Proceedings of the 4th International Symposium on Spatial Data Handling, Zurich, 1990, pp. 151-160.

CALCULATING MAP OVERLAY POLYGONS' AREAS WITHOUT EXPLICITLY CALCULATING THE POLYGONS — IMPLEMENTATION

Wm. Randolph Franklin

Electrical, Computer, and Systems Engineering Dept., 6026 JEC, Rensselaer Polytechnic Institute, Troy, NY, 12180, USA

Internet: wrf@ecse.rpi.edu, Bitnet: wrfrankl@rpitsmts.bitnet

Telephone: (518) 276–6077, Telex: 6716050 RPI TROU, Fax: (518) 276–6261

Abstract

The implementation of OVERPROP, an algorithm for quickly calculating the areas of all the polygons resulting from overlaying two input maps, is presented. OVERPROP works from a reduced representation of each map as a set of "half-edges" with no global topology. It does not first calculate the overlaid map. In fact, finding the areas of all the output polygons is simpler than finding the polygons themselves. This method is exact within the arithmetic precision of the machine, and does not use raster, Monte Carlo, or other sampling techniques. It is well suited to a parallel machine, and could be extended to overlaying more than two maps simultaneously, and to determining other properties of the output polygons, such as perimeter or center of mass. OVERPROP is useful when the sole purpose of overlaying two maps is to find some mass property of the resulting polygons.

Introduction

In the map overlay problem, two maps, or planar graphs, are combined to make a third, each of whose polygons represents the intersection of one polygon from the first input map with one from the second. For instance, if the first map's polygons are countries and the second watersheds, then one output polygon would be that part of France that drains into the Atlantic. The overlay problem is one of the more difficult computational issues in GIS since the algorithm is complex, the data are large, and the numerical inaccuracy of computers can confound even correct algorithms [10, 11].

Our algorithm, OVERPROP, presents a simpler method, more efficient for large databases, that is less sensitive to numerical errors since they cannot cause topological difficulties. It is based on our simpler, local topological data structures for representing polygons and polyhedra. Its speed comes from the uniform grid data structure. These concepts have been described in the context of computer aided design (CAD) most recently in [4]. A preliminary description of this algorithm was presented in [3]. In the rest of this paper, we will see, first, the logical data structures, then the theoretical formulae underlying the algorithm. We will see the algorithm described, first, briefly, ignoring efficiency considerations, and then in more detail. Control of numerical errors, design tradeoffs, and possible extensions complete of the paper.

Major Data Structures

Conceptually, each input map is a planar graph composed of vertices and edges partitioning the E^2 plane into polygons. The graph may be disconnected with the components either nested or disjoint. Each polygon has an identification number, which may be non-unique when one logical region is composed of several polygons. By convention, the outside is polygon #0. In our map data structure, the vertices and polygons are not represented explicitly, but could be recreated if desired.

We assume that each input map is a set of tagged edges: $\{(x_1, y_1 x_2, y_2, p_l, p_r)\}$ where (x_1, y_1) are the Cartesian coordinates of the edge's first endpoint, and(x_2, y_2) are the second. (p_l, p_r) are the numbers of the polygons to the left and right.

Each edge of each polygon has two endpoints, each of which defines a "half-edge", e. Since a data structure edge borders two polygons, it decomposes into four half-edges. Although the half-edges are not explicit data structure elements, they can be derived easily, and are important in the following algorithm. A half-edge contains the information $(v_x, v_y, t_x, t_y, n_x, n_y)$ where (v_x, v_y) is the location of the endpoint, (t_x, t_y) is a unit tangent vector from the endpoint along the edge, and (n_x, n_y) is a unit normal vector perpendicular to the tangent, and pointing into the polygon.

A polygon number of the output map is an ordered pair (p_1, p_2) of the two input polygons whose intersection forms this particular output polygon.

Theoretical Basis of the Algorithm

The algorithm is built on the following principle. Many properties of a polygon, such as area, center of mass, and moments of inertia, may be calculated separately for each half-edge of the polygon and then summed. These are "extensive" properties in a thermodynamic sense, in that the total value for an object may be found by integrating over its area or volume. This concept is also related to Greene's theorem in calculus, where an integral over an area is transformed into an equivalent integral over the boundary of the area. Formally, let polygon *P* have the set of half-edges *E*. Let *F(P)* be some desired function of the polygon, such as area. Then for these *F*, there exists a functional that returns a new function f_F such that $F(P) = \sum_{e \epsilon E} f_F(e)$.

For example, $f_{area} = \frac{1}{2}(v_x t_x + v_y t_y)(v_x n_x + v_y n_y)$ and $f_{perimeter} = -(v_x t_x + v_y t_y)$

So if we can find the half-edges of the output polygons we can find the areas. The output half-edges needed by the formulae are of two types:

a) derived from an endpoint of an original edge of one of the input maps, or
b) derived from an intersection of two input map edges.

For each original endpoint we also need to know which polygon of the other map this point is in. For each intersection of two edges we already know the numbers of the relevant polygons.

Brief Algorithm

A sketch of the algorithm, ignoring efficiency considerations, now appears as follows.

1. Initialize a data structure to contain a list of triples, *(p_1, p_2, partial-area)*, to accumulate the parts of the areas of the output polygons. p_1 and p_2 are polygons from the first and second input maps respectively.
2. Find all intersections of any edge of map 1 with any edge of map 2. Each intersection will generate 8 half-edges, as shown in figure 1 where the tangents and normals of three are labeled.

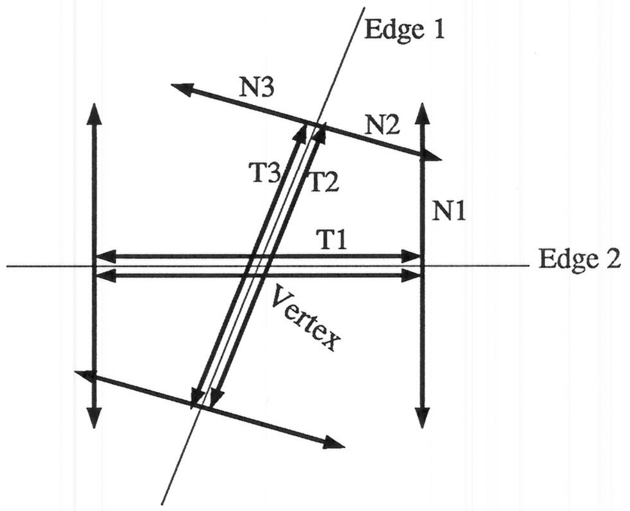

Figure 1: The 8 Half-Edges Derived from One Intersection

3. For each half-edge resulting from an intersection, find the tangent and normal vectors and the relevant polygons from each input map. Note that in figure 1 there are two polygons from map 1, one on each side of edge 1, and also two from map 2, one on each side of edge 2. Each of the 4 possible pairs of the input polygons will apply to 2 of the 8 half-edges. For example, half-edges 1 and 2 are part of the output polygon that is the intersection of the polygon on map 1 to the right of edge 1 with the polygon on map 2 above edge 2.
4. Apply the formula to each half-edge to calculate its contribution to the relevant output polygon's area.
5. Now we must calculate and process the half-edges not resulting from an intersection. For each endpoint of each edge of map 1, determine which polygon of map 2 contains it. Repeat for the edges of map 2.
6. Split each edge of each input map into four half-edges.
7. Apply the formula to these half-edges and store their partial areas.
8. Extract the final area of each non-empty output polygon, which was the goal.

Detailed Efficient Algorithm

The algorithm presented in the last section is of value only if it is efficient enough. For instance, if finding all intersections of input edges takes quadratic time in their number, then maps with 10,000 edges, and thus 50,000,000 possible intersection, will be difficult to process. This section tells how to make the algorithm practical.

General numerical speed: On almost every computer, floating point computations are slower than integer, sometimes by a factor of three or more. Initially, when using 4 byte integers, we scale the map coordinates to the range [—M, M] in X and Y, where M=20000. This allows us to calculate edge equations exactly and then substitute a point into the equation to determine which side of the edge it is on, again without any roundoff, since the largest number computed will be absolutely less than $4M^2$.

Storing (cell, edge) information: We use a square array with one extendible array per cell. The cell is a pointer to an array initially 5 long, with the first element the allocated size and the second the number of elements in use, leaving 3 to store edge numbers. If that is insufficient, then the array is reallocated with double the size, the old elements copied, and the pointer from the square cell array changed to point to the new one. Statistics are kept of the number of times that a reallocation is necessary. Contrary to appearances, even if a cell's array is reallocated many times, each edge number is copied very few times. If the array size is grown by a factor α each time (here α=2) from one element to a large size, then each element is copied an average of $1/(\alpha-1)$ times (here 1).

Finding intersections: We use Franklin's uniform grid method [4, 6].

Locating the polygon containing each edge endpoint: The naive method compares each point against each edge of the other map. However, using a uniform grid again, this can be done in expected time $\Theta(N+M)$ if there are N points to test against a map with M edges. We use an optimized version of an implementation of a one-dimensional grid method by V. Sivaswami that has the following performance on a Sun 3/60 workstation.

1. The first case tested 10,000 points against a random polygon with 100,000 edges. Preprocessing took 190 μsec/edge, or 19 seconds CPU time, and classification took 7.6 msec/point, or 76 seconds.

2. In the second test, we took a map of the USA with 915 edges in 166 chains and 49 polygons, and located 930 random points. Preprocessing took 440 μsec per edge and location took 3.8 msec per point, for a total preprocessing and query time of 3.9 seconds. The sample data are shown in figure 2.

In these tests, the total time varied less than 25% as the linear grid resolution varied over a range of 25:1.

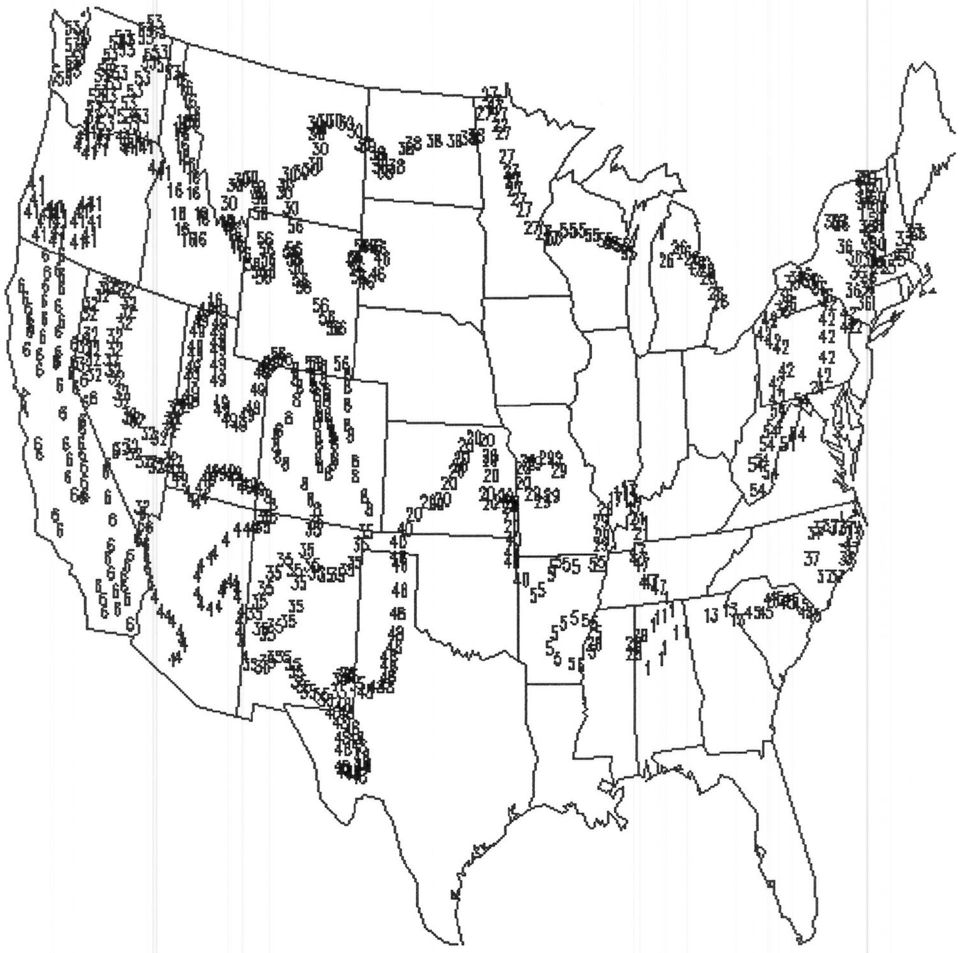

Figure 2: Point Classification in a Planar Graph. Each Point is Labeled with its Calculated Polygon Number.

If the user is feeling pessimistic about the data, and worried that the uniform grid might fail, then there do exist more complicated but worst-case optimal methods [8]. A typical time for these might be $\Theta((N+M) \log M)$.

Creating the one dimensional slab structure:

1. Each slab is one column of grid cells. The edges in the cells in a column are gathered together and sorted primarily by their minimum Y coordinates, and secondarily by their number. This comparison predicate assures that two edges will not compare the same unless they are the same edge. This results in duplicates of the same edge being adjacent after the sort, so they can be deleted. Duplicates occur when the same edge falls in more than one cell in the column.

2. Finally a selection sort is performed on the unduplicated edges. The criterion is the partial order where one edge is less than another if, from a viewpoint of $(0,-\infty)$, the former edge (at least partly) obscures the latter one.
3. To determine the polygon containing a point, first the slab containing the point is determined. Then a ray is fired up from the point to the first edge that it hits, in the order that the edges are stored in the slab. Then the point is in this edge's left or right neighboring polygon depending on whether the edge's first endpoint is to the right or left of its second endpoint, respectively.

Degenerate cases: These occur when a vertex of one map falls on an edge or vertex of the other. Although this has previously been a very difficult problem, there is now a complete theoretical, and practical, solution to the problem of degeneracies, the Simulation of Simplicity technique of Edelsbrunner and Mucke [2]. Briefly, this pretends to add an infinitesimal to the coordinates of the second map. By definition, all first order infinitesimals are less than all positive finite numbers. Different orders of infinitesimals may be used; all second order ones are less than all first order ones, etc. A delightful book on this is by Knuth [7]. The effect of infinitesimals is to prevent any comparison tests from returning an equality. We do not actually work with infinitesimals, but instead determine what their effect would be on any test, and code a modified test accordingly. The modified test is longer but generally not significantly slower. For example, suppose that we are testing a point (x, y) against a line $ax + by + c = 0$, and that we are using simulation of simplicity to pretend to shift the point by (ϵ, ϵ^2). Thus the test becomes as follows.

```
    d=ax+by+c;
    if (d≠0) return signum(d);
    else if (a≠0) return signum(a);
    else return signum(b);
```

`signum` returns —1, 0, or 1 according to the sign of its argument. Note that the extra tests are called only in the degenerate case. Now, one might deduce this particular test without resort to simulation of simplicity. However this technique provides a general, theoretically well-founded, method for creating all such tests.

Error Control

Numerical roundoff errors, and the numerous sliver polygons generated even if there is no roundoff, help make the polygon overlay problem so hard. With OVERPROP, slight errors in calculating intersections vertex coordinates will cause proportionally slight errors in the areas. However, although OVERPROP does not explicitly use the topology of the polygons, it implicitly assumes that it is consistent in the following ways.

a) There are no missing edges, and no gaps between edges of a polygon.
b) The polygon numbers stored with the edges are all correct.

In fact the sensitivity of OVERPROP to these input errors may be used to test for consistency by the following method.

a) Repeatedly randomly translate and rotate the input maps.
b) Find the areas with OVERPROP.
c) If the answers are different by more than numerical roundoff, then the input is bad. If the answers agree after several random operations, then the input is almost certainly internally consistent.

OVERPROP does not care about roundoff errors that cause one chain of edges to cross another erroneously as shown in figure 3. The will cause the output to be off by an amount proportional to the geometric error in spite of this topological error.

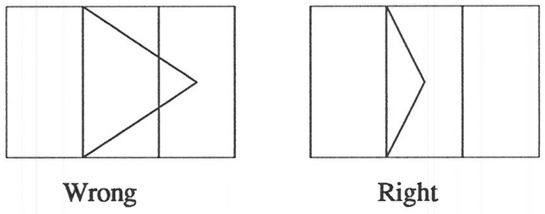

Figure 3: Topological Error Caused by Numerical Inaccuracy

Design Tradeoffs

The data structures and algorithms used here were selected from many possibilities. Factors considered included efficiency of execution and efficiency of implementation, i.e. simplicity. Some decisions included the following.

Storing the vertices in a separate array and having the edges contain vertex numbers instead of vertex coordinates: This is the choice between immediate versus indirect data. Storing the vertex coordinates immediately in the edges makes processing the edges cheaper. Depending on wordlengths, the storage will be the same or greater. If a vertex position takes 2 bytes each for x and y, and a vertex number takes 4 bytes, then the storage is identical. On the other extreme, if each coordinate takes 4 bytes, but a vertex number takes only 2, then the indirect method will be smaller. With memory prices for workstations at $100 per megabyte and falling, then memory is not so important as before. However, big programs will always execute more slowly than otherwise equivalent small programs because of bus bandwidth limitations, and because the big programs will use the cache less efficiently.

Using integers instead of reals for coordinates: This saves space and execution time. However, attention must be paid to scaling since that is no longer automatic.

Splitting long chains of edges that form a common boundary into the separate edges: This somewhat controversial choice makes the data structure bigger, but each data element is now fixed length instead of variable. This is a similar concept to a relational database satisfying Codd's third normal form [9]. Also the algorithm is simpler with fixed length items, and parallelizes better.

Implementation

A preliminary version of OVERPROP is now up and running, coded in C on a Sun 3/60, a 4 MIPS machine with an MC68020 processor. The biggest test case, shown in Figure 2, involved overlaying a map of the coterminous United States with 1081 vertices, 915 edges, and 49 polygons against a map of July isotherms at 10°F intervals, with 920 vertices, 892 edges, and 6 polygons. With a 100×100 grid, the biggest cell had only 14 edges from the 2 maps. The actual histogram of number of cells with n edges for $0 \le n \le 14$ was 5259 2900 1065 493 159 66 31 15 8 0 0 0 2 1 1. Among the 10000 cells there were only 287 reallocations for more space. 1357 pairs of edges were tested for intersection to yield 111 intersections and another 146 duplicate intersections. When the slabs were created from the edges, the biggest slab had under 100 edges from both maps together.

The time was measured using the Free Software Foundation *gcc* compiler with optimizing enabled. Reading the input files took 4.9 CPU seconds, and after that, everything else, including applying the grid, finding the intersections, locating the vertices of each map in the other, and calculating the areas of all the output polygons, took another 7.2 seconds. We expect the total time to shrink in the future, and also expect to test much bigger data sets.

Possible Extensions

Overlaying multiple input maps simultaneously: These ideas extend to finding the areas of the results of overlaying more than two input maps simultaneously. As with two maps, the output half-edges are derived from input edge endpoints and intersections of edges. What is new is that the location of each half-edge must be determined in every input map. In this case, the advantage of OVERPROP compared to actually finding a sequence of more and more complicated intermediate maps would be even greater. There would also be no artifacts resulting from the order in which the several input maps were processed, since they would all be used together.

Execution on parallel machines: OVERPROP is designed also to be implementable on a parallel machine. The component operations are generally one of two types: map a function over a set of elements to calculate a new set, or sort a set of elements, perhaps combining adjacent elements. Mapping is easily executable even on a SIMD (single instruction, multiple data stream) machine. Sorting is a little harder, but parallel sort algorithms exist.

This ease of parallel execution contrasts to many other, more complicated, geometric algorithms that are essentially sequential. We have not implemented this complete algorithm in parallel yet, but Chandrasekhar Narayanaswami and Mohan Kankanhalli have implemented edge intersections in parallel, as well as other other algorithms, such as Boolean combinations of polyhedra, [1], and hidden surface determination, [5]. The machines used have been a 16 processor Sequent Balance 21000 and a 32 processor hypercube. Both are MIMD machines. The Sequent has a common memory, runs Unix, and makes it easy to allocate different processors to different processes, or to different iterations of the same loop of the same process. As many as 15 of the 16 processors can be used in this case. Typical speedups when using 15 processors have been ten times faster than when using one processor.

The Hypercube has a separate, rather small, memory for each processor, and communication between processors is somewhat slow. Since one processor has too little memory to run a reasonable sized problem, we couldn't compare a sequential and a parallel time. Therefore we measured the communication time compared to the total CPU time, and the CPU time of the slowest processor compared to the average of all the processors. The communication time was 1/3 or less of the total time, and the slowest processor took no more than twice the average processor. Therefore if the load were perfectly balanced over the 32 processors and had no communication cost, it would be under three times faster, which is a respectable performance.

Summary

OVERPROP calculates the areas of the intersection polygons resulting from overlaying two maps very efficiently. Unlike other polygon overlay algorithms, it uses no explicit global topology. There is no tracing chains of edges. Polygons completely contained inside other polygons are not a problem, in fact are not even recognized; however the correct answer is produced with the area of the outer polygon excluding the inner area. Polygons may also have multiple separate components. OVERPROP is a result of an investigation into how little topology we actually need explicitly to store, and concurrently, by how much can special cases be reduced. Sometimes a more complete topology is needed, for instance to draw the output polygons, but for mass calculations it is not.

Acknowledgements

This work was supported by NSF Presidential Young Investigator grant CCR-8351942. Partial support for this work was provided by the Directorate for Computer and Information Science and Engineering, NSF Grant No. CDA-8805910. We also used equipment at the Computer Science Department and Rensselaer Design Research Center at RPI. Part of this work was conducted using the computational resources of the Northeast Parallel Architectures Center (NPAC) at Syracuse University, which is funded by and operates under contract to DARPA and the Air Force Systems Command, Rome Air Development Center (RADC), Griffiss Air Force Base, NY, under contract # F306002-88-C-0031. Part of the research reported here was made possible through the support of the New Jersey Commission on Science and Technology and the Rutgers University CAIP Center's Industrial Members.

Bibliography

[1] Chandrasekhar, N., and Franklin, W. R. A fast practical parallel convex hull algorithm. Tech. rep., Electrical, Computer, and Systems Engineering Dept., Rensselaer Polytechnic Institute, 1990.

[2] Edelsbrunner, H., and Mucke, E. P. Simulation of simplicity: A technique to cope with degenerate cases in geometric algorithms. In *Symposium on Computational Geometry* (1988), pp. 118–132.

[3] Franklin, W. R. Overprop - calculating areas of map overlay polygons without calculating the overlay. In *Second National Conference on Geographic Information Systems* (Ottawa, 5-8 March 1990).

[4] Franklin, W. R., Chandrasekhar, N., Kankanhalli, M., Akman, V., and Wu, P. Y. Efficient geometric operations for CAD. In *Geometric Modeling for Product Engineering*, M. J. Wozny, J. U. Turner, and K. Preiss, Eds. Elsevier Science Publishers B.V. (North-Holland), 1990, pp. 485–498.

[5] Franklin, W. R., and Kankanhalli, M. Parallel object-space hidden surface removal. In *Proceedings of SIGGRAPH'90 (Dallas, Texas) in Computer Graphics* (August 1990), vol. 24.

[6] Franklin, W. R., Narayanaswami, C., Kankanhalli, M., Sun, D., Zhou, M.-C., and Wu, P. Y. Uniform grids: A technique for intersection detection on serial and parallel machines. In *Proceedings of Auto Carto 9: Ninth International Symposium on Computer-Assisted Cartography* (Baltimore, Maryland, 2-7 April 1989), pp. 100–109.

[7] Knuth, D. E. *Surreal numbers: how two ex-students turned on to pure mathematics and found total happiness: a mathematical novelette*. Addison-Wesley, 1974.

[8] Preparata, F. P., and Shamos, M. I. *Computational Geometry: An Introduction*. Texts and Monographs in Computer Science Springer-Verlag, 1985.

[9] Ullman, J. D. *Principles of Database and Knowledge-base Systems*. Principles of computer science series, 14. Computer Science Press, 1988.

[10] White, D. A new method of polygon overlay. In *An Advanced Study Symposium on Topological Data Structures for Geographic Information Systems* (Cambridge, MA, USA, 02138, 16-21 October 1977), Laboratory for Computer Graphics and Spatial Analysis, Harvard University.

[11] White, D., Maizel, M., Chan, K., and Corson-Rikert, J. Polygon overlay to support point sample mapping: The national resources inventory. In *Proceedings of Auto Carto 9: Ninth International Symposium on Computer-Assisted Cartography* (Baltimore, Maryland, 2-7 April 1989), pp. 384–390.