
15-15 (5)	63	53	47	48	49	51	49	48	39	44	48	56	71	68	75	82
14-11 (5)	92	63	55	48	39	41	48	50	43	44	44	70	53	61	77	82
14-08 (2)	91	53	48	37	37	42	47	57	66	51	45	58	54	57	65	70
14-21 (2)	84	56	53	37	33	46	46	52	51	50	49	44	59	79	80	53
14-22 (2)	90	59	56	56	48	54	63	82	59	45	44	53	53	57	53	54
07-28 (3)	64	46	50	47	42	41	40	57	79	73	79	87	76	74	90	89
07-21 (3)	70	58	49	48	48	44	38	44	73	85	76	88	87	91	90	76
07-15 (4)	91	73	69	63	76	81	77	90	89	87	80	79	79	70	68	72
06-11 (4)	92	71	66	60	75	75	96	94	93	92	88	89	92	90	90	91
06-20 (1)	89	77	68	46	46	48	48	63	81	94	92	84	84	88	91	89
06-61 (1)	85	76	66	54	47	66	71	77	80	89	91	91	86	76	76	72
01-21 (5)	89	51	48	45	41	39	40	45	51	53	53	68	85	87	88	86
01-10 (3)	76	63	55	51	50	48	48	53	59	59	91	89	89	90	90	85
34-15 (4)	96	79	77	75	61	78	84	77	88	85	84	70	66	66	72	70
33-23 (1)	99	93	70	72	59	68	68	63	64	90	87	87	84	82	77	76
33-28 (4)	92	69	63	70	68	70	80	80	89	85	81	73	78	87	89	89
33-10 (4)	93	68	63	66	63	67	79	77	79	82	79	79	73	71	66	79
29-20 (5)	99	63	48	46	41	42	41	58	66	70	68	77	72	80	84	79
29-22 (1)	96	63	56	49	50	58	68	69	76	79	82	93	95	95	94	92
25-27 (4)	85	74	73	68	76	85	91	91	86	79	91	88	80	89	84	81
25-08 (4)	93	68	60	66	78	77	93	94	88	75	73	75	67	75	85	85
25-26 (4)	86	84	76	84	85	76	57	71	77	81	72	76	76	62	51	55
21-17 (5)	87	70	56	47	51	43	45	43	43	41	60	66	64	73	80	83
21-28 (5)	93	62	55	45	57	69	66	49	76	74	62	57	63	74	82	81
21-21 (2)	67	58	50	48	41	42	55	84	68	44	42	68	70	63	68	72
21-27 (3)	70	54	52	55	48	49	60	61	66	77	85	72	68	62	55	68
19-20 (2)	97	65	64	47	49	52	49	36	35	33	33	40	48	46	43	40
19-09 (5)	88	70	76	52	39	51	61	61	66	63	49	55	72	80	78	80
19-23 (1)	94	64	56	46	42	82	73	84	69	85	91	94	91	87	49	48
16-26 (2)	95	68	38	36	48	76	60	59	69	55	52	68	63	48	51	45
16-14 (5)	92	67	48	39	32	33	40	51	53	53	53	54	64	91	92	89
50-10 (4)	86	59	47	53	68	68	79	81	85	89	88	82	81	61	69	80
50-22 (3)	68	53	54	53	49	43	58	57	47	49	89	91	87	90	92	94
50-23 (1)	76	53	47	48	45	45	63	78	89	89	91	92	79	80	83	53
50-17 (4)	91	53	47	61	66	76	79	93	90	91	96	90	91	78	85	84
48-14 (3)	60	50	49	49	45	39	40	72	76	89	90	88	94	95	91	93
48-23 (3)	61	55	56	49	51	48	42	39	59	66	78	94	90	91	90	91
48-21 (1)	70	64	62	60	59	73	77	70	91	94	92	85	83	90	87	89
45-09 (1)	98	70	57	51	51	67	68	72	59	74	85	86	84	88	83	76
45-10 (1)	97	82	50	56	58	70	79	65	71	89	82	75	81	88	93	94
44-09 (1)	91	61	55	49	45	94	76	73	72	81	79	76	77	81	82	79
44-23 (1)	94	69	60	53	75	70	79	61	90	90	89	87	90	89	92	91
44-14 (4)	89	57	48	57	73	76	82	80	91	81	84	76	74	73	71	74
40-08 (3)	56	48	66	41	33	32	33	35	36	52	58	76	87	86	87	86
40-28 (3)	60	51	51	46	44	40	39	39	66	77	76	78	83	84	81	87
40-16 (1)	93	91	85	50	53	66	62	68	87	83	87	91	83	86	86	69
37-15 (1)	92	64	62	65	58	63	81	74	83	85	84	86	90	87	87	79
37-11 (3)	52	55	51	50	52	51	53	48	47	71	75	77	84	84	80	79
34-28 (4)	85	77	72	69	68	68	79	80	81	89	81	76	77	81	82	86

[a] By volume percent in 5-cm sections from 0-80 cm.

water from the thawed region on the warm side of the lens and the magnitude of heat flow toward the frozen region. In this system the soil water potential gradients produced by the presence of freezing fronts in the soil are several orders of magnitude larger than the local gravity potential and, thus, gravity effects are insignificant. The process has been validated in the field by observations of heave occurring in soils at subfreezing temperatures in both arctic and temperate regions (Mackay, Ostuck, Lewis, and MacKay, 1979; Czeppe and Widacki, 1973).

In review, two conditions are necessary for ice lens growth: (1) A wet thawed region with a hydraulic connection to, (2) a freezing-frozen region with a weak thermal gradient, which permits water to accumulate on the warm side of the growing lens. An ice lens is, thus, similar to lake or river ice as it is growing from a pool of free water. Ice segregation will be terminated by either a dessication of the water supply on the thawed side of the lens or a steepening of the thermal gradient into the frozen region.

Here, thawed, frozen, and freezing–thawing regions are defined. The freezing–thawing region is the location of potential ice segregation where water occurs both as a solid and liquid. During the period of summer thaw, soil water from the thawed region is drawn toward the frozen region, producing ice lens growth in the freezing region. Simultaneously, the downward advance of both the 0°C isotherm and the subfreezing isotherm that marks the freezing–frozen region boundary deepens the active layer. The complexity of this process is not apparent to an observer on the surface tracking the advance of the active layer by probing for the frozen boundary or estimating the depth of the isotherm from temperature measurements.

After snowmelt the active layer deepens rapidly, but the rate of deepening slows through the summer because the active layer depth is a function of the square root of elapsed time from snowmelt. At autumn freeze-up, two freezing fronts are generally present in wet soils, one moving from the surface downward and the other from the permafrost upward. There is the potential for ice segregation in the freezing–thawing regions associated with both of these fronts. However, the lower region is a more favorable site for massive lens growth because the thermal gradients near the base of the active layer are much weaker than those near the surface. The distance covered by downfreezing is between one and two orders of magnitude greater than the distance of upfreezing.

At wet sites, and especially where a horizontal connection to a water reservoir is maintained by the thawed region, one would anticipate the development of an ice-rich zone at the base of the active layer marking the location of the freezing-thawing thermal region during freeze-up. At drier sites and in locations without horizontal water advection during freeze-up, the pronounced ice-rich region at the base of the active layer should be absent.

At all sites, the geometry of freeze-up should produce a dessicated zone in the center of the active layer. In the final states of freeze-up, horizontal advection is eventually cut off at all sites because water moves both upward and downward away from the center of the active layer. The ice content of the uppermost core section indicates the degree of surface wetness at freezup.

Thus, as the original profiles are assigned to 2, 3, 4, and then 5 groups or clusters, the following representative characteristics should be present in the mean group profiles. First, dessication of the active layer should be an attribute of most profiles. Second, some groups should display the prominant base of the active-layer, ice-rich zone. Third, it is anticipated that a profile group or groups may indicate the effects of site dryness by exhibiting erratic behavior in comparison to the two previously described patterns.

THE ALGORITHM

With the design task in mind, we were able to rationally and defensibly choose, from among the wide variety of clustering techniques, the KMEANS algorithm of Helmut Spath. To enrich the printout with associated statistics and labels, the computer program originally published by Spath (1975), was modified. We briefly review the basic principle and algorithm here.

Denote with $X(i, j)$ the percentage of ice from profile (i) in section (j). In our case, (i) ranged from 1 to 51, and for each of these 51 profiles, (j) ranged from 1 to 16, representing profile core sections 5 cm in length extracted from the surface to a depth of 80 cm. Denote with $G(1, 1, j)$ the average percentage of ice in depth range (j) calculated over all cores, that is, over all profile sample sites. In one sense, this is the ice content pattern typical of the study area.

Denote with $S(1, 1, j)$ the sum over all 51 cores of the squared deviation, $[G(1, 1, j) - X(i, j)]$**2. Thus $S(1, 1, j)$ indicates how well $G(1, 1, j)$ typifies the ice content pattern at depth (j). This is the familiar "sum of squared error" approach. We denote the sum of $S(1, 1, j)$ over all depths (j) with $SS(1, 1)$, which is the total "squared error" associated with the typical ice content pattern $G(1, 1)$.

But there may be two or more distinct ice content profile patterns represented by the 51 cores. To discover whether this is plausable, we devide the 51 cores into two groups with typical ice content patterns $G(2, 1, j)$ and $G(2, 2, j)$. The first subscript denotes the number of groups into which cores (profiles) have been divided; the second subscript arbitrarily enumerates these groups; and the last subscript indicates the depth of the core section. We can now calculate the associated parameters $S(2, k, j)$ and $SS(2, k)$. We seek to divide the collection of core-profiles into two groups so that $SSS(2) = SS(2, 1) + SS(2, 2)$ is a minimum, that is, so that the sum of squared error around the two typical ice content profiles is minimized.

We repeat the process, dividing the 51 cores into three groups, with typical ice content patterns $G(3, 1, j)$, $G(3, 2, j)$, and $G(3, 3, j)$, so that $SSS(3) = SS(3, 1) + SS(3, 2) + SS(3, 3)$ is minimized. In the same way, we calculate typical ice content patterns for four and finally five groups. We were not confident that meaningful physical interpretations within the limits of the sample size could be extended to more than five groups.

The number of ways to divide the study collection of 51 cores into two groups is approximately 2 raised to the 51st power. The number of ways to divide these cores into three groups is several orders of magnitude greater, and even greater for more groups. It is not feasable, even with a computer, to evaluate our optimality criterion for each possible grouping. Therefore, the computer program employs a valley-descending heuristic that attempts to guess the best grouping. The number of groups desired is prespecified, and the program begins by using a random number generator to divide the cores into that number of groups. The optimality criterion $SSS(k)$ is then evaluated. One-by-one, cores (profiles) are transferred from one group to another, and at each transfer the optimality criterion is evaluated and compared with its previous value. If the new value is higher, the core is returned to its former group; if lower, the core is retained in its new group. The program continues until no core can be moved into another group. The squared error is now locally minimum. However, it is possible that a quite different group assignment might have an even lower squared error. Thus, the procedure finds a local minimum that is a function of the initial group assignment. It is possible that a local minimum may not be the global minimum. To minimize this possibility, the program was executed 12 times from different initial random group assignments for each number of groups we sought.

In the case of two groups, the same two optimal groups were produced by each of the 12 executions of the procedure. When three groups were specified the procedure also produced the same optimal group in each of the 12 executions. When four or five groups were specified, searching among a staggering number of possible groupings, the same optimal groups occurred in over half of the trials, but some different groupings with a higher squared error were determined in a few trials. We are reasonably confident that the groupings we present here have squared errors that are at, or near, the global minimum.

Because each grouping into a specified number of groups is done independently, the optimal series of the best two, the best three, the best four, and the best five groups need not be hierarchical. When two or more groups from a larger number in this series can be combined to form a single group from a lower number in the series, it suggests that a typical pattern decomposes into typical subpatterns. Because hierarchical relationships are not forced by the procedure, when they do occur they can be interpreted as indicating an underlying hierarchical structure in the data.

Another indication of the reliability of the groupings at each level is the errors $SS(k, m)$ themselves, or these errors divided by one less than the number of cores in the group (k, m). $SSS(k)$ is a decreasing function of k, but the difference $[SSS(k) - SSS(k - 1)]$ may prove to be a statistic whose approximate distribution could provide significance levels for the revealed groups. Without simulating them, it is not clear to us how to approximate the distributions of this or other statistics associated with our analysis. Therefore, at this time, we are not providing significance levels.

THE GROUPS REVEALED

The mean profiles of the groups produced by the algorithm are presented as Fig. 1. In this figure each element of the horizontal histogram represents the ice content, percent by volume, of a 5-cm core section. The vertical scale runs from

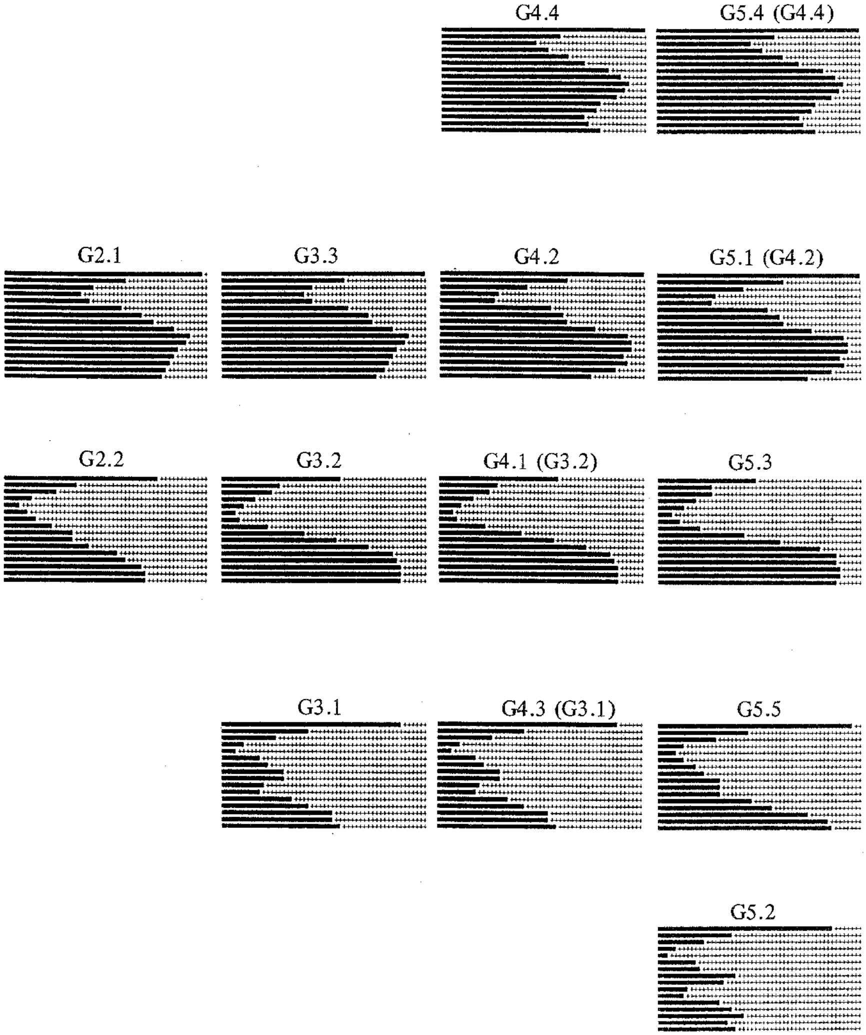


Fig. 1. Group histograms.

the surface at the top of the histogram to a depth of 80 cm at the bottom of the histogram, and the horizontal scale ranges from 40% at the left to 90% at the right. The groups are designated by the notation GK.I, which is equivalent to *G*(*k*, *i*). Where a higher-level groups is identical to a lower-level group, the relationship is indicated by parentheses.

The two groups are based on the presence in G2.1 of an ice-rich region at the base of the active layer at a depth of 50 cm. Both G2.1 and G2.2 display the anticipated dessication of the active layer. The presence of a wet surface in G2.1 may be physically correlated with the presence of the basal lens, which requires

Table 2

	Thaw depth (cm)	
	Minimum	Maximum
Group 1		
06–20	29	40
06–21	34	38
33–23	28	37
29–22	28	39
19–23	34	55
50–23	38	45
48–21	31	40
45–09	24	38
45–10	29	36
44–09	29	41
44–23	29	35
40–16	26	35
37–15	31	39
Mean	30	40
Group 2		
15–14	50	64
15–26	41	51
14–08	41	53
14–21	50	63
14–22	39	60
21–21	38	51
19–20	50	59
16–26	26	38
Mean	42	55
Group 3		
07–28	44	49
07–21	40	48
01–10	38	45
21–27	34	42
50–22	41	48
48–14	39	48
48–23	43	49
40–08	54	58
40–28	47	54
37–11	47	52
Mean	41	51
Group 4		
07–15	26	37
06–11	21	30
34–15	28	40
33–28	27	40
33–10	30	41
25–27	19	29
25–08	28	37
25–26	25	35
50–10	25	42
50–17	25	34
44–14	26	29
34–28	28	34
Mean	26	36
Group 5		
15–15	48	60
14–11	50	52
01–21	50	53
29–20	29	50
21–17	30	52
21–28	36	44
19–09	36	60
16–14	40	50
Mean	39	53

a wet active layer for its growth. Note that the active layer depth is inversely correlated with the total ice content. This is attributed to the increased volumetric heat of fusion in ice-rich soil and the corresponding reduction in the effective thermal diffusivity with increased wetness (Brown, 1969). Further division follows two paths after this initial split.

In the upper part of Fig. 1, G2.1 becomes G3.3 with the loss of profile 48-14. G3.3 in turn divides into G4.4 and G4.2, and these groups remain unchanged as G5.4 and G5.1. All the upper diagram profiles show active layer dessication and a basal lens. The major differences between G5.4 and G5.1 are the deeper dessicated zone in G5.1, and different trends of ice content at the base of the profiles.

In the lower portion of the diagram, G2.2 divides into G3.2 and G3.1. The lower part of G3.2 is ice rich, which might be interpreted as basal ice below a 60-cm active layer. It is possible that the division between the upper and lower profiles is primarily based on active layer depth, which in turn is caused by site wetness. Note that the surface ice content of all profiles in the upper portion of the diagram is 90%. G3.2 and G3.1 are transformed intact into G4.1 and G4.3. Profile G5.3 is quite similar to G4.1. Profile G4.3 splits into G5.5, which in-

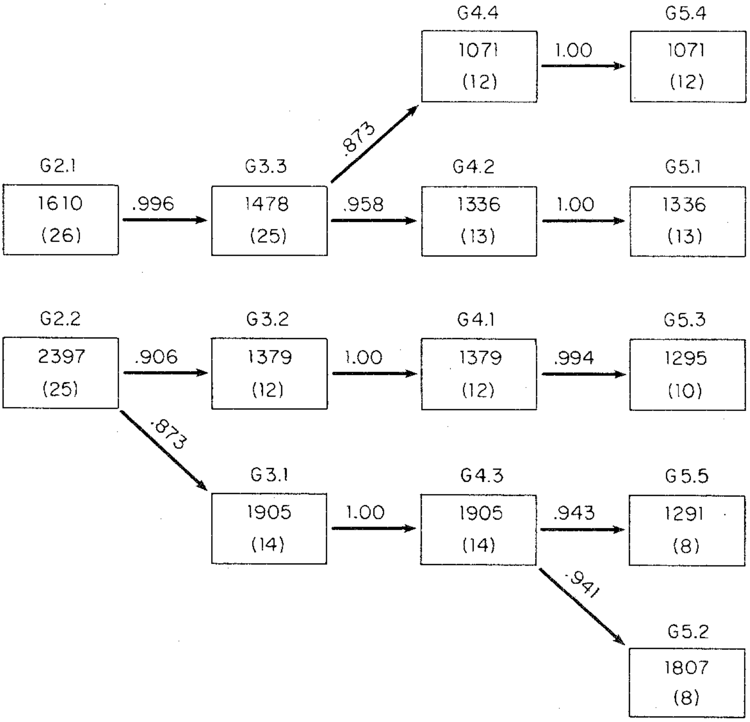
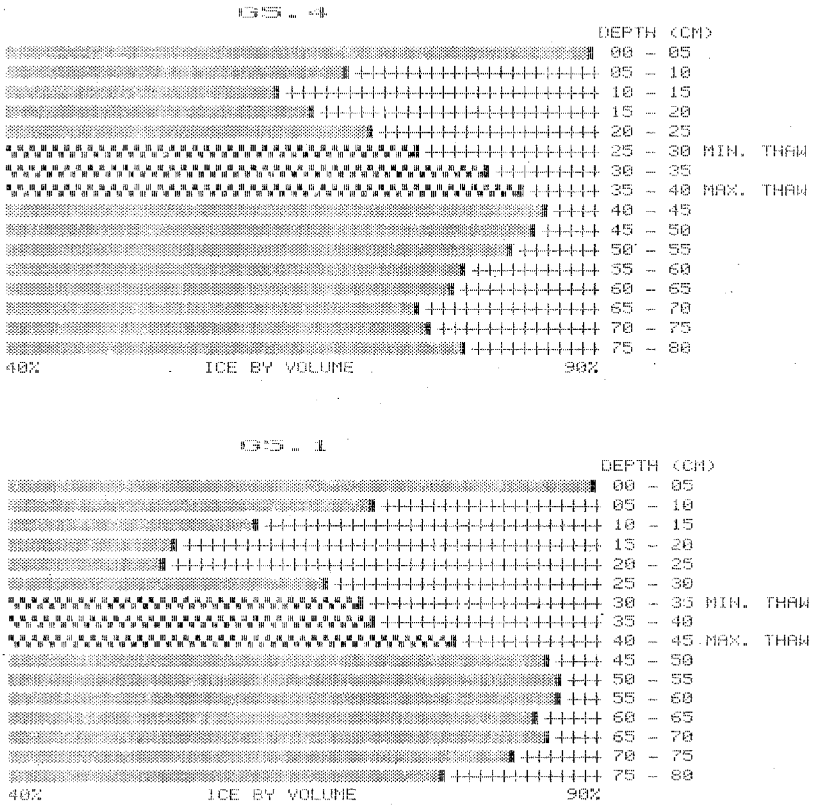


Fig. 2. Group statistics.

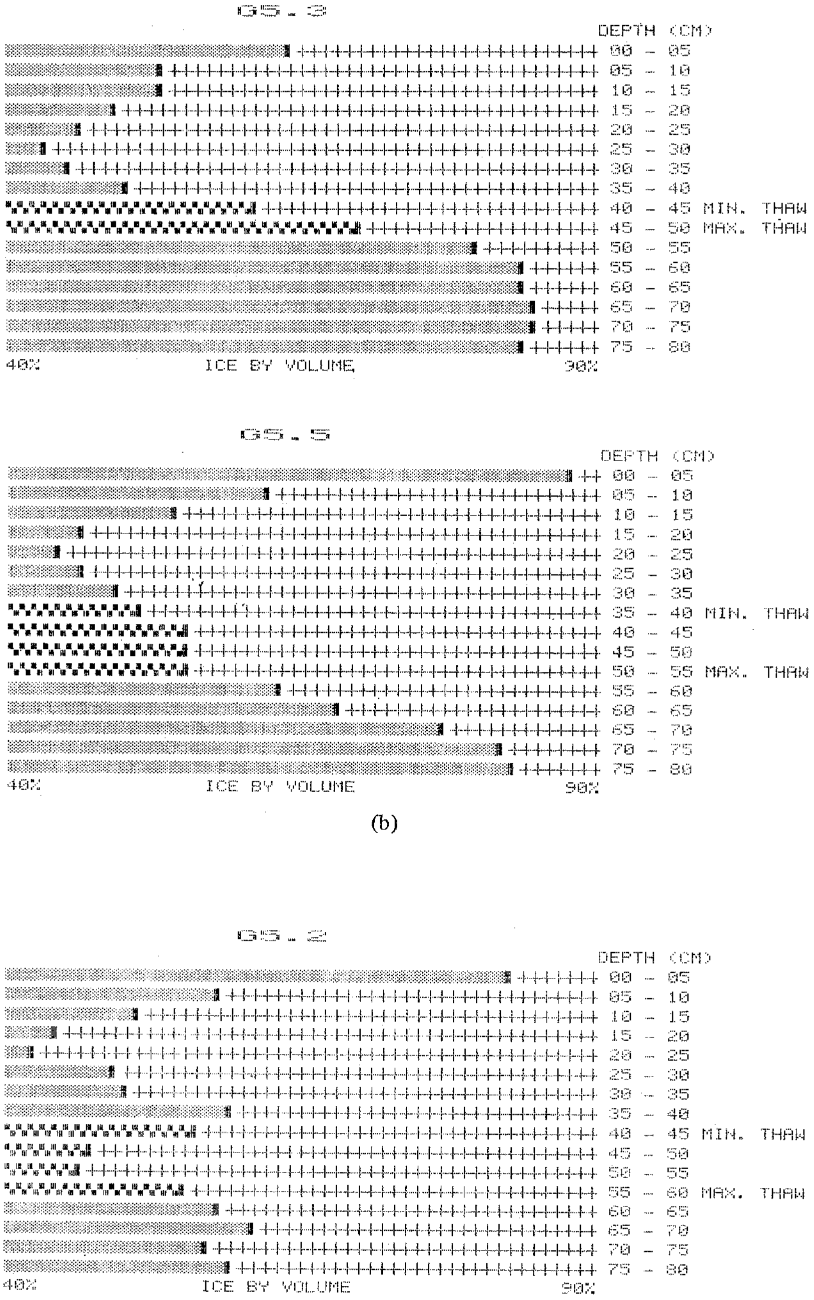
cludes sites with a wet surface but deep dessication, and G5.2, which seems to be a dump group for erratic patterns. This last group may contain sites whose normal pattern of active layer dessication and basal ice enrichment has broken down due to extreme site dessication effects.

A five-year minimum and maximum active layer depth for each core included in the study was calculated from data tables in an unpublished report by Brown. These statistics by group at level 5 are summarized in Table 2 and displayed as Fig. 3. From these data it is apparent that the basal ice-rich zone in group 5.4 is a product of the present active layer regime abstracted in Table 2. The patterns of deeper basal ice in G5.1 and G5.3 are also in adjustment with the present summer thaw regime. The major difference between G5.1 and G5.3 is the much dryer surface and more pronounced dessication in G5.3. Typical profile G5.5 exhibits extreme dessication and may represent an extreme case of climatological or geomorphic dessication in the past. Once deep thaw occurs and



(a)

Fig. 3. Level 5 group profiles.



(b)

(c)

Fig. 3. Continued.

water is drained away from a site, patterns like G5.2 and G5.5 may persist with the annual frost and thaw confined to the dessicated region accompanied by a low amplitude heave regime. However, core patterns similar to G5.4 indicate that high amplitude soil heave is occurring at the site under its present Klima-Geomorphic regime.

GROUP STATISTICS

The statistics associated with Fig. 1 are displayed in Fig. 2. The arrow labels connecting the group blocks are the correlation coefficients of the group core section values. Within the blocks the upper number is the within-group variance and the lower number in parentheses is the number of profiles included in the group. Note that G5.2 contains the highest variance at level 5 and is derived from groups that contain the highest variance at all preceding levels. This pattern indicates that this group may be a dump for erratic profiles.

CONCLUSION

The clustering procedure, in structuring groups that minimize the sum of within-group squared deviations from typical group profiles, has produced groups that indicate hierarchial structure in the data consistent with physical theory. An estimate of the magnitude of annual surface heave may be available through an analysis of ice stratigraphy as the magnitude of surface heave increases with site wetness (Benedict, 1976). This analysis has demonstrated that total ice volume and the presence of massive ice at the base of the active layer are indicators of wetness and, thus, high amplitude surface heave.

ACKNOWLEDGMENT

As we used it, the Spath algorithm was implemented in a Fortran program by Dr. Neal Oden, Division of Biological Science, SUNY Stony Brook, and subsequently modified by G. F. Estabrook. We are indebted to Dr. Jerry Brown, Chief, Earth Sciences Branch, CRREL, Hanover, New Hampshire, for the use of his data in our example.

REFERENCES

Benedict, J., 1976, Frost creep and gelifluction: A Review: Quat. Res., v. 6, p. 55–76.

Brewer, M. C., 1958, Some results of geothermal investigations of permafrost in Northern Alaska: Amer. Geophys. Union Trans., v. 39, p. 278–284.

Brown, J., 1969, Soil properties developed on the complex tundra relief of northern Alaska: Biuletyn Periglacjalny, v. 18, p. 153–167.

Czeppe, Z. and Widacki, W., 1973, The probable influence of humidity on frost movements of soil at Ciezkowice near Tarnow: Studia Geomorphologicia Carpatho-Balanica, v. 7, p. 105–108.

Lachenbruch, A. H. and Marshall, B. V., 1969, Heat flow in the Arctic: Arctic, v. 22, p. 300–311.

Mackay, J. R., Ostrick, J., Lewis, C. P., and MacKay, D. K., 1979, Frost heave at ground temperatures below OC, Inuvik, Northwest Territories, *in*, Science and Technology, Notes in Current Research: Geological Survey of Canada, Paper 79-1A, p. 403–405.

Outcalt, S. I., 1980, A simple energy budget model of ice segregation: Cold Reg. Sci. Technol., v. 3, p. 145–151.

Spath, H., 1975, Cluster-analyse algorithmen: Oldenbourg-Munich.