# WHIRLPOOL:
# A GEOMETRIC PROCESSOR FOR POLYGON COVERAGE DATA

James Dougenik
Laboratory for Computer Graphics
and Spatial Analysis
Harvard University
48 Quincy Street
Cambridge, Massachusetts 02138

## I. Introduction

WHIRLPOOL is one of the family of ODYSSEY programs developed for the manipulation and display of polygon coverage data. Together these programs have a broad range of capabilities from creation of cartographic base files to interactive display of thematic maps. (Dutton 1977) (Teicholz 1979).

By itself WHIRLPOOL has a variety of functions. In the creation of new base files, the program can plan an important role, automatically correcting some common digitizing errors and detecting others. Another important function is the geometric overlay of polygon coverages to produce a combined coverage for further use in modeling and analysis.

The program has other useful capabilities which are necessary to perform geometric polygon overlay and can also be used alone. It can calculate polygon areas and produce a file containing the area and perimeter of each polygon. The technique is applicable even for highly complex coverages whose polygons may be too complicated for other programs. A list of location points can be processed producing a file containing, for each point, the polygon identifier of the polygon within which the point is located. The base file can

also be generalized to remove excess detail points.

The program was written in FORTRAN by James Dougenik and Nick Chrisman. It has been installed on a PDP-10 and on an IBM 370 with the CMS operating system, and is currently being installed on other systems.

Polygon Coverages

One of the unifying concepts underlying the ODYSSEY family of programs is that of a common cartographic file organization. Called a "chain file", it is an abstract representation of the network of boundaries present in a polygon coverage. (Puecker and Chrisman 1975).

In a polygon coverage, the area of interest is partitioned into non-overlapping polygons using nominal (or numerical) categories. Examples are governmental jurisdictions such as municipalities or states, land use categories, or zoning classifications. The "chain" abstraction represents the boundaries between polygons as sequences of line segments approximating the true boundary to the accuracy necessary. The lists of boundary segments are often referred to as chains or arcs, and where they meet are called nodes or junctions.

The chain file consists of all the chains for a study area. Each chain consists of a list of its segments and the names or identifiers of the two polygons of which it is a boundary and the two nodes at which it begins and ends. The node identifiers allow easy assembly of the network again (Puecker and Chrisman 1975) (Corbett 1975).

II. Major Functions

Polygon Overlay

The statement of the problem of polygon overlay is deceptively simple. Given two polygon coverages of the same study area, make them into one polygon coverage. Its solution is a powerful tool. It gives the information necessary to answer specific questions such as "What regions possess a certain combination of properties?", or "What combinations of properties does this region possess?" and more general questions such as "What properties are commonly (rarely) present together?".

When performing polygon overlay, certain major tasks can be identified. (White 1977). To make the two coverages into one, geometrically, requires finding all the locations where polygon boundaries from one coverage intersect the other. Finding the intersections is not too difficult, but finding them efficiently is challenging, and is the key to an efficient polygon overlay program. Shamos and Hoey (1976) describe some techniques applicable to finding intersections efficiently.

To answer the questions about the combined coverages, the correspondences between the polygons of the combined coverage and the separate coverages must be established. Some created polygons contain boundaries from both original coverages; the correspondences here are immediately available. Some original polygons are not intersected by the other coverage; more complicated techniques are necessary to find within which polygon of the other coverage they lie.

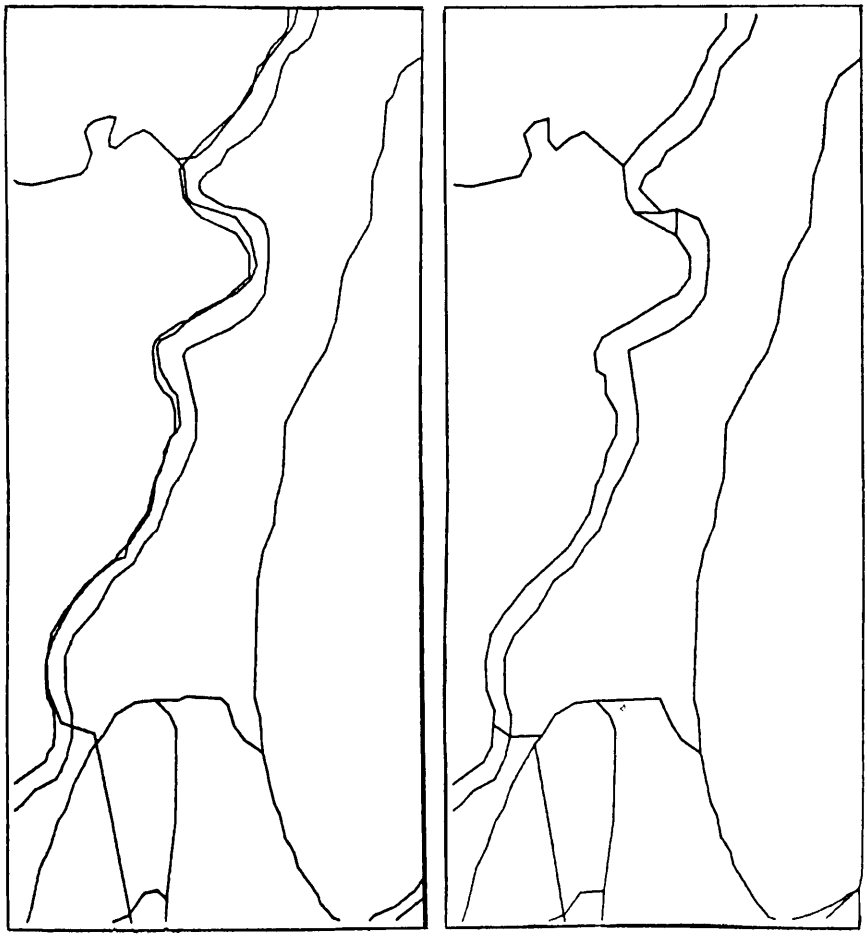
The great difficulty that is encountered when attempting to solve the polygon overlay problem is not at all obvious. This is because it is really the separate problem of data error. The polygon boundary locations are not exact. Errors in surveys, the source maps, the digitizing process all combine to make the boundaries only approximate locations. The difficulty arises when the same boundary is represented slightly differently in the two different coverages. The result is "sliver" polygons, small areas resulting from where the two descriptions do not match. (Goodchild 1977) They are small in area, yet can greatly increase the size of the combined coverage file. Also, the more detailed the boundary descriptions the more slivers are produced.

WHIRLPOOL can solve this problem by using its error distance, which is described in more detail later. This distance is used as the maximum distance which any location can be moved. By giving the program the freedom to move coordinates, it can remove many sliver polygons because they are narrower than the error distance. See Figure I.

Digitizer Data

An important function in any polygon coverage manipulation system is the creation of polygon coverages. A digitizer will record the coordinates of the polygon

Figure I. On the left, a map of zoning classifications has been overlayed with a map of soil types. A boundary from the zoning map closely follows the west bank of a river from the soils map, producing numerous sliver polygons. On the right, WHIRLPOOL has removed the slivers using an appropriate error distance. The triangular polygon in the river results from where the zoning boundary was closer to the other bank of the river, due to digitizing error. The procedure cannot handle errors which are larger than the error distance.



boundaries, but the polygon identifiers have to be entered and the boundary endpoints at the same junction coerced to the same coordinate location before a correct coverage has been created. When digitizing manually, segments are sometimes duplicated, and boundaries are occasionally digitized without stopping at junctions resulting in errors which must be corrected.

There are many approaches to these problems. A common solution is that boundary endpoints are given numbers and all endpoints with the same number can then be given the same location. The two polygon identifiers are also entered when the polygon boundary is digitized. This type of solution has the drawback that each time a number or identifier has to be entered more than once the chance of making a mistake increases.

A different technique is as follows. After digitizing the boundaries, a point is digitized for each polygon and the identifier is entered with it. A point can then be used to name a polygon in the coverage by finding within which polygon it is located. The boundary endpoints can be rectified if the different digitized locations fall within an error distance of each other. This process removes the necessity of specifying identifiers multiple times and eliminates node number entirely. Digitizing operations must then be aware of the error distance when preparing highly detailed maps.

Digitized data can be processed through WHIRLPOOL using this technique. Errors in digitizing such as forgetting to stop at boundary endpoints, and digitizing segments more than once are automatically corrected by finding intersections and eliminating sliver polygons.

Implementing the Error Distance

The error distance plays a major role in the polygon overlay and digitizer input processes of WHIRLPOOL. First, the basic concept of intersection is generalized to that of extended intersection or "fuzzy" intersection. Basic intersections occur if two line segments share a common point. Extended intersections occur if a point from each line is within the error distance of each other. A basic intersection is the special case of an extended intersection with the error distance at zero. Segments are then broken at their point of intersection.

The other important role the error distance plays is when the boundary endpoints are coalesced. The error distance is used to find endpoints which are closer than this distance. These are collected together and a clustering analysis is performed. This process selects which points are moved and which stay fixed. These clusters can get quite complicated if the error distance is relatively large. Then a point can be close to a point which is close to another point, and so on. As points are moved, checks are made to determine when two boundary segments have become identical.

Spatial Filter

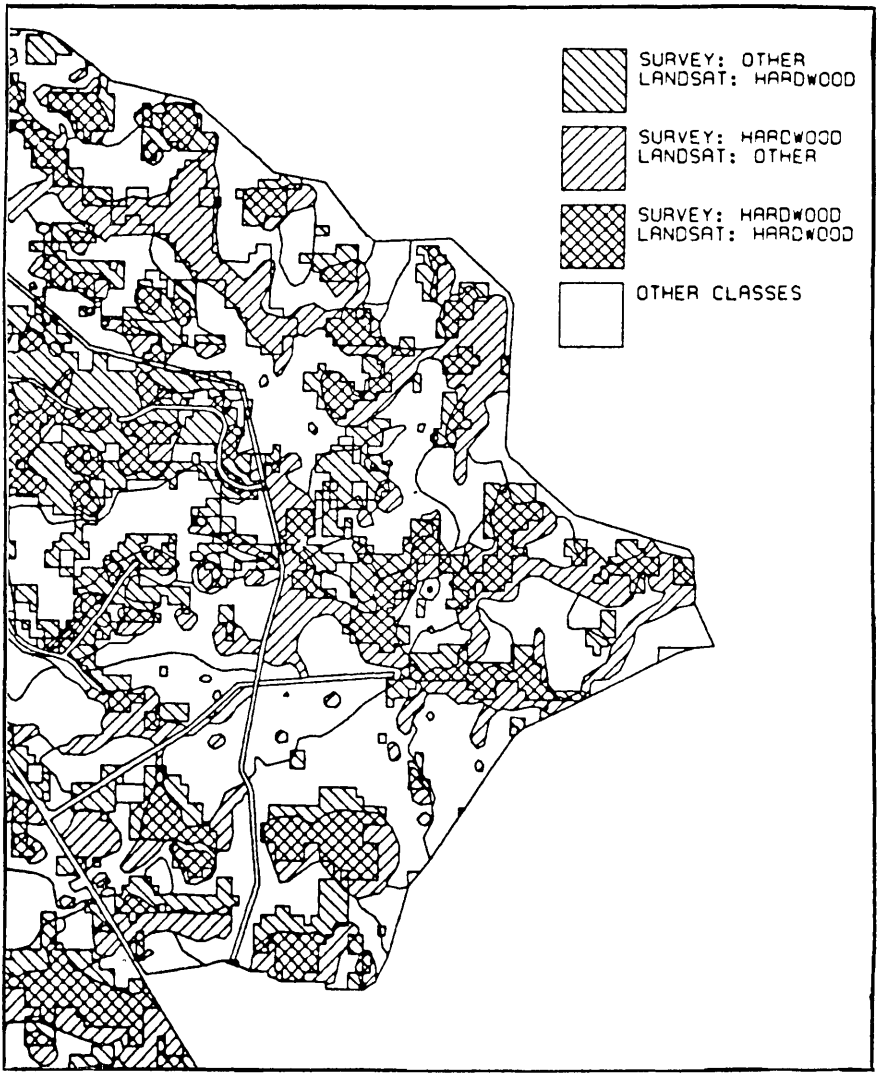
A byproduct of the implementation of the error distance is its use as a spatial filter. By making the error distance larger when processing a digitized file, two things happen. Detail is reduced along boundary lines and closely packed features begin to coalesce. The properties of this filter technique are largely unexplored but its results to date have been pleasing for removing excessive detail and numerous small polygons.

## III. Applications

The program has been used in a number of test and practical applications. A file of census tracts for Montreal was digitized and processed by WHIRLPOOL to create the base file used for an epedemiological study. (Nisen and Hunt 1979)

An important application was a test project sponsored by a lumber company. (Morehouse and Dougenik 1979) A polygon coverage classified by forest types was digitized from maps compiled by traditional survey methods of ground survey and aerial photographs. This was overlaid with a classified LANDSAT matrix which had been converted to a polygon coverage. The results were then analyzed for agreement of forest types. The results were very interesting with many large areas of agreement. Some major areas of disagreement were explained through lack of map updating to reflect recent harvests of hardwood trees. An example of this is found in the upper portion of Figure II.

Figure II. An example from an overlay of a LANDSAT derived polygon coverage and a similarly classified coverage digitized from survey maps.
(The shaded map was produced by the POLYPS program.)



## REFERENCES


Corbett, J. (1975). "Topological Principles in Cartography," AUTO-CARTOGRAPHY II.

Dutton, G. H. (1977). "Navigating ODYSSEY," Harvard Papers on Geographic Information Systems, vol. 2. Reading, Mass.: Addison-Wesley.

Goodchild, M. F. (1977). "Statistical Aspects of the Polygon Overlay Problem," Harvard Papers on Geographic Information Systems, vol. 6. Reading, Mass.: Addison-Wesley.

Morehouse, S. and Dougenik, J. (1979). "Polygon Overlay with LANDSAT Data: An Application of ODYSSEY," Unpublished L.C.G.S.A. Internal Report.

Nisen, W. G. and Hunt, A. M. (1979). "Dual Techniques for the Determination of Spatial Clustering of Mortality in Montreal, Quebec - 1972," Paper presented at Auto-Carto IV, Reston, Va.

Puecker, T. K. and Chrisman, N. R. (1975). "Cartographic Data Structures," American Cartographer, vol. 2, pp. 55-69.

Shamos, M. I. and Hoey, D. (1976). "Geometric Intersection Problems," Seventeenth Annual IEEE Symposium on Foundations of Computer Science, pp. 208-215.

Teicholz, E. (1979). "Overview of the Harvard Geographic Information System: Odyssey Project," Conference Proceedings of the American Society of Civil Engineering.

White, D. (1977). "A New Method of Polygon Overlay," Harvard Papers on Geographic Information Systems, vol. 6. Reading, Mass.: Addison-Wesley.