ELSEVIER

Computers & Geosciences 33 (2007) 325–334

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

# Fast algorithms using minimal data structures for common topological relationships in large, irregularly spaced topographic data sets

Thomas H. Meyer*

*University of Connecticut, Department of Natural Resources Management and Engineering, Storrs, CT 06269-4087, USA*

Received 20 January 2005; received in revised form 14 July 2006; accepted 17 July 2006

**Abstract**

Digital terrain models (DTMs) typically contain large numbers of postings, from hundreds of thousands to billions. Many algorithms that run on DTMs require topological knowledge of the postings, such as finding nearest neighbors, finding the posting closest to a chosen location, etc. If the postings are arranged irregularly, topological information is costly to compute and to store. This paper offers a practical approach to organizing and searching irregularly spaced data sets by presenting a collection of efficient algorithms ($O$(N), $O$(lg N)) that compute important topological relationships with only a simple supporting data structure. These relationships include finding the postings within a window, locating the posting nearest a point of interest, finding the neighborhood of postings nearest a point of interest, and ordering the neighborhood counter-clockwise. These algorithms depend only on two sorted arrays of two-element tuples, holding a planimetric coordinate and an integer identification number indicating which posting the coordinate belongs to. There is one array for each planimetric coordinate (eastings and northings). These two arrays cost minimal overhead to create and store but permit the data to remain arranged irregularly.


*Keywords:* Digital terrain model; Irregularly spaced data; Topological relationships; Triangulated irregular network (TIN)

## 1. Introduction

Topographic data sets are sets of triplets containing two planimetric coordinates and one vertical coordinate. These coordinates are either measured by automatic methods such as scanning laser altimeters (LIDAR) (Flood and Gutelius, 1997; Baltsavias, 1999), interferometric synthetic aperture radar (IFSAR) (Hodgson et al., 2003; Gamba and Houshmand, 2000; Mercer and Schnick, 1999), or by manual compilation with methods like photogrammetry and ground surveying. The terrain samples are called *postings*. Automatic terrain sampling methods produce irregularly spaced samples either by design or simply due to uncontrollable environmental factors such as the wind turbulence jostling the aircraft carrying an instrument. Samples collected by manual methods are frequently arranged irregularly by choice in order to capture breaklines and other important features that define the shape of the topography; irregularly spaced postings capture the shape of the terrain and the features thereon better than gridded postings

*Tel.: +1 860 486 0145; fax: +1 860 486 5408.
*E-mail address:* thomas.meyer@uconn.edu.

0098-3004/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2006.07.002

(Makarovic, 1977; Gould, 1981; Douglas, 1986). Additionally, some applications require irregularly spaced data. For example, the US National Geodetic Survey maintains a database of high-accuracy survey control markers and provides web-based applications that allow a user to query the database to find all markers within a certain distance of a point of interest. The published coordinates of these markers must not be changed by their representation in the database; they must remain irregularly spaced. Also, gridding the data can impede feature detection (Cooper and Cowan, 2004). Digital terrain models employing irregularly spaced postings are common and useful; the triangulated irregular network (TIN) is probably the most common example of the type.

Many terrain analysis algorithms depend on the topological relationships between the postings. In particular, many algorithms require *neighborhoods* of postings that are close to one another in some sense. For example, the computation of gradients (Meyer et al., 2001), curvature (Shary, 1995; Ozkaya, 2002), semi-variograms (Isaaks and Srivastava, 1989), kriging (Hessami et al., 2001), roughness metrics (Philip and Watson, 1986), cluster analysis (Gebhardt, 2001), feature recognition (Cooper and Cowan, 2004), and fractal dimensions (DeCola, 1989; Cheng, 1999) are defined over neighborhoods. For gridded data, two typical neighborhoods are the four cardinal postings around the point of interest or the cardinal postings plus the diagonals. For irregularly spaced data, the situation is less clear. One popular way to determine sets of nearest neighbors for irregularly spaced data is to construct the Delaunay tessellation of the postings. Then, for some posting $p$, take the nearest neighbors of $p$ to be those postings that share an edge in the tessellation with $p$. This solution is elegant and satisfies the goal of "letting the data speak for themselves" (Gould, 1981), but a Delaunay tessellation requires considerable time to compute and space to store. These problems can be intractable given the size of many topographic data sets. For example, as of the time this article was written, at least one commercial LIDAR sensor can collect samples at 70 000 Hz with sub-meter posting spacing (Optech, 2003). At this rate, a one-hour flight of this sensor would collect more than $2.5 \times 10^8$ samples.

The time needed to compute the inherent topology of data sets as large as these would be prohibitive to most users. Therefore, large topographic data sets are usually gridded and the resulting loss of accuracy is simply accepted. This paper offers an alternative, a way to have the accuracy of irregularly spaced data without unacceptable computational and storage burdens of complicated data structures such as Delaunay tessellations (Mortenson, 1985, p. 317), quadtrees (Samet, 1990; de Berg et al., 1998), k-d-B-Trees (Bentley, 1975; Robinson, 1981), hB-Trees (Lomet and Salzberg, 1989, 1990) or R-Trees (Guttman, 1984); see Nievergelt and Widmayer (1991) for a survey. This paper presents several simple and efficient algorithms that compute the basic topological relationships needed for algorithms requiring neighborhoods for inputs. These algorithms depend only on two simple data structures, namely, two sorted arrays.

## 2. Supporting data structure

The following discussion depends on sets and the elements thereof. The $i$th element of a set $P$ is denoted $P_i$. Conversely, we denote that element itself with $p^i$. Thus, $P_i = p^i$.

An individual postings is typically a set of values including three spatial coordinates plus other ancillary information such as an intensity value, a time stamp, a return number, etc. Define a posting to be a set $p^i = \{e^i, n^i, u^i, \alpha^i, \beta^i, \ldots\}$, where $e^i, n^i, u^i \in \mathbb{R}$ are the posting's easting, northing, and height (up) coordinates, respectively. $\mathbb{R}$ denotes the set of reals, and $\alpha^i$, etc. are additional attribution fields holding ancillary information of no particular type. Let $p^i_e$, $p^i_n$, $p^i_u$ denote the easting, northing, and up coordinate of posting $p^i$ and $P = \{p^1, \ldots, p^{\mathrm{N}}\}$ denote the given posting data set. Thus, $P_{i,e}$ is the easting of $p^i$. Define the index set over $P$ to be $I = \{1, \ldots, \mathrm{N}\}$.

Our strategy is to decompose $P$ into three arrays. One of the arrays, $\mathscr{N}$, is a sorted array of northings together with an index indicating which posting that northing came from. $\mathscr{E}$ is a sorted array of eastings together with an index into $\mathscr{N}$ indicating which northing that easting was paired with. The last array, $\mathscr{P}$, contains the attribution fields of $P$ plus an index into $\mathscr{E}$, thus forming an index loop: knowing an easting leads to the northing associated with that easting; knowing a northing leads to the attribution information and a pointer to the associated easting; and knowing a posting leads to its easting. Therefore, given any tuple in any of $\mathscr{E}, \mathscr{N}, \mathscr{P}$ allows the entire original posting to be reconstructed. For notation, let $\mathscr{N}_i = \{n^i, \eta^i\}$, meaning $\mathscr{N}_i$ is the $i$th tuple of the sorted northing array, $n^i$ is the northing coordinate