A COMPUTER ORIENTED GEODETIC DATA BASE; AND A NEW TECHNIQUE IN FILE SEQUENCING

G. M. Morton, Advisory Systems Engineer

International Business Machines Co. Ltd.,
150 Laurier Avenue, West,
Ottawa 4, Ontario,
Canada.

To the best of my knowledge, the documentation in my work entitled "A Computer Oriented Geodetic Data Base; and a New Technique in File Sequencing" is free of any proprietory, secret or confidential information belonging to a person or organization outside the IBM Company. Where I have used the work, plans, procedures, system, program or names of companies or individuals, I have obtained their permission to do so and a copy of their permission is attached.

G. M. Morton

G. M. Morton March 1st, 1966

ABSTRACT

Title A Computer Oriented Geodetic Data Base; and a New Technique in File Sequencing.

Author G. M. Morton

Date March 1st, 1966

Direct Inquiries to:

Mr. G. M. Morton,
International Business Machines Co. Ltd.,
150 Laurier Avenue, West,
Ottawa 4, Ontario,
Canada.

Phone - 613-236-0271

Abstract

A coordinate system for use with map or survey data is introduced and developed. Although this system can be applied to any sized area on any part of the Earth's surface, particular reference is made to Canada. The problems of using data spanning such a large area are discussed, and a new technique of file sequencing is introduced to counter these problems. Some of the properties of such a file sequence are used to demonstrate the flexibility and power to be gained by use of this technique.

TABLE OF CONTENTS

A Computer Oriented Geodetic Data Base; and a New Technique in File Sequencing

1.0 Introduction

1.1 The Canada Land Inventory

The project known as the Canada Land Inventory, or CLI, comes within the scope of the Agricultural Rehabilitation and Development Administration (ARDA). The purpose of the CLI is to gather, for some 600,000 square miles of Canada, information concerning the land and its use. This covers all land presently settled, and also land which can be developed for some useful purpose, such as forestry or recreation. This information will be used by ARDA for current and future planning in land resource allocation and development.

1.2 The Data

The data available to the CLI are currently in the form of map sheets. A complete set of map sheets covering the area of interest and portraying one aspect of the land is termed a coverage. The coverages which will be available will include present land use, soil suitability for agriculture, soil suitability for forestry, land suitability for recreation, political and administrative units. Most maps are drawn at a scale of 1:50,000, that is, slightly larger than 1 inch to the mile, which is equivalent to 1:63,360. Some idea of the amount of data with which the CLI is concerned can be gained by imagining one coverage extending over all of Canada. The maps for such a coverage would more than cover an acre! IBM Canada has been awarded the task of providing a computer solution to the gathering, storage, manipulation and retrieval of the data from these maps.

1.3 Data Input

The maps will be read on a unit called the Drum Scanner, which has been developed by IBM, Kingston. This device records the presence or absence of a line on a map every .004 inches. Since the maps to be scanned are at a scale of 1:50,000, then 1/250th of an inch on the map's surface represents 200 inches on the earth's surface. Hence any system used to store these maps should be capable of maintaining at least such resolution.

1.4 Resolution

Resolution of one point every 200 inches is not, however, sufficient. A map in a larger scale, say of 1:25,000 or 1:10,000 would require better resolution. So too, would better resolution be required if a scanner capable of scanning 1000 points to the inch were used. In another field, survey data requires far greater accuracy than can be obtained from 200 inches per point.

1.5 The Coordinate System Problem

The coordinate system chosen for the storage of maps or survey data must be one in which great accuracy can be maintained if needed. That great accuracy is not required at all times can be seen by observing that nobody draws climatological maps at a scale of 1:50,000! Because maps are drawn at different scales, the system should allow the storage of maps to be done at different scales, and still maintain flexibility in the manipulation of differently scaled data.

All map data within such a file must be kept in one projection and one orientation. Any projection of the surface of the Earth onto a plane results in some distortion. In some projections, a localized rectangular coordinate system allows tolerable, though not excellent, accuracy. These localized systems, which can seldom extend farther from their origin than 10 miles, have two serious drawbacks: one, adjacent systems do not mesh; and two, it is impossible to express accurately the position of a point in one system, with reference to the origin of another. In other words, any calculations which required either the crossing of coordinate system boundaries, or the referencing of data in another coordinate system would cause grave difficulties.

2.0 The Coordinate System

2.1 Theory of the Coordinate System

The use of a computer provides an answer to the problem. In the past, map projections have always been chosen such that the measurement of interest has not been distorted on the plane surface. Hence we have maps in which shape, direction, length or area are preserved and can be measured directly from the map (knowing its scale). The amount of accuracy lost in representing the map in a localized rectangular coordinate system is more than made up for by ease of measurement. In fact, when using actual maps, no other feasible means exist. When a computer is available however, accuracy can be maintained without loss of capability.

A simple but hard to find answer to the problem is developed in the paper by F. J. Jankulak entitled "A Reference Coordinate System for Computer Mapping." To quote from his conclusion - "The Geographic Coordinate System of parallels and meridians has been shown to be the best choice for use as a reference coordinate system for computer mapping." Since the length of 1° of either latitude or longitude varies by the latitude, the system is no longer rectangular, and the distance between two points on the grid depends on the position of the points within the system. (See Appendix 1, Fig 1). Local corrections for area and distance can be applied to correct for this distortion as needed. Exploration of the technique used for this correction and calculations demonstrating its accuracy can also be found in the above paper. It is sufficient here to state that these correction factors can be carried to any desired accuracy.

2.2 Implementation of the Coordinate System

2.2.1 The Computer

Best advantage can be taken of this system by the use of a binary computer, and for the purposes of this paper, such a computer is assumed. It is also assumed that the computer has a word length of 31 bits plus sign, and a half-word length of 15 bits plus sign. In no case will the sign position be used, so the half-word range is 0-32,767 and the full word range is 0-2,147,483,647.

2.2.2 Scope

Although the system can be applied to any sized area on any part of the globe, this paper will assume that the area of interest is the extent of Canada.

2.2.3 Axes

All of Canada must be maintained in the one coordinate system. We choose the total coordinate system as being M degrees square, where M is a power of 2, with any origin such that the X and Y axes (which will be a parallel and a meridian respectively) do not touch any of the land area of Canada. It is convenient that Canada be wholly contained in the first quadrant of the system (so that all points can have positive x, y values): an origin of 145°W, 40°N is suitable. Canada extends some 90° eastward of this origin, and since the next power of 2 greater than 90 is 128, each axis is 128° in length. (See Appendix 1, Fig. 2)

2.2.4 Extent

The total system (in this case) is thus defined to be all points within the first quadrant of a coordinate system whose origin (0,0) is at 145°W, 40°N and whose axes extend 128° from the point of origin along the meridian at 145°W and along the parallel at 40°N.

2.2.5 Unit Scale

Each point in the system must be representable by a number pair (x, y). Using full words for x and y, they can be in the range 0-2,147,483,647. Since each direction scans 128°, this allows for each point to be separated by slightly less than .00000006 of one degree. At 45° latitude, this, in actual distance on the Earth's surface, is just over one quarter of one inch in the meridional direction. This distance, expressed correctly as $(1/2^{24})^{\circ}$, is termed the unit scale.

2.2.6 Intermediate Scales

2.2.6.1 Reason for Intermediate Scales

Since it is obviously impractical to keep all (or perhaps any!) data at this scale, allowance must be made for other scales. As we have seen earlier, differently scaled maps call for different scales of data storage; moreover, the amount of data required to store a map should be kept at a minimum.

2.2.6.2 Data Units

From the data processing aspect, it is also imperative that the data be divided into logical units for processing; these logical units being of a size suitable for input to a computer. Later it will be seen how the total system is divided into these logical units or frames.

2.2.6.3 Allowable Intermediate Scales

Data within a single frame can be kept at any scale provided that the unit length in such a scale is a power of 2 times the unit scale length. In other words, the scale can be 1, 2, 4, 8, 16 or M times the unit scale where $M=2^s$ and s is any positive integer.

2.2.6.4 s-Factor

The value of s is termed the s-factor of a frame. The s-factor, of course, should not be of a size such that the unit length in the scale exceeds the bounds of the total system. Hence $0 \leq s \leq 31$, though at s=31 since there are only 4 points in the total system, the value of such a scale must be in doubt!

2.2.7 Flexibility of Scale Structure

Because of the binary nature of the scale, a point in one scale exists in each finer scale. Also, the nearest point in any coarser scale can be found readily by a simple rounding technique. Hence, if two adjacent frames have different s-factors, no inconvenience is encountered in going from one frame to another.

3.0 The Frame Structure

3.1 Frames

Since we are dealing - for the CLI - with areal data, it is logical that each record in the file should represent some unit of area. In keeping with the technique of the coordinate system, such records or frames will be "square" in shape and be enclosed by parallels and meridians.

3.2 The Unit Frame

What is the size of the unit frame? In the same way that each point was represented by a number pair, so each unit frame can be represented by a number pair. There is no need to have as many frames as points, so half-words rather than full words may serve. This would allow the unit frame to be $(128/32,768)^{\circ}$ on each side. This is just smaller than [illegible] of about 1,450 feet by 1030 feet. At the finest [illegible] would be a total of 2^{32} grid points within this frame. [illegible] however, or any scale approaching it, discrete points [illegible] 1:50,000 scale maps read on the drum scanner can be stored with an s-factor of 9; and at this scale, the unit frame consists of 2^{14} or 16,384 points only. This, then, is a satisfactory unit frame size.

3.3 The Largest Frame

It is conceivable that the amount of data required for a particular coverage can be so sparse as to allow all the data for Canada to fit into a single record. An example of such data is the location of each weather station across Canada. In this case, one frame (with any required s-factor) would consist of the total system.

3.4 Intermediate Frames

From the unit frame to the largest frame is quite a step, so intermediate frame sizes must be allowed. In a way analogous to the allowable scale factors, we let frame sizes take on the values of U x $(2^f)^2$ where U is the unit frame size and f is an integer called the f-factor. f must be in the range $0 \leq f \leq 31$. An f-factor of 0 denotes the unit frame, an f-factor of 31, the total system. Each frame is made up of four frames of the immediately preceding frame size. If a frame is of an intermediate size, then its location within the system must be such that, if the total system were divided, starting at the origin, into frames with the same f-factor then the frame in question would coincide exactly with one of the frames so drawn. Figure 3 in Appendix I shows a frame of f-factor 5 which has been further divided into other frames. The f-factor of each frame is noted.

4.0 Frame Sequencing

4.1 The Frame Sequencing Problem

It was stated above that each unit frame could be represented by a number pair (x, y). This identification scheme has two serious drawbacks. Firstly, when placed in identification sequence, of, for example, all x within y, the result is very narrow strips of data from coast to coast. Therefore the location of data which are close together on the Earth's surface could be far apart in the file. Whether the file is on tape or disc, this is a disadvantage. Secondly, the numbering scheme is interrupted whenever different f-factor frames are encountered.

4.2 The Frame Sequencing Requirement

Regions of the Earth's surface are the units by which information is required. Frame numbering and sequencing must bear with such a philosophy. Data on areas close together on the Earth's surface must as much as possible be close together in the file, and the sequencing technique should have flexibility enough to allow for intermediate frame sizes.

4.3 The Frame Sequencing Solution

4.3.1 Basic Sequence

To begin with, suppose the total system to be divided into unit frames. From the origin, the first four frames are numbered thus:-

1	3
0	2

0, 0

The next four frames to be numbered bear the same relationship, as a whole, to the previous four frames as frame 1 bore to frame 0, viz:

5	7
4	6
1	3
0	2

Similarily, the next four, and the next four after them, maintain this relationship.

5	7	13	15
4	6	12	14
1	3	9	11
0	2	8	10

In other words, each frame is put together by using frames with the next lower f-factor, in the same manner as the first four unit frames were put together to form the first f-factor 1 frame. Appendix 1, Figure 4, shows the position of the first 256 unit frames. How are intermediate frames identified? They have the same number as the unit frame which would have been in the upper right hand corner of the intermediate frame, if such a unit frame had existed. Consider frame 15. This could be the number of a frame with f-factor 0, 1 or 2. To avoid ambiguity, the f-factor of each frame is appended to the frame number. Hence 15-0, 15-1 and 15-2 refer to three different frames. A file kept in frame number sequence meets the requirements of map data retrieval.

4.3.2 Area Requirement Met

Generally speaking, areas are contained in frames whose sizes are proportional to the size of the areas-large areas, large frames-small areas, small frames. Hence most areas will be contained in a single frame, nearly all the remainder in two frames, some in three or four frames, and a very few in more than four frames. For single frame areas, one sequence is as good as another. For areas contained in two or more frames analysis of the technique shows that the probability of requiring arm movement if the file is on a DASD is less than with other techniques (See Appendix 2).

4.3.3 Flexibility Requirement Met

More important is the flexibility afforded by this sequencing technique. A file in frame sequence retains this sequence regardless of the frame sizes. For example, all the unit frames contained within an intermediate frame occupy the same position in the file relative to other frames as does the intermediate frame. Put in other words, all data belonging to a frame of any size appear in the file, contiguously, regardless of current frame sizes within the file. Also, if a frame must be subdivided into smaller frames, no unnecessary dislocation of data takes place.

5.0 Other Properties

5.1 Relationship Between Unit Frame Numbers and Normal Coordinates

It is necessary to convert unit frame numbers into the number pair (x, y) so that latitude and longitude of individual frame origins can be determined. The first four unit frames (numbers 0, 1,2 and 3) expressed in binary are 00, 01, 10 and 11, and the origins of these frames are 0, 0; 0, 1; 1, 0 and 1, 1. (Here we are referring to points in a frame system, not points within our total system of geodetic coordinates. To arrive at the latter, the "frame" coordinates must of course be multiplied by 2^{16}). Because of the method of construction, by which this positional relationship is propogated through successively higher frame sizes, the correspondence between the binary representation of a frame number and its origin holds good for larger frame numbers. The rule can be stated as follows:-

(a) Express the frame number in an even number of binary digits.

(b) Starting with the leftmost bit, write down every alternate bit.

(c) Starting with the second leftmost bit, again write down every alternate bit.

(d) The number written in (b) is the binary representation of the x coordinate, in (c) the binary representation of the y coordinate.

For example, what is the origin of frame 50? 50 in binary is 110010 which "splits" into 101 and 100 or 5, 4 which is indeed the origin of frame 50.

5.2 Origin of Intermediate Sized Frames

To arrive at the origin of a frame with f-factor other than zero, we first find the unit frame which has the same origin. To do this, simply subtract $(2^f)^2 - 1$ from the intermediate frame number, where f is the f-factor.

5.3 Finding the Number of a Contiguous Unit Frame

When processing data within one frame, quite often adjacent frames must be referenced. If we are dealing with unit frames, the finding of adjacent frame numbers can be accomplished as follows. The unit frame origin is found as above. One is either added or subtracted (depending on direction) to either one or both of the coordinates of the origin (again depending on direction). The coordinates of the new origin, which is of the required frame, are then "merged" together to form the required frame number.

5.4 Contiguous Intermediate Frames

Suppose we are following a map line in an intermediate frame when the frame border is reached. We wish to follow the line further. But we do not know the frame size of the neighbouring frame and hence we do not know its number. However, we can calculate the unit frame number which would contain the desired line. If this unit frame does not exist, then the frame with the next higher number contains the desired data! It is for this reason that intermediate sized frames are numbered by the highest rather than the lowest unit frame within them. Otherwise the frame with the next lower number would have contained the data, which is contrary to normal file searching practices.

5.5 Independance on File Size or Extent

This technique in file sequencing is not dependant on either file size or extent. The rules for sequence number manipulation are the same regardless of whether the file represents Canadian data or Scottish data. Also, similar technique is used for very dense and very sparse data.

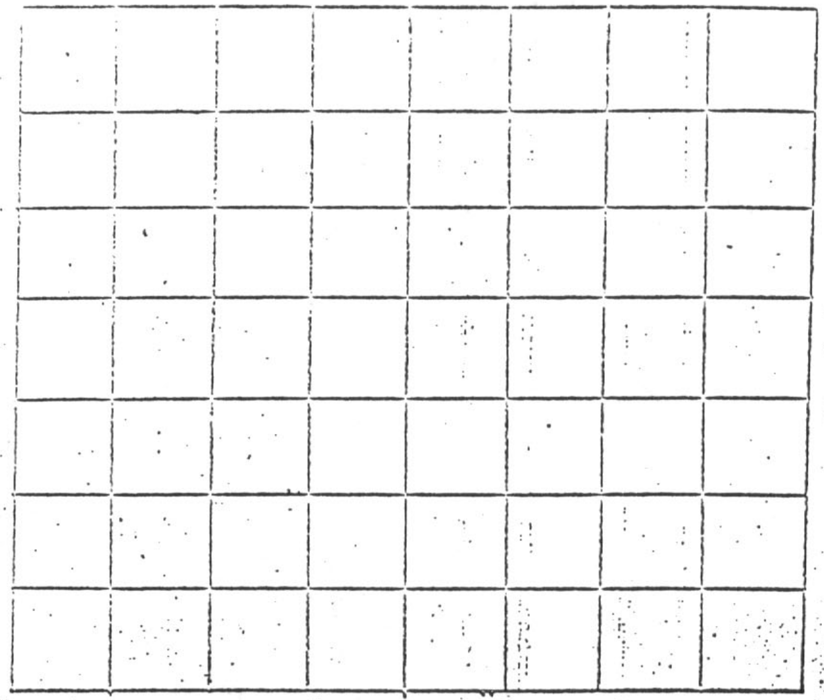
5.6 Geographic Coordinate and Frame Number

The unit frame to which a geographic coordinate belongs is readily discernible. The first half-words of the x and y coordinates respectively of the point are the x and y "frame" coordinates of the unit frame to which the point belongs.

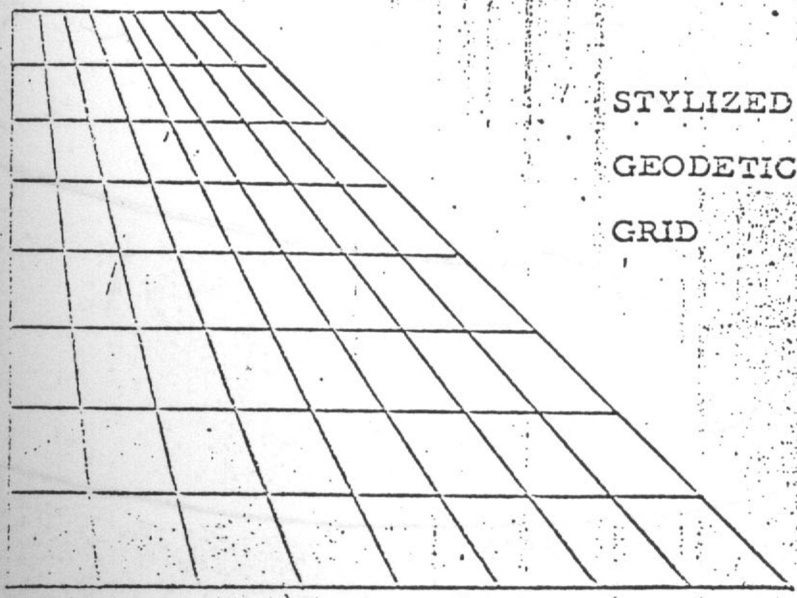
6.0 Conclusion

Both the geographic coordinate system and the frame structuring and sequencing techniques are admirably suited to the processing of areal or other geographic data.

Appendix I Figure 1



RECTANGULAR GRID



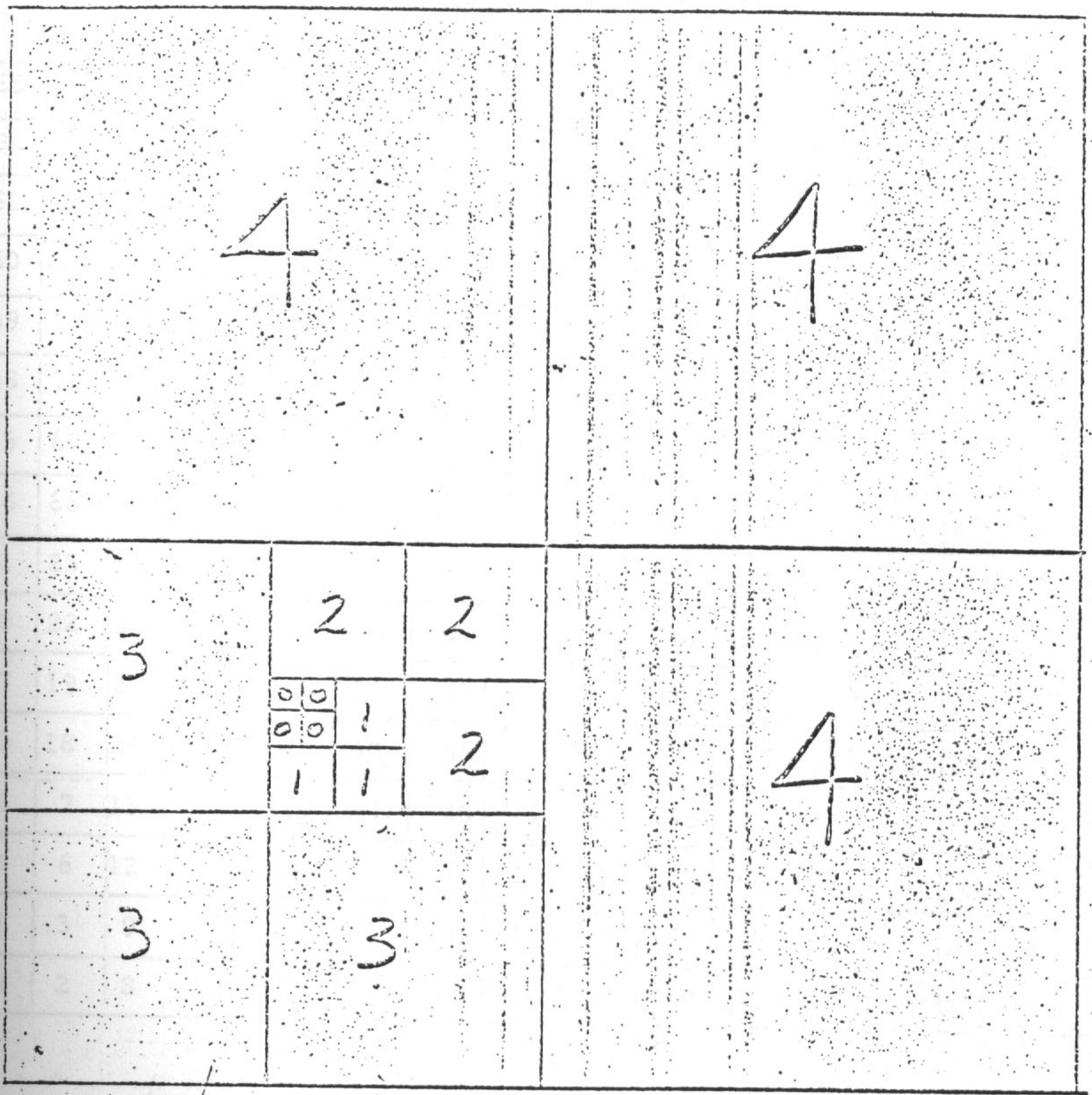
STYLIZED
GEODETIC
GRID

Appendix 1 Figure 2

145°W
55°W
40°N

CANADA

Appendix I Figure 3



AN f - FACTOR 5 FRAME

10
11
100

Appendix 1 Figure 4

11 22

12 12

85	87	93	95	117	119	125	127	213	215	221	223	245	247	253	255
84	86	92	94	116	118	124	126	212	214	220	222	244	246	252	254
81	83	89	91	113	115	121	123	209	211	217	219	241	243	249	251
80	82	88	90	112	114	120	122	208	210	216	218	240	242	248	250
69	71	77	79	101	103	109	111	197	199	205	207	229	231	237	239
68	70	76	78	100	102	108	110	196	198	204	206	228	230	236	238
65	67	73	75	97	99	105	107	193	195	201	203	225	227	233	235
64	66	72	74	96	98	104	106	192.	194	200	202	224	226	232	234
21	23	29	31	53	55	61	63	149	151	157	159	181	183	189	191
20	22	28	30	52	54	60	62	148	150	156	158	180	182	188	190
17	19	25	27	49	51	57	59	145	147	153	155	177	179	185	187
16	18	24	26	48	50	56	58	144	146	152	154	176	178	184	186
5	7	13	15	37	39	45	47	133	135	141	143	165	167	173	175
4	6	12	14	36	38	44	46	132	134	140	142	164	166	172	174
1	3	9	11	33	35	41	43	129	131	137	139	161	163	169	171
0	2	8	10	32	34	40	42	128	130	136	138	160	162	168	170

5 6

THE FIRST 256 UNIT FRAMES

APPENDIX 2.

Given an n x n matrix, whose elements are sequenced as in the map file sequencing technique, it is desired to find Px, the probability that an element immediately above or to the right of another element, differs in sequence number from it by less than some other given number, x. For example, in the 16 x 16 matrix shown in Appendix 1, Figure 4 what is the probability that adjacent (as defined above) elements are less than 16 apart?

In the north (or upward) direction, each element differs from its predecessor by either 1, 3, 11 or 43. In the general case, the difference is , for l = 1, 2, -----

$$\frac{(2^l)^2 + 2}{6} \qquad (1)$$

In the east (or right) direction, the differences are 2, 6, 22 or 86, which generalizes to

$$\frac{(2^l)^2 + 2}{3} \qquad (2)$$

What possible differences are less than the given number x? l must satisfy, and (l + 1) must not satisfy, the inequalities - for north and east respectively -

$$x > \left[(2^{l_1})^2 + 2\right]/6 \qquad (3)$$

$$x > \left[(2^{l_2})^2 + 2\right]/3 \qquad (4)$$

l_1 and l_2 are shown since they need not necessarily coincide. From (3) and (4) is derived

$$l_1 = \left|\left[\log_2(6x-2)\right]/2\right| \qquad (5)$$

$$l_2 = \left|\left[\log_2(3x-2)\right]/2\right| \qquad (6)$$

Further examination of the sample matrix shows that there are 128 differences of 1, 64 differences of 3, 32 differences of 11 and 16 differences of 43. From this we derive the series for the actual number of elements with particular differences 1, 3, 11 etc. This series is

$$\frac{n^2}{2} + \frac{n^2}{4} + \frac{n^2}{8} \cdots\cdots + \frac{n^2}{2^j} + \cdots$$

or

$$n^2\left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots\cdots + \frac{1}{2^j} + \cdots\right) \qquad (7)$$

The same series gives the number of elements in the easterly direction. Hence the total number of differences less than x is given by

$$n^2\left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots + \frac{1}{2^{l_1}}\right) + n^2\left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots \frac{1}{2^{l_2}}\right) \qquad (8)$$

where l1 and l2 satisfy (5) and (6) respectively. This expression simplifies to

$$n^2\left[\left(1 - \frac{1}{2^{l_1}}\right) + \left(1 - \frac{1}{2^{l_2}}\right)\right] \qquad (9)$$

or

$$n^2\left(2 - \frac{1}{2^{l_1}} - \frac{1}{2^{l_2}}\right) \qquad (10)$$

Since there are a total of 2n (n - 1) differences, this gives, for P, the function

$$P_x = \frac{n^2\left(2 - \frac{1}{2^{l_1}} - \frac{1}{2^{l_2}}\right)}{2n(n-1)} \qquad (11)$$

or

$$P_x = \frac{n}{n-1}\left(1 - \frac{1}{2^{l_1+1}} - \frac{1}{2^{l_2+1}}\right) \qquad (12)$$

Larger n implies larger l_1 and l_2 hence both $\frac{n}{n-1}$ and $\left(1-\frac{1}{2^{l_1+1}}-\frac{1}{2^{l_2+1}}\right)$ tend to unity as n increases. Hence for large n, Px tends to unity, or near certainty.

If we take as an example the sample matrix and x = 16, we
find P_{16} to be 13/15. If the matrix were arranged in the conven-
tional strip manner, then this probability would be 1/2. Hence the
map file sequencing technique gives better probability for less movement.

Similar analysis shows the same result for 3-frame areas, 4-frame areas, etc.