JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 101, NO. B4, PAGES 8741–8743, APRIL 10, 1996

A global, self-consistent, hierarchical, high-resolution shoreline database

Pål Wessel
School of Ocean and Earth Science and Technology, University of Hawaii at Manoa, Honolulu

Walter H. F. Smith
NOAA Geosciences Laboratory, National Ocean Service, Silver Spring, Maryland

Abstract. We present a high-resolution shoreline data set amalgamated from two databases in the public domain. The data have undergone extensive processing and are free of internal inconsistencies such as erratic points and crossing segments. The shorelines are constructed entirely from hierarchically arranged closed polygons. The data can be used to simplify data searches and data selections or to study the statistical characteristics of shorelines and landmasses. The data set can be accessed both electronically over Internet and from the National Geophysical Data Center, Boulder, Colorado; it comes with access software and routines to facilitate decimation based on a standard line-reduction algorithm.

Introduction

With ever-increasing amounts of remotely sensed data, it often is necessary to perform intricate data searches and selections based on multiple criteria. A particular criterion is whether or not the data represent values over land or water. In many studies, shoreline data are used to construct logical data masks in order to manipulate and "mask out" parts of a large data set. In other cases, the intersection between data profiles and shorelines must be determined. Finally, in some studies the shoreline data themselves are the object of study. A systematic approach to these types of data retrieval and processing requires a self-consistent, hierarchical, high-resolution shoreline database. By self-consistent, we mean that all shorelines are represented as continuous closed polygons and the data are free of shoreline intersections or other artifacts caused by data inaccuracies. By hierarchical, we mean that the shorelines are ordered so that the polygons representing ocean-land boundaries may be distinguished from those outlining land-lake boundaries and also that each polygon can be ranked according to how much area it encloses.

We present a digital data set that fulfills these requirements. It was constructed from two well-known, public domain data sets. The World Data Bank II (WDB; also known as CIA Data Bank) contains coastlines, lakes, political boundaries, and rivers. These data have an approximate working scale of 1:3 million, meaning the features are considered to be accurately located on maps using that scale or smaller. The other data set is the World Vector Shoreline (WVS), which only contains shorelines along the ocean/land interface (i.e., no land-locked bodies of water). The WVS data set is superior to the WDB data set in quality and resolution (its working scale is approximately 1:100,000), but it lacks lakes. Although not explicitly given, the precision of the WDB data appear to be in the 500–5000 m range, while the precision of WVS is an order of magnitude better. We produced our data set using the WVS data when possible and supplementing it with WDB data. We obtained these data sets over the Internet. They are also available on CD-ROM from the National Geophysical Data Center (NGDC) [1994].

Paper number 96JB00104.
0148-0227/96/96JB-00104$05.00

Processing

To facilitate land/water determinations, it is necessary that the shoreline data be organized in closed polygons. Both the WVS and WDB data consist of unsorted line segments; no information is provided with them to indicate which line segments belong to the same polygon. In addition, polygons enclosing land must be differentiated from polygons enclosing water (e.g., land-locked lakes) since they may be used in different contexts.

The WVS and WDB together represent more than 100 Mb of binary data and close to 15 million data points. The large amount of data necessitated automatic procedures for data manipulation. Our first processing step was to remove point duplicates (repeated values) and outliers (identified as single points along the shoreline whose two immediate neighbors were identical.) In nature, no shoreline can cross another shoreline, but the digitized representations of shorelines often do cross; correcting such artifacts becomes a complicated processing step. Crossing segments were automatically edited, provided that only a few points had to be deleted. We determined crossover locations using the crossover routines of *Wessel* [1989]. Crossovers most likely arose because manual digitization often produces slight overlaps instead of exact closure. We found that the majority of segment crossovers were near the segment's endpoints. Hence endpoints were automatically removed until no crossings remained. In a few hundred cases the crossover and editing algorithms would have eliminated more than 5% of the points in a segment. In these cases we visually examined the data to determine (subjectively) which points to manually edit in order (1) to avoid crossings and (2) to keep the segment as close to its original shape as possible.

Next, we examined all loose segments to determine which segments should be joined to produce closed polygons. Because

most of the segments did not join exactly (i.e., there were nonzero gaps between some segments), we had to find all possible combinations of groupings and choose the simplest combinations (i.e., that gave the smallest segment separation). The WVS segments joined to produce more than 180,000 polygons, the largest being the complete Africa-Eurasia polygon which has more than 1.4 million points. The WDB data resulted in a smaller database, about 20% the size of WVS.

The next step was to combine the WVS and WDB databases. The main difficulty in this step was the presence of duplicate polygons: obviously, most of the features in WVS are also in WDB. However, because the resolution of the data differs, it is nontrivial to determine which polygons in WDB to include and which ones to ignore. We used two techniques to address this problem. First, we looked for crossovers between all possible pairs of polygons. Because of the crossover processing discussed above, we knew that there were no remaining crossovers internally within WVS and WDB; thus crossovers could occur only between WVS and WDB polygons. If crossovers were detected, they could indicate one of two scenarios: (1) A slightly misplaced WDB polygon crosses a more accurate WVS polygon, both representing the same geographic feature, or (2) a small WDB polygon representing a coastal lake crosses the more accurate WVS shoreline. We distinguished between these cases by comparing the area and centroid of the two polygons. In almost all cases it was obvious when we had duplicates; a few cases had to be inspected visually. Unfortunately, on many occasions the WDB duplicate polygon did not cross its WVS counterpart but was either entirely inside or outside the WVS polygon. In those cases we relied on the area-centroid tests.

We next had to assign a hierarchical level to each polygon. Here, level 1 polygons represent ocean boundaries, level 2 polygons represent lake boundaries, level 3 polygons represent island-in-lake boundaries, and level 4 polygons represent pond-in-island-in-lake boundaries. Level 4 was the highest level encountered in the data. To automatically determine the hierarchical levels, we compared all possible pairs of polygons to find how many polygons a given polygon was inside.

Once the hierarchical levels of the polygons were determined, we enforced a common handedness for all polygons, i.e., we arranged them so that when one moves along a polygon's perimeter from beginning to end, the area immediately to one's left is land. Thus level 1 and 3 polygons go counterclockwise, while level 2 and 4 polygons go clockwise. At this step we also computed the area of all polygons.

Examples

Shorelines are used in a variety of situations. For instance, an application in satellite radar altimetry might require a land/water mask made from all polygons enclosing an area larger than the altimeter's "footprint," that is, the region which backscatters radar energy, or an area perhaps 50 km^2 in area. The hierarchical and oriented polygons make possible graphics fill operations of features with holes in them, so that land areas may be painted with a mask which allows water-covered areas to show through, regardless of whether they are marine or lacustrine areas. This feature is useful for plotting maps of gridded data which are only valid over wet areas [e.g., *Sandwell et al.*, 1995].

Full resolution of the data set is vital when working in relatively small areas but becomes impractical for regional and global applications. It therefore is desirable to obtain reduced versions of the complete database, corresponding to different resolutions. Such decimation can be carried out using the Douglas-Peucker line-reduction algorithm [*Douglas and Peucker*, 1973]. The routine works to reduce the richness of texture along lines by removing points from the shoreline segment, thus giving a straighter line segment. This process depends on a tolerance value: if the removal of a point causes the resulting straight-line segment to depart more than Δ km from the actual data points originally between the segment's new endpoints, then the point is kept. Figure 1 illustrates the effect of the line-reduction algorithm on the shoreline of the island of Sardinia, Italy. Here, we have used values of $\Delta = 0.2$ km, 1 km, 5 km, and 25 km which typically lead to ~20% reduction in data size for each step in resolution. These five data sets, derived from the data set discussed in this note,

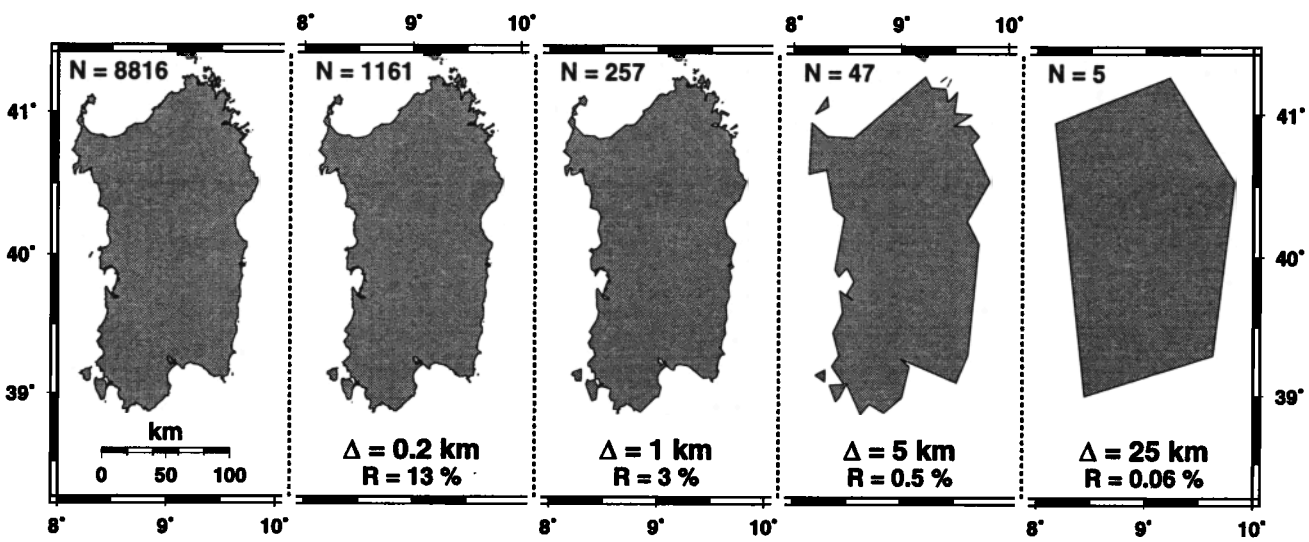

Figure 1. Example of how the very detailed polygon representing the island of Sardinia (8816 points) may be reduced by choosing various tolerances. *N* indicates the number of points in the polygon. *R* represents the percentage of points in relation to the original full resolution polygon (on the left).

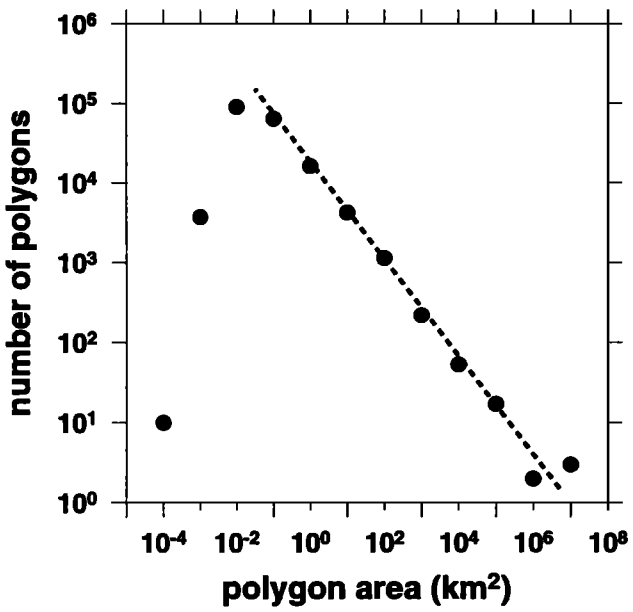

Figure 2. Land polygons follow a powerlaw distribution over a large range of polygon areas. Departure from the linear trend at small areas most likely represents undersampling.

make up the binned shoreline data distributed with the Generic Mapping Tools (GMT) [*Wessel and Smith*, 1995]. (The GMT software package also contains tools to create data masks based on these five resolutions, with the options to ignore small features as discussed above. However, GMT is not used to access the data discussed in this note.)

The user may create reduced data sets of arbitrary resolution using software archived with the data described here. The line-reduction algorithm may produce segments that cross one another, so that if it is applied without further processing as outlined above, self-consistency of the results cannot be guaranteed.

Statistics

The complete database contains 188,628 polygons representing 10,222,509 data points. The mean point separation is 178 m, with values ranging from a few meters to an extreme of 24 km in Antarctica. The largest polygon represents the combined Eurasia-Africa continents; it contains 1,435,084 points. The smallest polygon is a small arctic island near Queen Elizabeth Islands off northern Canada; it is made up of only 4 data points and has a size of approximately 175 m^2. For the most part, the distribution of land areas (level 1 WVS polygons) follows a power law (Figure 2), as one might expect of the contours of a fractal surface. The departure from this power law at areas less than 0.1 km^2 probably reflects undersampling of such small features.

Appendix

The shoreline data file is a single 89 Mb binary file using a simple, straightforward integer format that is described in the accompanying documentation; sample programs distributed with the data show users how to access the file as well as decimate it using the line-reduction routine above. Extracting data from the file can be done on a personal computer with very modest memory. However, the line-reduction routine (as implemented) requires ~36 Mb of memory to process the largest polygon; we therefore provide the four lower resolutions used in Figure 1 in addition to the full resolution set. All files and programs may be obtained from the World Wide Web at http://www.soest.hawaii.edu/wessel/wessel.html or from the National Geophysical Data Center, Boulder, Colorado.

Acknowledgments. R. E. Arvidson, D. Merritts, and J. H. Willemin provided thorough reviews. This work was supported by grant EAR-9302272 from the National Science Foundation. School of Ocean and Earth Science and Technology contribution 4048.

References

Douglas, D. H., and T. K. Peucker, Algorithms for the reduction of the number of points required to represent a digitized line of its caricature, *Can. Cartogr., 10*, 112–122, 1973.

National Geophysical Data Center (NGDC), Global Relief Data CD-ROM, Boulder, Color., 1994.

Sandwell, D. T., M. M. Yale, and W. H. F. Smith, Gravity anomaly profiles from ERS-1, Topex, and Geosat altimetry, *Eos Trans. AGU, 76*(17), Spring Meet. Suppl., S89, 1995.

Wessel, P., XOVER: A cross-over error detector for track data, *Comput. Geosci., 15*, 333–346, 1989.

Wessel, P., and W. H. F. Smith, A new version of the Generic Mapping Tools (GMT), *Eos Trans. AGU, 76*(33), 329, 1995.

W. H. F. Smith, NOAA Geosciences Laboratory, N/OES12, National Ocean Service, Silver Spring, MD 20910-3281. (e-mail: walter@amos.grdl.noaa.gov)

P. Wessel, Department of Geology and Geophysics, SOEST, University of Hawaii at Manoa, 2525 Correa Road, Honolulu, HI 96822. (e-mail: wessel@soest.hawaii.edu)

(Received July 11, 1995; revised December 6, 1995; accepted December 28, 1995.)