AN ATTEMPT TO ELIMINATE IDENTICAL LOGS IN A LARGE SOIL DRILLING LOG DATABASE

Hideyasu Asahi[1], **Toshikazu Kuroshima**[1] , **Kunio Kawauchi and Shinji Hirai**[1]

[1]Muroran Institute of Technology

Mizumoto-cho 27-1, Murorann City, 050-8585 Japan

Email: aasan98@mmm.muroran-it.ac.jp

ABSTRACT

It is often found the identical logs in soil drilling log database. In this paper, we propose a method to detect the identical logs in a large drilling log database to compile reliable database.

The detections are performed by the method as mentioned below.

1) *The soil names in soil log are converted to number.*
2) *The numerized drilling-log are converted to the spectra of Walsh-Hadamard function.*
3) *Sort a few (4 or 5) spectra using a spreadsheet program software such as Lotus 123 or MS Excel.*
4) *Identical drilling logs are found in the table side by side .*

1 INTRODUCTION

A considerable amount of time and effort is spent accumulating drilling logs in order for geological consultant companies, public research institutes and universities to construct drilling log databases. These organizations spend a considerable amount of time and effort obtaining drilling logs and constructing databases for the reuse of log information.

The problem is that identical logs are found in most drilling log databases. In some cases more than 10% of identical logs may be found in the database. This problem is difficult to avoid even if a lot of work is spent on the detection and exclusion of identical logs when the editor constructs a drilling log database. This is because it is difficult to compare identical logs from different locations by visual inspection.

In the past, small databases were rationalized using cluster analysis. But this technique requires large amounts of computer memory and high speed processors. The comparison of about 3000 logs is probably the limit on a PC. Furthermore, complex visual inspection is still needed after cluster analysis in order to completely eradicate the duplicates.

In this paper we discuss an alternative solution to the problem using Fourier-type spectra of rectangular waveform functions.

2 IDEA

2.1 Numerical treatments

We substitute numerical values for the drilling log soil names, then sample the numerical value at equal intervals by depth. Then we perform Fourier-type transformation on the sampled sequences using the orthogonal rectangular waveform function system. The sorting and comparison of these spectra assist the discovery of identical logs in the database. Drilling log databases of Noboribetsu City and the central Hokkaido area of Japan were used in the analysis (Figure 4). These databases contain 850 and 13,000 logs respectively (including duplicates).

2.2 Numerization of soil names in the drilling log database

In order to conveniently process the logs, the soil names are grouped into 13 categories for the Noboribetsu City database. But, as described later, no categorization occurs in the central Hokkaido database and all 119 soil names are numerized individually. Then we sample the numerical value sequence **f** = ($f_0\ f_1\ f_2\cdots f_l\cdots f_{n-1}$) at equal interval points ($x_0\ x_1\ x_2\cdots x_l\cdots x_{n-1}$) in the depth direction of n finite numbers as in Figure 1. We perform discrete Fourier-type transformation on the data sequence f_l using the complete system of orthogonal and normal functions $\mathbf{h}_i = (h_{i0}\ h_{i1}\ h_{i2}\cdots h_{il}\cdots h_{i(n-1)})$ (Figure 3) :

$$c_i = \sum_{l=0}^{n-1} h_{il} \cdot f_l \qquad (1)$$

As we know, the relationship between $f_l(x_l)$ and the spectrum c_i can be expressed by following Perseval's theorem:

$$\sum_{i=0}^{n-1} c_i^2 = \sum_{l=0}^{n-1} f_l^2 \quad (=T) \qquad (2)$$

Total T is a constant value for each drilling log. It is equal to all the information on the drilling log. The information ratio that spectrum value c_i shares is as follows:

$$p_i = c_i^2 / T \qquad (i : 0, 1, 2, \cdots, n\text{-}1) \qquad (3)$$

As for the magnitude of each spectrum c_i is the square root of the information ratio in data sequence f. Furthermore, the cumulative information ratio H_i is calculated by the

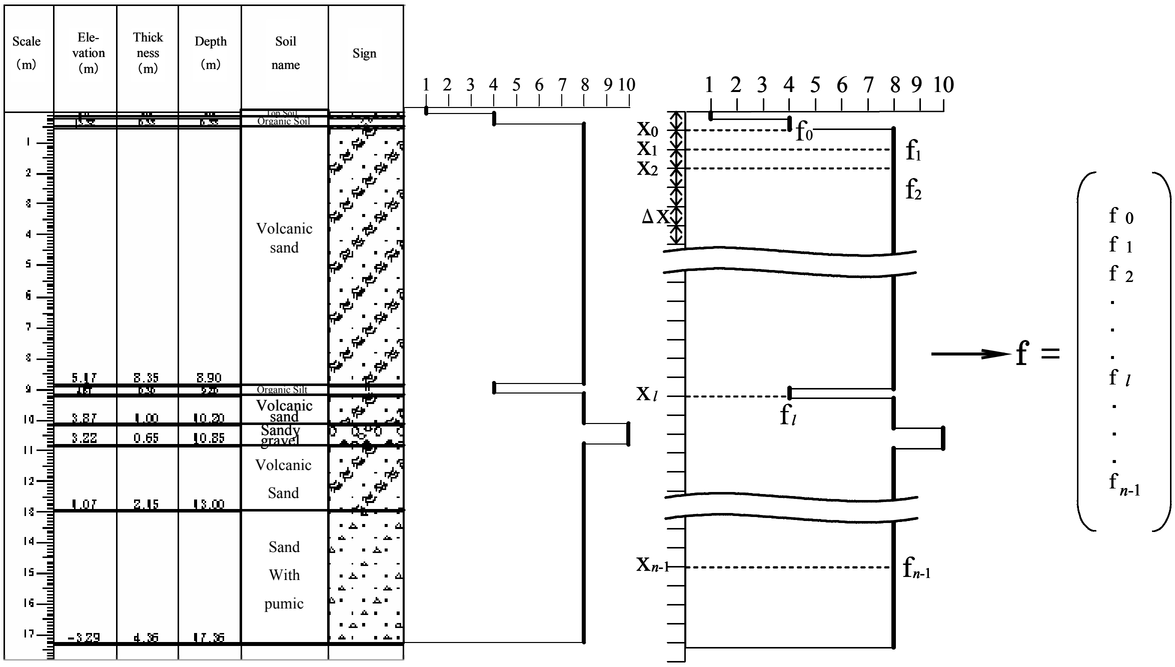


Figure 1. Schematic diagram of Numerization of drilling log

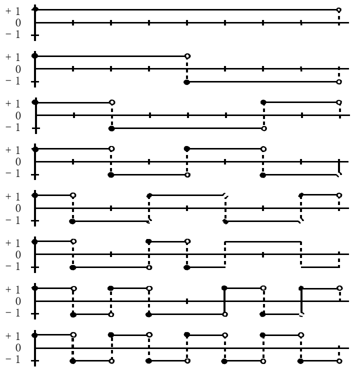


Figure 2. Continuous Walsh-Hadamard Function system, n = 8

$$H_8 = \begin{pmatrix} \mathbf{h}_0 \\ \mathbf{h}_1 \\ \mathbf{h}_2 \\ \mathbf{h}_3 \\ \mathbf{h}_4 \\ \mathbf{h}_5 \\ \mathbf{h}_6 \\ \mathbf{h}_7 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{pmatrix}$$

Figure 3. Discrete Walsh-Hadamard Function system, n = 8

following equation（Asahi, 1995）:

$$H_i = \sum_{j=0}^{i} p_j \qquad (4)$$

These values are shown as a percentage.

Incidentally, in this paper we use spectra from the Walsh-Hadamerd function system as in Figure 2 and Figure 3.

Each spectrum encapsulates information throughout the data sequence. The direct current component especially contains a useful and important piece of information about the drilling log data sequences. So, the comparison of the DC component spectra and the other few spectra make it possible to distinguish different logs and detect identical ones. These values are shown as a percentage.

3 Application

3.1 Cumulative information ratio

The soil names in the Noboribetsu drilling log database were grouped into 13 categories. They are surface soil, high organic soil, volcanic ash soil, organic soil, clay, silt, sandy soil, sand, gravely soil, gravel, tuffs（the Quaternary formation）, hard soil（the Tertiary formation） and rock. Numerical values of 1 to 13 were applied to each category.

In the central Hokkaido database, however, unexpected results were produced as a result of this grouping. One third of the logs became identical when only 13 soil categories were used. We therefore applied numerical values of 1 to 119 to these soil names.

Figure 5 shows the distribution of cumulative information ratios with sequence order in the Noboribetsu database. These cumulative information ratios are calculated using the 1244 logs of more than 10m depth.

As shown in Figure 5, the direct current components account for almost 97% of the entire numerized drilling log information. This same tendency has been found before（Asahi et al., 1995）. The direct current component ratio greatly depends on the numerical values of the soil categories. However, we do not exclude the direct current components on the grounds that they are statistically equivalent to the mean.

The detection process of identical logs is as follows:

1）Calculate the spectra using numerical formula（1）.

2）Import the spectra into spreadsheet software such as MS Excel or Lotus 123.

3）Sort the sets of spectrum sequences in order of the direct current components.

4）Subtract each direct current component from the adjoining one.

5）A result of 0 indicates that the spectrum（log） is probably a duplicate, so move it to the top of the spreadsheet.

6）Repeat steps 3） to 5） for the spectra of the following sequences.

7）Examine the original drilling logs for the suspected duplicates to confirm that they are in fact duplicates.

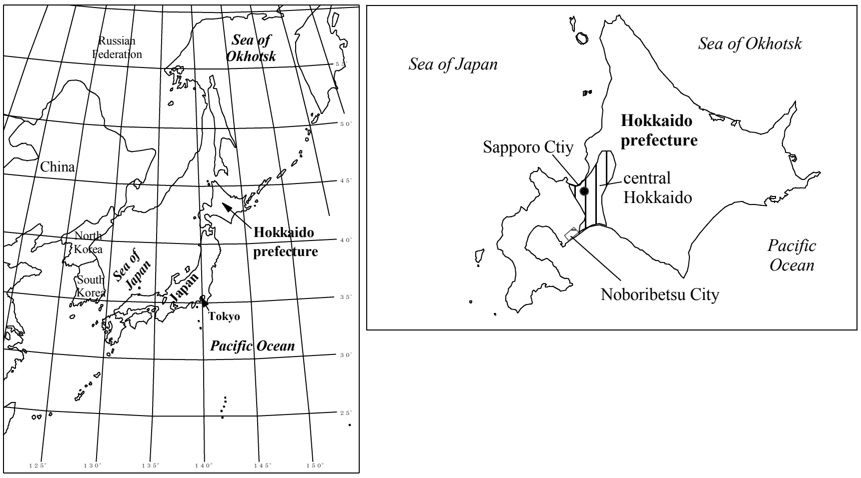


Figure 4. Location of the central Hokkaido and Noboribetu City

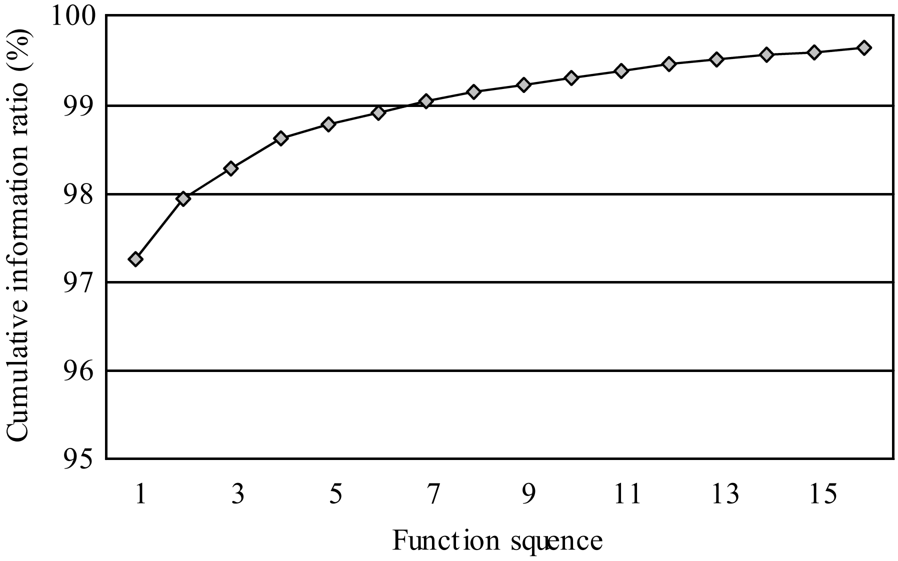
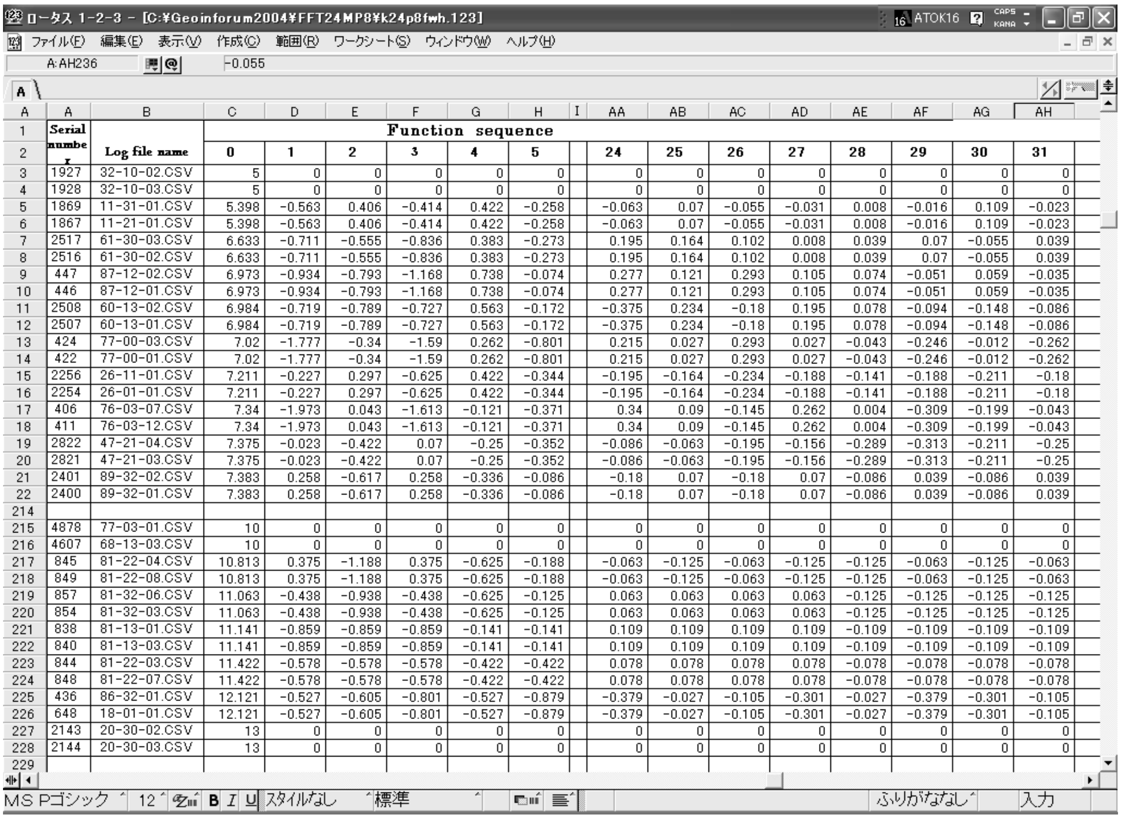
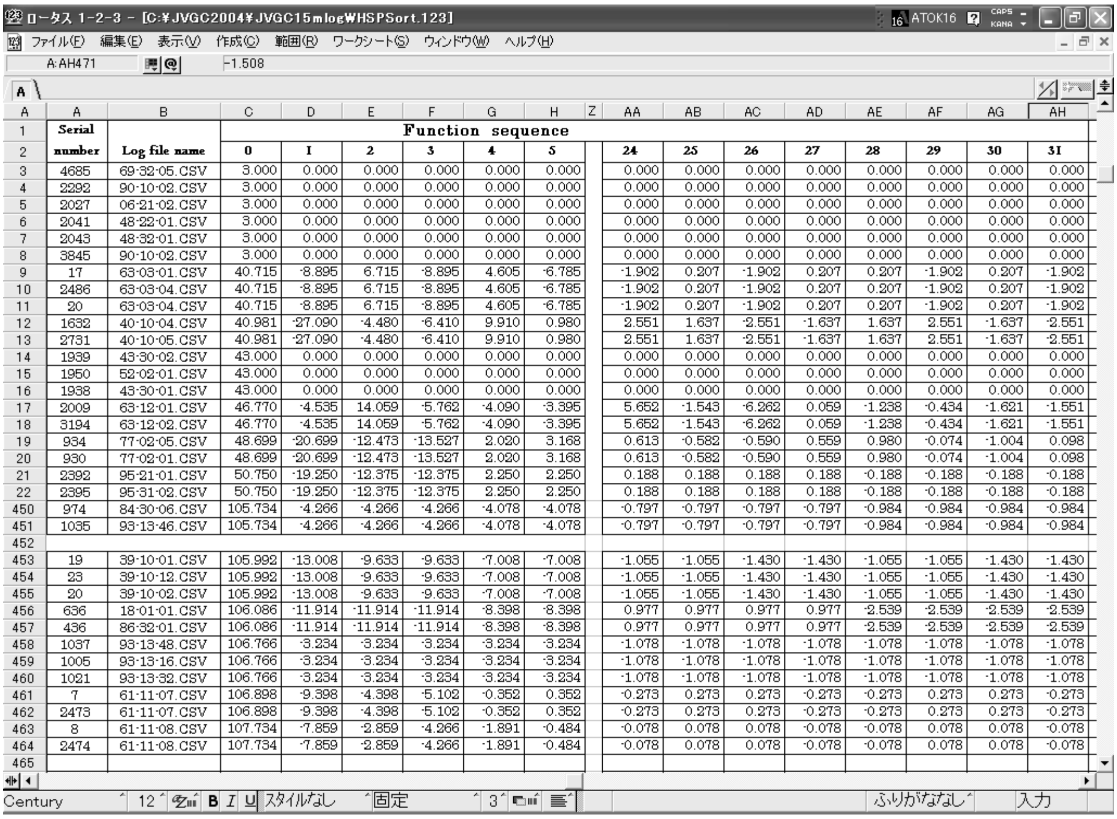


Figure 5. Cumulative information ratio


ロータス 1-2-3 - [C:¥Geoinforum2004¥FFT24MP8¥k24p8fwh.123]
ATOK16
CAPS
KANA
ファイル(F) 編集(E) 表示(V) 作成(C) 範囲(R) ワークシート(S) ウィンドウ(W) ヘルプ(H)
A:AH236
-0.055
A
MS Pゴシック
12
B I U
スタイルなし
標準
ふりがななし
入力

A	A	B	C	D	E	F	G	H	I	AA	AB	AC	AD	AE	AF	AG	AH
1	Serial number		Function sequence														
2		Log file name	0	1	2	3	4	5		24	25	26	27	28	29	30	31
3	1927	32-10-02.CSV	5	0	0	0	0	0		0	0	0	0	0	0	0	0
4	1928	32-10-03.CSV	5	0	0	0	0	0		0	0	0	0	0	0	0	0
5	1869	11-31-01.CSV	5.398	-0.563	0.406	-0.414	0.422	-0.258		-0.063	0.07	-0.055	-0.031	0.008	-0.016	0.109	-0.023
6	1867	11-21-01.CSV	5.398	-0.563	0.406	-0.414	0.422	-0.258		-0.063	0.07	-0.055	-0.031	0.008	-0.016	0.109	-0.023
7	2517	61-30-03.CSV	6.633	-0.711	-0.555	-0.836	0.383	-0.273		0.195	0.164	0.102	0.008	0.039	0.07	-0.055	0.039
8	2516	61-30-02.CSV	6.633	-0.711	-0.555	-0.836	0.383	-0.273		0.195	0.164	0.102	0.008	0.039	0.07	-0.055	0.039
9	447	87-12-02.CSV	6.973	-0.934	-0.793	-1.168	0.738	-0.074		0.277	0.121	0.293	0.105	0.074	-0.051	0.059	-0.035
10	446	87-12-01.CSV	6.973	-0.934	-0.793	-1.168	0.738	-0.074		0.277	0.121	0.293	0.105	0.074	-0.051	0.059	-0.035
11	2508	60-13-02.CSV	6.984	-0.719	-0.789	-0.727	0.563	-0.172		-0.375	0.234	-0.18	0.195	0.078	-0.094	-0.148	-0.086
12	2507	60-13-01.CSV	6.984	-0.719	-0.789	-0.727	0.563	-0.172		-0.375	0.234	-0.18	0.195	0.078	-0.094	-0.148	-0.086
13	424	77-00-03.CSV	7.02	-1.777	-0.34	-1.59	0.262	-0.801		0.215	0.027	0.293	0.027	-0.043	-0.246	-0.012	-0.262
14	422	77-00-01.CSV	7.02	-1.777	-0.34	-1.59	0.262	-0.801		0.215	0.027	0.293	0.027	-0.043	-0.246	-0.012	-0.262
15	2256	26-11-01.CSV	7.211	-0.227	0.297	-0.625	0.422	-0.344		-0.195	-0.164	-0.234	-0.188	-0.141	-0.188	-0.211	-0.18
16	2254	26-01-01.CSV	7.211	-0.227	0.297	-0.625	0.422	-0.344		-0.195	-0.164	-0.234	-0.188	-0.141	-0.188	-0.211	-0.18
17	406	76-03-07.CSV	7.34	-1.973	0.043	-1.613	-0.121	-0.371		0.34	0.09	-0.145	0.262	0.004	-0.309	-0.199	-0.043
18	411	76-03-12.CSV	7.34	-1.973	0.043	-1.613	-0.121	-0.371		0.34	0.09	-0.145	0.262	0.004	-0.309	-0.199	-0.043
19	2822	47-21-04.CSV	7.375	-0.023	-0.422	0.07	-0.25	-0.352		-0.086	-0.063	-0.195	-0.156	-0.289	-0.313	-0.211	-0.25
20	2821	47-21-03.CSV	7.375	-0.023	-0.422	0.07	-0.25	-0.352		-0.086	-0.063	-0.195	-0.156	-0.289	-0.313	-0.211	-0.25
21	2401	89-32-02.CSV	7.383	0.258	-0.617	0.258	-0.336	-0.086		-0.18	0.07	-0.18	0.07	-0.086	0.039	-0.086	0.039
22	2400	89-32-01.CSV	7.383	0.258	-0.617	0.258	-0.336	-0.086		-0.18	0.07	-0.18	0.07	-0.086	0.039	-0.086	0.039
214																	
215	4878	77-03-01.CSV	10	0	0	0	0	0		0	0	0	0	0	0	0	0
216	4607	68-13-03.CSV	10	0	0	0	0	0		0	0	0	0	0	0	0	0
217	845	81-22-04.CSV	10.813	0.375	-1.188	0.375	-0.625	-0.188		-0.063	-0.125	-0.063	-0.125	-0.125	-0.063	-0.125	-0.063
218	849	81-22-08.CSV	10.813	0.375	-1.188	0.375	-0.625	-0.188		-0.063	-0.125	-0.063	-0.125	-0.125	-0.063	-0.125	-0.063
219	857	81-32-06.CSV	11.063	-0.438	-0.938	-0.438	-0.625	-0.125		0.063	0.063	0.063	0.063	-0.125	-0.125	-0.125	-0.125
220	854	81-32-03.CSV	11.063	-0.438	-0.938	-0.438	-0.625	-0.125		0.063	0.063	0.063	0.063	-0.125	-0.125	-0.125	-0.125
221	838	81-13-01.CSV	11.141	-0.859	-0.859	-0.859	-0.141	-0.141		0.109	0.109	0.109	0.109	-0.109	-0.109	-0.109	-0.109
222	840	81-13-03.CSV	11.141	-0.859	-0.859	-0.859	-0.141	-0.141		0.109	0.109	0.109	0.109	-0.109	-0.109	-0.109	-0.109
223	844	81-22-03.CSV	11.422	-0.578	-0.578	-0.578	-0.422	-0.422		0.078	0.078	0.078	0.078	-0.078	-0.078	-0.078	-0.078
224	848	81-22-07.CSV	11.422	-0.578	-0.578	-0.578	-0.422	-0.422		0.078	0.078	0.078	0.078	-0.078	-0.078	-0.078	-0.078
225	436	86-32-01.CSV	12.121	-0.527	-0.605	-0.801	-0.527	-0.879		-0.379	-0.027	-0.105	-0.301	-0.027	-0.379	-0.301	-0.105
226	648	18-01-01.CSV	12.121	-0.527	-0.605	-0.801	-0.527	-0.879		-0.379	-0.027	-0.105	-0.301	-0.027	-0.379	-0.301	-0.105
227	2143	20-30-02.CSV	13	0	0	0	0	0		0	0	0	0	0	0	0	0
228	2144	20-30-03.CSV	13	0	0	0	0	0		0	0	0	0	0	0	0	0
229																	

Figure 6. Sorted spreadsheet for the Noboribetsu database


ロータス 1-2-3 - [C:¥JVGC2004¥JVGC15mlogWHSPSort.123]
ATOK16
CAPS
KANA
ファイル(F) 編集(E) 表示(V) 作成(C) 範囲(R) ワークシート(S) ウィンドウ(W) ヘルプ(H)
A:AH471
-1.508
A
Century
12
B I U
スタイルなし
固定
3
ふりがななし
入力

A	A	B	C	D	E	F	G	H	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	**Serial**		**Function sequence**														
2	**number**	**Log file name**	**0**	**I**	**2**	**3**	**4**	**5**		**24**	**25**	**26**	**27**	**28**	**29**	**30**	**3I**
3	4685	69-32-05.CSV	3.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	2292	90-10-02.CSV	3.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
5	2027	06-21-02.CSV	3.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	2041	48-22-01.CSV	3.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	2043	48-32-01.CSV	3.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
8	3845	90-10-02.CSV	3.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
9	17	63-03-01.CSV	40.715	-8.895	6.715	-8.895	4.605	-6.785		-1.902	0.207	-1.902	0.207	0.207	-1.902	0.207	-1.902
10	2486	63-03-04.CSV	40.715	-8.895	6.715	-8.895	4.605	-6.785		-1.902	0.207	-1.902	0.207	0.207	-1.902	0.207	-1.902
11	20	63-03-04.CSV	40.715	-8.895	6.715	-8.895	4.605	-6.785		-1.902	0.207	-1.902	0.207	0.207	-1.902	0.207	-1.902
12	1632	40-10-04.CSV	40.981	-27.090	-4.480	-6.410	9.910	0.980		2.551	1.637	-2.551	-1.637	1.637	2.551	-1.637	-2.551
13	2731	40-10-05.CSV	40.981	-27.090	-4.480	-6.410	9.910	0.980		2.551	1.637	-2.551	-1.637	1.637	2.551	-1.637	-2.551
14	1939	43-30-02.CSV	43.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
15	1950	52-02-01.CSV	43.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
16	1938	43-30-01.CSV	43.000	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
17	2009	63-12-01.CSV	46.770	-4.535	14.059	-5.762	-4.090	-3.395		5.652	-1.543	-6.262	0.059	-1.238	-0.434	-1.621	-1.551
18	3194	63-12-02.CSV	46.770	-4.535	14.059	-5.762	-4.090	-3.395		5.652	-1.543	-6.262	0.059	-1.238	-0.434	-1.621	-1.551
19	934	77-02-05.CSV	48.699	-20.699	-12.473	-13.527	2.020	3.168		0.613	-0.582	-0.590	0.559	0.980	-0.074	-1.004	0.098
20	930	77-02-01.CSV	48.699	-20.699	-12.473	-13.527	2.020	3.168		0.613	-0.582	-0.590	0.559	0.980	-0.074	-1.004	0.098
21	2392	95-21-01.CSV	50.750	-19.250	-12.375	-12.375	2.250	2.250		0.188	0.188	0.188	0.188	-0.188	-0.188	-0.188	-0.188
22	2395	95-31-02.CSV	50.750	-19.250	-12.375	-12.375	2.250	2.250		0.188	0.188	0.188	0.188	-0.188	-0.188	-0.188	-0.188
450	974	84-30-06.CSV	105.734	-4.266	-4.266	-4.266	-4.078	-4.078		-0.797	-0.797	-0.797	-0.797	-0.984	-0.984	-0.984	-0.984
451	1035	93-13-46.CSV	105.734	-4.266	-4.266	-4.266	-4.078	-4.078		-0.797	-0.797	-0.797	-0.797	-0.984	-0.984	-0.984	-0.984
452																	
453	19	39-10-01.CSV	105.992	-13.008	-9.633	-9.633	-7.008	-7.008		-1.055	-1.055	-1.430	-1.430	-1.055	-1.055	-1.430	-1.430
454	23	39-10-12.CSV	105.992	-13.008	-9.633	-9.633	-7.008	-7.008		-1.055	-1.055	-1.430	-1.430	-1.055	-1.055	-1.430	-1.430
455	20	39-10-02.CSV	105.992	-13.008	-9.633	-9.633	-7.008	-7.008		-1.055	-1.055	-1.430	-1.430	-1.055	-1.055	-1.430	-1.430
456	636	18-01-01.CSV	106.086	-11.914	-11.914	-11.914	-8.398	-8.398		0.977	0.977	0.977	0.977	-2.539	-2.539	-2.539	-2.539
457	436	86-32-01.CSV	106.086	-11.914	-11.914	-11.914	-8.398	-8.398		0.977	0.977	0.977	0.977	-2.539	-2.539	-2.539	-2.539
458	1037	93-13-48.CSV	106.766	-3.234	-3.234	-3.234	-3.234	-3.234		-1.078	-1.078	-1.078	-1.078	-1.078	-1.078	-1.078	-1.078
459	1005	93-13-16.CSV	106.766	-3.234	-3.234	-3.234	-3.234	-3.234		-1.078	-1.078	-1.078	-1.078	-1.078	-1.078	-1.078	-1.078
460	1021	93-13-32.CSV	106.766	-3.234	-3.234	-3.234	-3.234	-3.234		-1.078	-1.078	-1.078	-1.078	-1.078	-1.078	-1.078	-1.078
461	7	61-11-07.CSV	106.898	-9.398	-4.398	-5.102	-0.352	0.352		-0.273	0.273	0.273	-0.273	-0.273	0.273	0.273	-0.273
462	2473	61-11-07.CSV	106.898	-9.398	-4.398	-5.102	-0.352	0.352		-0.273	0.273	0.273	-0.273	-0.273	0.273	0.273	-0.273
463	8	61-11-08.CSV	107.734	-7.859	-2.859	-4.266	-1.891	-0.484		-0.078	0.078	0.078	-0.078	-0.078	0.078	0.078	-0.078
464	2474	61-11-08.CSV	107.734	-7.859	-2.859	-4.266	-1.891	-0.484		-0.078	0.078	0.078	-0.078	-0.078	0.078	0.078	-0.078
465																	

Figure 7. Sorted spreadsheet for the central Hokkaido database

3.2 An example of the detection of identical logs by spectra comparison

Figure 6 shows the sorted spreadsheet for the top 10m of log depth in the Noboribetsu database. 228 out of 1244 logs are detected as identical ones.

Figure 7 shows the sorted spreadsheet for the top 15m of log depth in the central Hokkaido database. In this case, 462 out of 5058 logs are detected as identical ones.

4 Summary

The numerization of soil names in drilling logs and the use of spreadsheet software make it easy to compare a large number of logs. Software such as Lotus 123 and Microsoft Excel allow us to detect identical logs from a maximum of about 65,000 logs at a time.

In the case of more over 65,000 logs, we split the data into groups and apply the sorting process to each group.

5 References

Hideyasu Asahi, Kiyoshi Wadatsumi and Kiyoji Shiono, 1995. Spectral Analysis of Numrized Soil Drilling-logs — Application of Spectrral Map to Estimation of Soil Condition —. *Geoinformatics*, 6, 1, 1-11.

Hideyasu Asahi, 2000, A Function Generator for Walsh Order Hadamard Matrix and Fast Walsh-Hadamard Transform. *Geoinformatics*, 11, 1, 3-9.