OPTIMIZING RASTER STORAGE: AN EXAMINATION OF FOUR ALTERNATIVES

Michael F. Goodchild
Andrew W. Grandfield
The University of Western Ontario
London, Canada N6A 5C2

ABSTRACT

The performance of many algorithms in spatial data processing depends on the way in which spatial entities are ordered. The paper defines a general class of problems in which the objective is to preserve as far as possible the spatial relationships present in two dimensions. Applied to a raster, the problem leads to the N or Morton ordering and a new Pi-order is proposed based on a familiar Peano curve. An algorithm is given for defining Pi-order. These are compared empirically and analytically to the conventional row ordering, and a simple variant, using a number of standard images and using a class of indices which includes the spatial autocorrelation measures. The empirical results support Pi-order but the analytic results are mixed.

INTRODUCTION

Consider a set of spatial entities which are to be stored and processed in a spatial data system. In this paper the entities considered are the cells of a square raster, but they might equally be a set of non-overlapping polygons representing census tracts or counties. The order in which the entities are stored and processed is critical in many types of queries and sorting operations. The purpose of this paper is to consider the implications of such an order for storage, manipulation and output. The paper discusses various ways of defining an optimum ordering, indices for evaluating specific orderings, and algorithms for transforming one order into another.

Let the number of entities be n, and let r_i denote the position, or rank, of entity i in the ordering. The vector $\underline{R}$ with elements r_i, i=1,n must be a permutation of the integers 1 through n, and clearly n! such permutations are possible. Now consider a square n by n matrix $\underline{W}$ whose elements w_{ij},i=1,n,j=1,n represent the proximities or spatial relationships between each pair of entities. For example we might let w_{ij} = 1 if i and j are adjacent or share a common boundary, otherwise w_{ij} = 0. Or w_{ij} might be the length of common boundary or a decreasing function of the distance between the entities' respective centroids.

We can now define a class of indices

$$S = \sum_i \sum_j w_{ij} f(r_i, r_j) \qquad (1)$$

as measures of the spatial structure present in the ordering R. If f is an increasing function of differences, such as $(r_i - r_j)^2$ or $|r_i - r_j|$, then S can be said to measure the extent to which the one-dimensional ordering R preserves or destroys the two-dimensional spatial relationships present in the set of entities. Examples are given in the paper of cases where it is desirable to optimize S in order to preserve spatial relationships.

Indices having the same general structure are well known in the area of spatial autocorrelation. Both the Geary and Moran indices have this general form (Cliff and Ord, 1973), in each case normalized to produce a measure which is comparable across data sets of different n and $\underline{W}$. As simplified for the case where the entities' attributes are permuted integers, they are respectively:

$$I = 12 \sum_i \sum_j w_{ij}(r_i - \bar{r})(r_j - \bar{r}) / [(n^2 - 1) \sum_i \sum_j w_{ij}] \qquad (2)$$

where $\bar{r} = (n + 1)/2$

$$\text{and } c = 6 \sum_i \sum_j w_{ij}(r_i - r_j)^2 / [n(n + 1) \sum_i \sum_j w_{ij}] \quad \text{(Geary, 1954)} \qquad (3)$$

A low value of c is obtained by ensuring that two entities in close proximity to each other (high w_{ij}) are assigned similar integers. Thus we would minimize c in order to best preserve the spatial relationships present in the two-dimensional arrangement. Similarly maximizing I will ensure a strong correlation between the integers assigned to neighbouring entities. Note that different solutions are expected as the two indices are not complementary (I + c is not constant) except in special cases, for example when the weights are standardized such that $\sum_j w_{ij} = 1$ for all i.

The problem of optimizing I, c or more generally S falls into the general class of quadratic assignment first mentioned by Koopmans and Beckmann (1957). In this paper we consider one version of the problem which is immediately applicable to automated cartography: the question of the optimal method of storing raster data.

RASTER STORAGE

There has been considerable recent interest in the structuring of raster data, since options like quad-trees offer considerable advantage in searching and processing operations (see for example Klinger and Dyer, 1976, Samet, 1981). A given raster data set can occupy very different amounts of storage depending on how it is structured, particularly if run-length encoding is used and if the domain of the data is limited. The traditional ordering of a raster, row by row from the upper left corner (Figure 1a), may be less efficient than other orderings because of its property of rapidly traversing the image from one side to the other. An early and somewhat obscure paper by Morton (1966) described the use of a particular raster ordering in processing the frames of images in the Canada Geographic Information System (Figure 1c). Cells which are close together in space appear to be placed in similar positions in the sequence to a greater extent in the Morton ordering than in the conventional one. Recent papers by Tropf and Herzog (1981) and others have revived interest in the Morton ordering, which is recognized as an example of a space-filling or Peano curve. Any real image is likely to show strong spatial autocorrelation, in other words two pixels which are close together on the image are more likely to have similar data attributes than two pixels which are further apart. So we would expect a raster coded in Morton order to have greater average run length, and thus smaller volume, than one coded in conventional order.

Although Morton order has this intuitive advantage, only one move in two in the sequence is to a cell which is a rook's-case or 4-neighbour,

and as the raster increases in size the length of the longest move increases as well. In a raster of 2^N by 2^N cells the longest move is one column and $2^N - 1$ rows (or vice versa). However another member of the Peano family always has the property of moving to a 4-neighbour cell. We will refer to this curve as Pi-order (Figure 1d) and the Morton sequence as N-order in recognition of the basic 3-step moves from which the curves are constructed. Intuitively we would expect Pi-order to improve on the efficiency of both row order and N-order. The comparisons which follow will also include a revision of row-order in which every even row is reversed (Figure 1b). This will be referred to as row-prime order, and also has the property that each move is to a 4-neighbour.

The definition of Pi-order is somewhat more complex than the others. The construction of the curve can be expressed as an algorithm for returning the row number and column number (both between 0 and $2^N - 1$) of a Pi-ordered cell (numbered between 0 and $2^{2N} - 1$) in a 2^N by 2^N array. The algorithm is as follows:

1) Express the Pi-order number to base 4 as a vector of digits $\underline{P}$ with elements $p_i, i=1,N$, 1 being the most significant
2) $\underline{Q} = \underline{P}$
3) $i = 1$
4) If $p_i = 0$, then for $j = i + 1$ to N do:
   if $q_j = 3$ or $q_j = 1$ then $q_j = 4 - q_j$
   end

   If $p_i = 3$, then for $j = i + 1$ to N do:
   if $q_j = 0$ or $q_j = 2$ then $q_j = 2 - q_j$
   end
5) $i = i + 1$. If $i < N$ go to step 4.
6) Let $\underline{X}$ and $\underline{Y}$ be binary representations of the column and row number. Their elements are obtained from the following table:

q_i	x_i	y_i
0	0	0
1	1	0
2	1	1
3	0	1

The four orderings will now be compared in two ways: empirically, in terms of the volume of data generated for standard images, and analytically in terms of the mean difference in sequence between spatially adjacent cells through the performance of each order against S-indices.

EMPIRICAL COMPARISON

Clearly there is no standard image which can be used to obtain results which are totally generalizable. However several authors (Mandelbrot, 1982, Burrough, 1981, Goodchild, 1982) have argued that surfaces generated by fractional Brownian processes show substantial resemblance to a variety of real phenomena including terrain, to the extent that they support some limited degree of generalization. The surfaces each have the property that variance is a power function of distance, that is

$$E[z(\underline{x}) - z(\underline{x} + \underline{d})]^2 = |\underline{d}|^{2H} \qquad (4)$$

where $\underline{x}$ and $\underline{x} + \underline{d}$ are two points a distance $|\underline{d}|$ apart, z is the elevation of the surface, E denotes the statistical expectation and $0 \leqslant H \leqslant 1$ is a parameter. Surfaces with low H are locally rugged but

Figure 1: The Four Standard Orders

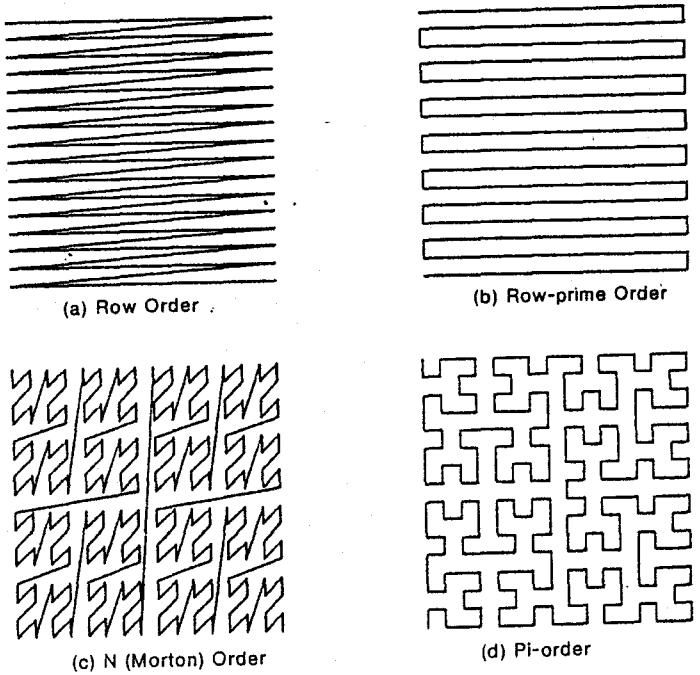
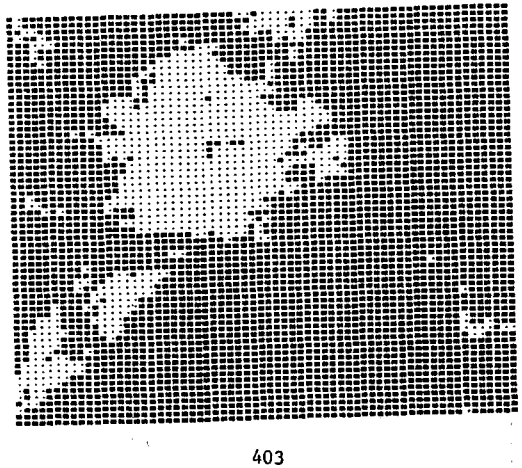


Figure 2: Example Image Created by Contouring a Fractional Brownian Surface (H=0.5) at the Mean Elevation



show little general trend, while surfaces with high H are locally smooth, but show a strong general 'drift'.

Nine surfaces were generated using values of H ranging from 0.1 to 0.9 in steps of 0.1. Each was then converted to a binary image by taking the mean elevation and coding each cell as either below (black) or above (white). An example is shown in Figure 2.

The results are shown in Table 1 in terms of the number of records required to code the image for each of the four orders. Since the maximum run length in a 2^N by 2^N image is 2^{2N}, each run of black or white can be coded as a record of 2N bits. At H = 0.0 each elevation is statistically independent so the number of runs can be obtained from a standard result, provided the numbers of black and white cells are equal.

For locally smooth surfaces (high H) Pi- and row^1- orders have a clear advantage, with a roughly 60% saving in storage over row-order and a 25% saving over N-order. At low H Pi-order is still the most efficient, but the advantage is small, and N-order appears less efficient than the conventional row-order.

TABLE 1: Data volume for binary images derived from fractional Brownian surfaces

		Order			
	H	Row	Row^1	N	Pi
	0.9	179	116	143	108
	0.8	155	92	123	86
	0.7	168	105	136	85
Contour	0.6	169	106	151	96
at	0.5	488	469	580	473
mean	0.4	398	364	462	378
	0.3	992	972	1050	980
	0.2	1477	1461	1539	1443
	0.1	1796	1779	1882	1707
Contour at median	0.0	2047	2047	2047	2047

ANALYTICAL COMPARISON

It was argued in the first section that the difference between the position in the sequence assigned to one cell and the positions assigned to its neighbours was a measure of the extent to which the sequence preserved the spatial relationships in the two-dimensional arrangement of cells. In this section we consider two S-indices for each of the four orders discussed above:

$$S_1 = \sum_i \sum_j w_{ij} |r_i - r_j| / \sum_i \sum_j w_{ij} \qquad (5)$$

and

$$S_2 = \sum_i \sum_j w_{ij} (r_i - r_j)^2 / \sum_i \sum_j w_{ij} \qquad (6)$$

S_1 can be interpreted as the mean difference in ordered positions between neighbouring cells and S_2 as the mean square difference. Each element w_{ij} is set to 1 if j is a 4-neighbour of i, otherwise 0.

Table 2 shows the analytical results obtained for S_1 and S_2 for cellular arrays of 2^N by 2^N cells. The expression for S_2 for Pi-order was obtained empirically rather than analytically and must be regarded as an estimate only.

As N becomes large the equations can be compared through the coefficients of their dominant terms alone. It is clear that row, row^1 and N are equal and superior to Pi against S_1, and that the conventional row order is best for large N against the S_2 criterion, since it has a smaller coefficient than row^1-order for 2^{2N}, and both N and Pi-order are dominated by terms in 2^{3N}. Pi-order seems worst on both criteria.

TABLE 2: Analytical results for S_1 and S_2 for the four orders

	S_1	S_2
Row	$(2^N + 1)/2$	$(2^{2N} + 1)/2$
Row-prime	$(2^N + 1)/2$	$(2^{2N+1} + 1)/3$
N (Morton)	$(2^N + 1)/2$	$\frac{5}{126}[2.2^{3N} + 2.2^{2N} + 16.2^N + 7]$
Pi	$[51.2^{2N} + 16.2^N + 16]/[84.2^N]$	$0.195\ 2^{3N}$ + lower order terms

OPTIMIZATION

In this section we consider the possibility of finding orders which optimize S-indices. Branch and bound (see for example Lawler, 1963, Pierce and Crowston, 1971) and a natural selection heuristic (Francis and White, 1974, p.377) were both applied to optimize the Geary and Moran indices to find orders which best preserve spatial relationships. The exact method was used for N=2 and the heuristic for N > 2.

Minimizing the Geary index for N=2 produced two solutions with the same optimum of c = 0.188. The first was row-order, which is consistent with the results obtained in Table 2, since the Geary index is proportional to S_2. Maximizing Moran for N=2 also gave two distinct solutions, with I = 0.729, but neither was the same on any of the four standard orders. For N=3 and N=4 the heuristic algorithm obtained solutions for both Moran and Geary which were substantially better than any of the four standard orders. We conclude that while the four orders behave relatively well, none is an optimal solution. A summary of the performance of the four orders is shown in Table 3, and it is clear that row-order is the best against the Geary criterion and N-order against the Moran.

CONCLUSIONS

Four raster orderings, including the traditional row order, the Morton or N-order, and a new order based on a Peano space-filling curve, have been evaluated against the criterion of preserving spatial relationships. Empirically it appears that Pi-order performs best against fractional Brownian surfaces of varying H, and these are arguably representative of a broad set of real images. Analytically, it appears that cases can be made for row-order, row^1-order and N-order, but that none of the four orders optimize simple measures of efficiency.

TABLE 3: Geary and Moran statistics for the four standard orders

	N	row	row-prime	N(Morton)	Pi
	2	0.188	0.243	0.202	0.368
Geary	3	0.047	0.062	0.074	0.165
	4	0.012	0.016	0.033	0.078
	2	0.667	0.608	0.706	0.545
Moran	3	0.857	0.842	0.881	0.801
	4	0.933	0.929	0.946	0.907

The general framework and measures used in this paper can be applied in a number of other areas. For example the order of U.S. states which minimizes S_1 or S_2 would be an appropriate order for tabulating statistical data, as it would preserve the spatial relationships on the map to the greatest possible extent. Different definitions of $\underline{W}$ could be used to preserve either local relationships within regions, or general relationships within the nation as a whole. Maximizing S_1 or S_2 has the effect of assigning neighbouring entities maximally different positions in the sequence. Such a strategy would be reasonable in the assignment of telephone area codes as it would ensure very different, and therefore hard to confuse, codes to neighbouring zones. Finally one might regard the design of a dart board as being a problem in assigning the integers 1 through 20 to maximize an S index.

REFERENCES

Burrough, P.A. 1981, Fractal dimensions of landscapes and other environmental data: Nature, 294, pp. 240-242.

Cliff, A.D. and Ord, J.K. 1973. Spatial Autocorrelation, Pion, London.

Francis, R.L. and White, J.A. 1974. Facility Layout and Location -- An Analytical Approach, Prentice Hall, Englewood Cliffs, N.J.

Geary, R.C. 1954, The contiguity ratio and statistical mapping: Incorporated Statistician, 5, pp. 115-141. Reprinted in Berry, B.J.L. and Marble, D.F. 1968, ed. Spatial Analysis, Prentice-Hall, Englewood Cliffs, N.J., pp. 461-478.

Goodchild, M.F. 1980, Simulation of autocorrelation for aggregate data: Environment and Planning A, 12, pp. 1073-1081.

Goodchild, M.F. 1982, The fractional Brownian process as a terrain simulation model: Modelling and Simulation, 13, pp. 1133-1137.

Klinger, A. and Dyer, C.R. 1976, Experiments in picture representation using regular decomposition: Computer Graphics and Image Processing, 5, pp. 68-105.

Koopmans, T.C. and Beckmann, M. 1957, Assignment problems and the location of economic activity: Econometrica, 25, pp. 53-76.

Lawler, E.L. 1963, The quadratic assignment problem: Management Science, 9, pp. 586-599.

Mandelbrot, B.B. 1982. The Fractal Geometry of Nature, Freeman, San Francisco.

Morton, G.M. 1966. A computer oriented geodetic data base, and a new technique in file sequencing, IBM Canada Ltd.

Pierce, J.F. and Crowston, W.B. 1971, Tree-search algorithms for the quadratic assignment problem: Naval Research Logistics Quarterly, 18, pp. 1-36.

Samet, H. 1981, An algorithm for converting rasters into quadtrees: IEEE Transactions on Pattern Analysis and Machine Intelligence, 3, pp. 93-95.

Tropf, H, and Herzog, J. 1981, Multidimensional range search in dynamically balanced trees: Angewandte Informatik, 2, pp. 71-77.