STRATEGIES FOR SPATIAL DATA COMPRESSION *

Keith C. Clarke
Hunter College
New York, NY 10021

ABSTRACT

As part of the research on remote workstations and networking for NASA's Land Data Pilot Information System the issue of data compression is being investigated in the context of spatial data. Spatial data compression can be logical, physical, information retaining or information reducing, and often requires different strategies than non-spatial data compression. Several different compression methods are considered in the context of polygonal land use data, and a test system is described which uses these methods to send data between a mini and a microcomputer. Spatial data compression is shown to have a number of potentially important applications in automated cartography.

INTRODUCTION

Digital data bases have increased in size at a remarkable rate since the emergence of the Geographic Information System as a management and research tool. This increase in data volume is the result of two forces; the conversion of the immense reserve of analog map data into digital form, and the new, usually higher resolution sensors and mapping instruments now coming into use. The transition of remotely sensed data, for example, from Multispectral scanners to the Thematic mapper to the French SPOT sattelite, and in the future to the Shuttle Imaging Spectrometer, represents a data density increase from 6.25 to 77.78 to 100.00 to 1422.2 bytes per hectare on the ground, a 227 fold increase.

Faster computers and new mass storage devices have alleviated the problem to some extent, but at many stages in the processing of cartographic information bottlenecks exist where the rate of data flow through communications links can fall to 1200 baud and less. Examples of these bottlenecks are magnetic tape and floppy disk input/output operations and communications between nodes in a network. One solution to the problem of low speed data transmission is the use of data compression. Data compression methods have long been used in the handling of textual and other non-spatial data, but few spatial data compression techniques are used on a regular basis. One notable exception is in image processing, where compression of array-type data is well understood and frequently used.

---

* Funds for the support of this study have been allocated by the NASA-Ames Research Center, Moffett Field, California, under Interchange No. NCA2-1R305-401

This paper examines the potential for data compression in spatial data processing, and considers the suitability of different data structures for compression. Finally, a test system for compression of polygonal data is described, as well as some of the algorithms behind the compression methods implemented in the system.

THE WORKSTATIONS CONCEPT

The National Aeronautics and Space Administration, as part of the Land Data Pilot Information System, is investigating the development of a computer network to provide data links between the NASA Technical centers, NASA headquarters, other NASA units, and principal investigators on land-based NASA research projects. Similar projects are already under way for the climatic, planetary, and oceanographic data from the manned and unmanned space programs. One of the problems specific to the Land Data Pilot Information system is the broad variety of needs for system users, and the huge volume of land-related data available.

The network has no shortage of computing power, but the major problem is seen as one of access at the local level, since investigators are geographically dispersed and may have a limited computing environment. At the users end, the working arrangement is assumed to be a microcomputer-based workstation, with limited storage capability (less than 12 megabytes) perhaps on a hard-disk, minimal graphics capability, and other assorted peripherals. As a standard at the node end, the VAX 11/780 has been suggested, with VAX VMS or Unix as the operating system. This favors microcomputers with Unix-like operating systems, or VAX VMS communications capabilities.

The communications link for the workstation would normally consist of microcomputer software which is commercially available and uses telephone lines at 300 or 1200 baud. Several packages exist, including the KERMIT system (DaCruz and Catchings, 1984). The standard communications link is asynchronous serial, using the RS-232C EIA standard, and may or may not involve timing, packet switching, buffering, or error-checking within the protocol.

Given these low transmission speeds, two data problems exist. First, users require the ability to browse through directories and get previews of selected data and data subsets. Secondly, when data have been selected, they have to be transmitted from the main node to the workstation, preferably with minimal error. It is the first problem to which the current research is addressed. However, this problem necessarily involves considering spatial data structures and how data volume reduction and data compression can be used to facilitate a preview capability for workstations.

DATA VOLUME REDUCTION AND DATA COMPRESSION

In computer science, data compression is subdivided into logical and physical. Logical data compression using spatial data involves changing the geocodes to minimize redundancy. An example might be

leaving off the first few digits of a U.T.M. map reference to limit the number of bytes necessary to represent the eastings and northings. The U.S.G.S., in its Digital Land Use GIRAS files, uses an arbitrary coordinate system based on a local origin to achieve the same goal. Into this category of logical compression also might fit the actual way that the digital map data are converted to computer readable files, perhaps using different file structures (random-access, sequential, etc.), or digital representations (ASCII, EBCDIC, binary).

Physical data compression involves an alteration in the logical data structure to reduce the total number of bytes needed to represent the data set. An example might be the conversion between vector and raster data structures, or the conversion of spatial data to a volume reducing transform.

Cartographic data does not fit neatly into these categories due to the added complications of data resolution and data dimensionality. In the context of textual data, where sequence is critical, leaving out the redundant blanks, punctuation and every n-th letter would reduce the information content to zero. Yet in the spatial context, point elimination is a necessary and valid way of reducing the volume of digitized line data, and a large number of point elimination methods have been devised (McMaster, 1983). Data volume reduction in a spatial context is equivalent to cartographic generalization. Generalization normally occurs when either of two processes take place, either data are sampled and/or data are reduced in dimensionality or scaling. Dimensionality involves the traditional division of cartographic data into point, line, area, and volume data. Scaling involves the type of measurement, either nominal, ordinal, interval, or ratio.

In terms of data transmission, generalization can be thought of as information loss. In the spatial context we can, and indeed must, generalize to clarify map information, so generalization may aid in the communication of cartographic information. This gives an increased flexibility to spatial data compaction, especially since we need only be concerned at this stage with a "preview" or "quick look" capability, which can be achieved with a high degree of generalization.

For this purpose, we will categorize spatial data compression as information loss compression and information retention compression. In many cases, similar techniques may be used for each, and therefore may be of use in both preview and full data transmission cases. The distinction is that information retention compression allows full reconstruction of spatial data after compression while information loss compression may preserve only spatial relations or generalizations.

The volume of spatial data may be considerably reduced by information loss compaction. However, this remains as physical compaction since no change has taken place in the actual data structure. Data structure is taken to mean the abstract representation of the cartographic objects and their attributes, which may or may not have an impact on the physical storage characteristics of the data. Physical spatial data compression, therefore, may be defined as an alteration in

the abstract representation of cartographic objects and their associated attributes, with the intent of reducing the total number of bytes needed for their physical digital representation.

POLYGONAL DATA COMPACTION

To investigate the problems associated with these various types of compaction, a single type of spatial data was considered in isolation. Since large quantities of data are encoded in the polygon, arc, node system (Fegeas et al., 1983) the basic unit of a plane polygon was selected. The polygon in this context is a nominal areal feature and is assumed to be a vector chain representation of a set of non-overlapping, contiguous, closed polygons which may or may not contain holes.

To contrast information loss/retention and physical/logical data reduction, consider a polygon consisting of three chains (figure 1). To logically compact this file, we could rewrite the file using Huffman coding (Rosenfeld and Kak, 1982; Held, 1983). This technique uses an adaptive compaction method to reduce the file size. The ASCII file containing the chain and chain relational (linkage) data is analyzed in terms of its content. If the file contained 50% ASCII code 32 (blank), the probability associated with encountering this code in the file would be 0.5. Similarly, if we had only one negative sign (-) in the file, perhaps representing a counter-clockwise chain link, and the file was one kilobyte in size, the probability of finding a negative sign (ASCII code 45) would be 1/1024.

Huffman coding starts by determining the lowest of these probabilities, and links this with the next lowest, summing the two probabilities to form a hierarchical link for each step (figure 2). This process continues until all possible values are linked and the total probability is unity. The hierarchical tree is then followed, coding a left turn as a one and a right turn as a zero. The string of digits leading to each ASCII code is then its Huffman code. The codes are arranged so that the lowest number of bits are associated with the most frequent ASCII codes and vice versa. A key property of the Huffman code is that the codes can be instantaneously translated upon reception by matching with the possible codes. This process uses the hierarchical tree backwards, with each bit leading the decoding closer to the actual ASCII code.

Some operating systems contain system utilities for adaptive Huffman coding, for example the "compact" command in Unix. Regular use of Huffman coding can greatly reduce the storage needed for files, and when both a transmitting and receiving network node have the compact and decompact utilities the data transmission bottleneck can be avoided to some extent. It may be that a more spatial Huffman coding could be developed, to take advantages of particular spatial data properties for example. Freeman codes for land use polygons in a north-south oriented township and range system, where fields follow the lines, may be overweighted in the orthogonal rather than the diagonal axes, and therefore could be compacted efficiently with Huffman codes. It is notable that

Figure 1: Types of Spatial Data Compression

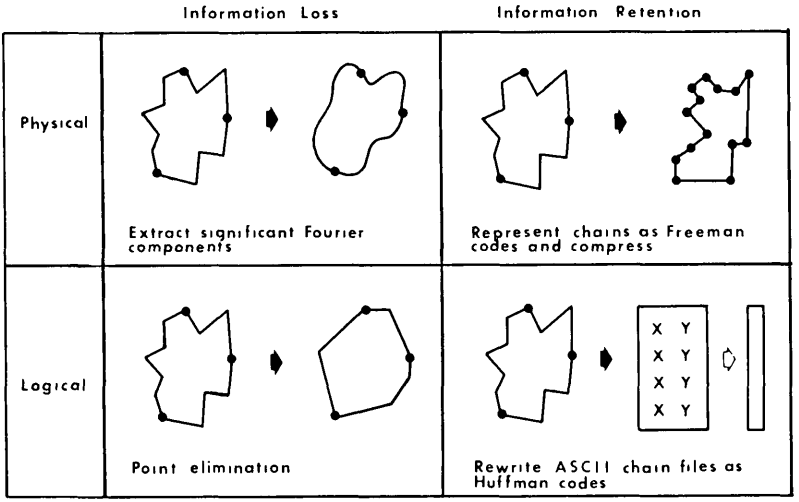
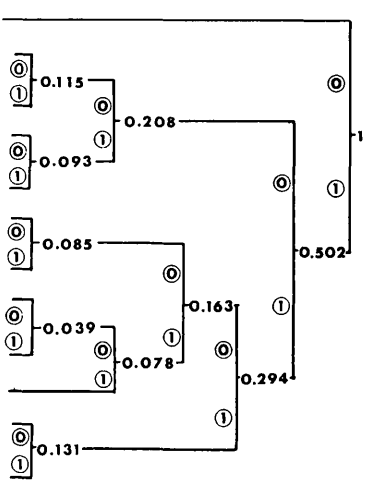


Figure 2: Huffman Coding

ASCII Code	Huffman Code	#times in file	Probability
32( )	0	512	0.500
50(2)	1000	57	0.056
55(7)	1001	60	0.059
52(4)	1010	50	0.049
54(6)	1011	45	0.044
56(8)	11000	43	0.042
57(9)	11001	44	0.043
53(5)	110110	39	0.038
45(-)	110111	1	0.001
49(1)	11011	40	0.039
51(3)	1110	62	0.061
43(0)	1111	72	0.070

Links [code1 for left branch, 0 for right]



the less random the data, the more useful the Huffman coding is for data reduction, since most spatial data are far from random.

An alternative (figure 3) is to convert the chains to Freeman codes (Freeman, 1974). To some extent this represents a resolution change, and is equivalent to a vector to raster conversion, since the Freeman code has a finite resolution. The Freeman codes can be represented in three bits, and so may produce a volume reduction. However, if the fourth bit is used as a flag, a zero flag can be used to mean a regular Freeman code, while a one flag can be used to signify a repetition of the previous code by the number of times represented in the next three bits. This means that sequences of seven identical Freeman codes can be reduced to eight bits, as opposed to the twenty-one bits necessary without compaction. Huffman and Freeman coding are information retention compression methods, though by decreasing the resolution (increasing the length of the vectors) Freeman coding can be made information reducing.

Logical information loss methods include line smoothing. It is possible to use a variety of methods, each of which uses different criteria for point elimination. A crude method is to leave out every n-th point. More sophisticated techniques consider points as "significant", (i.e. high information content) on the basis of the angle between successive segments or the displacement from a straight line between chain end points. McMaster (1983) contains a review and empirical evaluation of these methods.

Reducing the number of points in a chain is equivalent to generalization, just as the cartographer smooths a coastline by eye to reduce minor intricacies. It is also equivalent to the generalization produced by changing to a smaller map scale. If this scale change retains self-similarity, i.e. if the basic shape of the chains is repeated in smaller and smaller sub-segments as we go to larger scales, then it may be possible to invert the point elimination process by fractal enhancement. Dutton (1981) has proposed such a method. Although point elimination and inversion by fractal enhancement does represent data compression, and may indeed produce maps which are perfectly adequate for preview purposes, the final information is not the same as the original, and this method would have to be classified as information-loss compaction. In addition, not all types of lines on maps show self similarity at different scales.

An important point to note is that when the data are chain encoded, the end points of chains should not be included in the elimination process. This means that for a chain containing N points and N-1 segments, elimination of every n-th point will eliminate (N-1)/n points. For large numbers of points, the figure of merit for the compression (defined as the ratio of the length of the compressed to the length of the original data string) is approximately 1 - 1/n.

Physical information loss compaction techniques are numerous, but not all of them are invertible. Shape parameterization, such as the computation of form ratios, attempts to derive descriptive values for

Figure 3: Freeman Codes

Vector	Freeman Code		Index in array[x][y]
	decimal	binary	
→	0	0 0 0	x+1 , y
↗	1	0 0 1	x+1 , y+1
↑	2	0 1 0	x , y+1
↖	3	0 1 1	x-1 , y+1
←	4	1 0 0	x-1 , y
↙	5	1 0 1	x-1 , y-1
↓	6	1 1 0	x , y-1
↘	7	1 1 1	x+1 , y-1

Note: Code has resolution of one unit in x,y but $\sqrt{2}$ units diagonally.

Table 1: Compaction Estimates
Sample Land Use Data

	GIRAS format	Polygon List
Full Data Set	137,579 bytes	211,680 bytes
Point Elimination n=2	95,239 (69.2%)	118,520 (56.0%)
Huffman Coding	81,227 (59.0%)	124,976 (59.0%)
Freeman Code Compaction	86,140 (62.7%)	107,163 (50.6%)
Relations only + nodes	43,158 (31.4%)	-------
Fourier Abstraction	-------	83,700 (39.5%)

polygons, which are often single scalar values. Boots and Lamoureaux (1972) provided an annotated bibliography on these measurement methods. None of these methods involves inversion, so that although a polygon is reduced to a single real value, there is an incomplete mapping of this value back onto a particular shape. In addition, the parameters are designed to be independent of the size, location and orientation of the polygon, and these are all properties which a reconstruction of the polygon needs to preserve.

Considerably more information is contained in the end point and chain linkage information for the polygons. Regeneration of simple polygon networks from this data alone, equivalent to point elimination with $n = 1$, often can yield a satisfactory preview map. Topologically complex sets of polygons, however, are not well represented. Islands, for example, close onto points, and concave polygons may have crossed segments. One experiment, using the boundaries of the Departements of France, was able to reproduce a polygon set from contiguity and spacing information alone (Kendall, 1971) and may have potential as a method for spatial data compression.

One reversible shape-based technique was adapted from the work of Moellering and Rayner (1979), who considered a series of digitized points in terms of two real series and their Fourier transform. The real series constitutes the x and y values of the polygon outline, and should be of even spacing. In the transform, it is possible to extract the harmonics which contribute most to a polygon's shape, and to save the Fourier coefficients for these harmonics. An approximation of the polygon outline can then be reconstructed from the Fourier coefficients. Depending of the complexity of the polygon, comparatively few, sometimes as few as four or five, harmonics can produce a very good approximation of the polygon.

While the complex Fourier series representation is more elegant, the working approach taken here was to perform two separate Fourier transforms, one for the x and one for the y axis. The discrete Fourier transform was used, since the limitations on the string length imposed by the Fast Fourier transform algorithm was found to produce overflow into the padded part of the transform. The inverse transform involves summing over a reconstructed series, with zeros for non-significant Fourier coefficients, and the series length being passed as part of the compressed data.

TEST DATA SET

The above mentioned data compression methods are currently being tested on a U.S.G.S. digital land use data file, at the scale of 1 to 250,000. Section 4, a 20 minutes of latitude by 36 minutes of longitude area in south-western Massachusetts was used, since this area contained the smallest number of land use polygons of the 15 for the whole one degree by two degree quadrangle. The GIRAS data file for this area contains 465 land use polygons, consisting of 1,260 arcs or chains and 10,161 coordinate pairs.

The GIRAS files contain several non-essential attributes, such as the x and y ranges of each arc. Without these attributes, the GIRAS file in ASCII random access format, blocked to 80 characters per line with line feed characters, constitutes 137,579 bytes. Based on a simple topology, and leaving out the double-counted interior island polygons used within GIRAS, this file can be rewritten as closed polygon lists in 211,680 bytes.

Some preliminary and expected compactions and their figures of merits are listed in table 1. The figures of merit are stated as percentages of original data set size in bytes. For the GIRAS data, a low of 31.4% derives from the elimination of all points along each arc. The implications of this structure for data display make this a preview option only. Similarly, the 39.5% expected figure of merit for the Fourier abstraction method may produce gaps and slivers along the edges of the reconstructed polygons.

In the case of Freeman code compaction, a byte oriented approach was taken. A C-language program named SQUEEZE reads Freeman code files and writes sequential ASCII files. This strategy loses some of compaction of the bit-based methods, but allows a code to be repeated nine times instead of seven. Preliminary tests have given figures of merit of about 50%, and the sizes of the Freeman coded data sets have been approximately the same as the polygon list sets at the same resolution.

The Fourier abstraction method is still undergoing testing to determine the optimal number of spatial harmonics necessary for regeneration of polygon shapes. A single harmonic in x and y requires the storage of at least two pointers to the harmonics, and two real values. Including identifier and attribute information takes the number of bytes necessary to store a polygon with five harmonics up to 180. Five harmonics were used to generate the estimates, and no more than ten may be necessary to rebuild the polygon quite precisely. Finally, no algorithm has yet been devised to build polygons from relational information.

TEST CONFIGURATION

The test data sets reside on a VAX 11/730, runnning the Unix operating system. To simulate a network, the VAX is connected to a SAGE IV microcomputer running the IDRIS operating system, a Unix look-alike. Considerable differences in the two Unix versions have made the Unix "cu" (call Unix) command inoperative. Alternate communications software is currently being installed. The link will be by 1200 baud modem and telephone line, although the two computers are actually in adjoining rooms. Final displays will be produced on a pen plotter, with the polygon list as a destination format. Software for both the VAX and the SAGE is being written in the C-language, using portable calls so that transportation problems to other systems are minimized.

The figures of merit for the data compressions are only one of three sets of test figures being compiled. The other two are

communications errors and compaction/ reconstruction processing time. File transfer time will be assumed proportional to file size in bytes.

DISCUSSION

For networking, a preview capability may favor information loss compression, with the benefit of not having to transform between data structures. How effective a highly generalized cartographic image is is likely to determine the method chosen for the Land Data Pilot or any other compaction system. Fourier abstraction and reconstruction from relational data offer high degrees of compaction, but the "fuzzy" images they produce should be regarded as preview or throw-away only, and will not satisfy cartographic purposes.

Certainly, Huffman coding offers important savings in transmission time and storage at all levels. However, this method depends upon the compatibility of different operating systems, something which only occasionally exists. A double compaction, perhaps a combination of Freeman code and then Huffman compaction may be a highly desirable combination. Also of importance in the communications link is processing time. A suitable system should concentrate processing at the VAX end of the transmission, minimizing the work load on the slower workstation.

CONCLUSION

Clearly much work remains to be done, and the Land Data Pilot research is in its infancy. Already apparent, however, is the potential for applications of data compaction in the spatial sciences, where data sets are typically very large. In analytical cartography, the emphasis has often been upon the elegance of a data structure rather than its storage requirements. With such a rapid increase in the volume of spatial data, and the need to integrate these data into Geographic Information Systems, cartographers and spatial scientists must begin to incorporate this thinking into new systems. In particular, as the volume of archived, general purpose spatial data begins to expand, the need to compress spatial data will become self evident. Rather than waiting for the day when the sheer volume of data is overwhelming, cartographers should anticipate the need, and become involved in investigating the relative merits of the different strategies for spatial data compression.

REFERENCES

Boots, B. and Lamoureaux, L. (1972) Working Notes and Bibliography on the Study of Shape in Human Geography and Planning , Exchange Bibliography 346, Council of Planning Librarians.

DaCruz, F. and Catchings, B. (1984) "KERMIT: A file transfer protocol for universities", Byte , June 1984, 255-278.

Dutton, G. H. (1981) "Fractal enhancement of cartographic line detail", The American Cartographer , 8, 23-40.

Fegeas, R. G. at al. (1983) Land Use and Land Cover Digital Data Geological Survey Circular 895-E, Reston, Va., U.S.G.S.

Freeman, H. (1974) "Computer processing of line-drawing images", Computing Surveys , 6, 1, 57-97.

Held, G. (1983) Data Compression , New York, J. Wiley.

Kendall, D. (1971) "Construction of maps from odd bits of information", Nature , 231, 158-159.

McMaster, R. B. (1983) "A mathematical evaluation of simplification algorithms", Proceedings, AUTOCARTO 6 , Ottawa, Canada, 2, 267-276.

Moellering, H. and Rayner, J. N. (1979) Measurement of Shape in Geography and Cartography , Numerical Cartography Laboratory, Ohio State University, Columbus, Ohio.

Rosenfeld, A. and Kak, A. C. (1982) Digital Picture Processing , 2, Computer Science and Applied Mathematics, New York, Academic Press.