19. ACCURACY AND SPATIAL RESOLUTION: CRITICAL DIMENSIONS FOR GEOPROCESSING

Michael Goodchild

University of Western Ontario

Accuracy and Spatial Resolution have been referred to many times in the last couple of days and I would like to present some thoughts on that area, as I have been working in it for quite some time.

In any kind of of geographical data processing we probably have two sources of error. One is the conventional error that is present in the data, and will be present in any kind of data, whether we were taking about geographical data or not. The other kind is the error that is present because it is spatial data, and this is error in location, in the location of points, lines, or whatever. It is essential to realize that it is error of a particular kind because it is error that results only partly from data collection. People put things in the wrong places. They aren't quite accurate about where things are placed. But it is also error that is deliberate because it has resulted from the various processes of generalization that have been built in along the way, by cartographers drawing smooth lines where they are really jagged, and by the digitizing process. When we talk about generalization and accuracy it's a two edged thing. In some ways it's useful and in some ways it's not.

Spatial resolution crops up in many places and I would like to mention four of them. The first one is the case of point data -- data which refers to, say, oil well locations, where an error is present possibly because the location has not been specified precisely. But error may also be present because we are possibly using a point to represent something which is not a point. An oil well may have a point-like location at the surface, but because it meanders underground, it's real location is not simply that point. So, to some extent there is an error because the abstract concept of a point is being used to represent something which is actually spread over an area of space.

The second kind of error I would like to discuss is that relating to line data, where in most cases, when we are talking about vector data, we're talking about representing a line by a series of straight lines between points, and that in itself is probably some kind of generalization or abstract representation of the true line. The true line is itself an abstract representation of what in reality is not a line at all. It may, in fact, be some kind of zone. A fault zone may be represented by a simple line on a map, which is itself a generalization. So there are two stages here of generalization, first from the real entity which may be a zone to the cartographic representation as a smooth line, and then to the digital representation as a series of straight lines between points. Possibly one exception to that would be the legal description, where the reality that we are really trying to describe is itself an abstract line between points. The legal description line may not be subject to that first stage of generalization.

The third kind of error is in areal data. We're in the habit of drawing a zone on a map and implying that everything inside that zone is the same. That too is a generalization. There's an error present because the real boundary of the zone might not be the simple line, but might instead be a zone of graduation from quality A to quality B. What happens inside the zone may not be homogeneous, giving us a second source of error. And finally, in the digital representation the boundary is probably only represented, again, as a series of straight lines or a polygon, which is itself a misrepresentation of reality. So again there are a number of stages of misrepresentation and generalization, and what eventually gets represented in the computer may be a very long way from what we started with.

Now, let's compare what happens in digital cartography with what might happen in a manual process. In a manual process, the original entity is described on a map in a very well-informed manner, usually by some kind of expert, and so, although there is a generalization process going on it is usually an informed generalization. It is subject to processes which may be very complex because they represent a lifetime of professional experience. The same is likely to happen in the digital context because the original data for digital cartography is probably a manuscript which has been created in the field, again by a well-informed person. In the digital case, though, we now go through another stage where we take that original manuscript and place it in the computer. Now whether this is done by a digitizer operator or by some kind of scanning system, there is a major difference because we are going through a process of generalization which is not usually as well-informed. The digitizer operator is not usually a professional who understands the nature of the data. So, although we can see generalization taking place both in the manual and in the digital cases, in the digital case some of the generalization is not well-informed, whereas in the manual case professionals or at least experienced people, are usually in charge of the whole operation.

A couple of general points about spatial resolution. Why is spatial accuracy critical in any geographical data processing? It is critical because spatial resolution determines the number of bits that are required to contain each coordinate. It determines the number of coordinates that are required to represent a line, and so, ultimately, it determines the volume of data which is generated and therefore the processing costs. We therefore can say with some certainty that in 90 per cent of cases spatial resolution is the primary determinant of pro-

cessing costs and storage volume, or storage costs. So it is obviously a critical parameter for any kind of geographic data processing, and we can only minimize costs at the expense of accuracy.

Spatial resolution and generalization is a very complex topic. What I would like to do is to give you some idea of the scope of the topic by presenting some typical problems. I've picked out four problems which I would like to talk about briefly, and which arise in the geographical data processing context. Each of them carries a different message.

The first one is what we call the polygon overlay problem. Suppose we have a set of areal data, which contains a zone that has been labeled soil type "A". We might use a geographic data processing system in order to consider the combination of this particular source of information with some other map which might show different kinds of forests. Imagine a zone of forest type "B". The purpose of this exercise is to find out how much land is soil type "A" and forest type "B", with the purpose of measuring the zone of overlap.

The ability to do this is often presented as one of the major advantages of geographical data processing, particularly of polygon systems, that is, the ability to overlay and compute the amount of land which has joint characteristics. A problem which has cropped up in several tests of this kind of thing, is that the overlap zone, supposedly soil type "A" and forest type "B", is in fact not this. If you select a random point in that zone and go into the field and ask "Is it really soil type A and forest type B?", the probability of being correct can be very, very small. The problem is, of course, that the person who originally drew a line around soil type "A" was generalizing; the one who drew a line around forest type "B" was generalizing, and so in the intersection area the probability of having both these characteristics can be very small indeed.

What is message here? What does this illustrate? The point I would like to make is that this isn't really a digital problem -- it is a problem that arises because we are using digital technology, which is highly accurate, in the context of very dirty data. And, of course, the message is the old one of "garbage in, garbage out". We are pretending that the zone really does represent the area of soil type "A" -- when in fact it doesn't. What we tend to do then is to blame the input, and say that if that really was all soil type "A", the data processing system would be doing a good job. It's not because the input data is poor, however. The point is a bit more subtle than that. The problem here is that soil data really does tend to be poor data by its very nature. To pretend that we really could have better soil maps is really to be dreaming. If we accept that the data really must be poor, then the problem is really one of justifying this highly accurate system of geoprocessing. That, I think, is the message. Why have a highly accurate polygon overlay package to superimpose data which is fundamentally poor?

Number two is a very familiar problem with polygon data, or vector data in general. This is what we call the sliver polygon problem. This is the problem of having two versions of the same line, which have been digitized on separate occasions, or digitized by different people. Somewhere along the geoprocessing stream those two versions of the line are compared and fail to coincide. It is a problem which pervades polygon data systems. CGIS suffers from it, and so do similar systems. The paradox of the problem is that almost any digitizing method one cares to name produces, or tends to produce, close to the maximum number of sliver polygons. Furthermore the problem tends to get worse as the digitizing gets more and more accurate. I've done a lot of work on this, particularly from the point of view of trying to delete them. It's a two-sided problem, because on the one hand is the problem of detecting which of the polygons are slivers and which are real, which is a very subtle problem which we have tackled on the basis of area and shape and number of sides and geometric characteristics of the slivers. The second side of the problem is the problem of deleting them once they have been detected. I have worked on algorithms of this nature which replace the two sides of the polygon with a single line, with some success.

The third example I would like to mention is the raster problem, which is very different from the other two because rasters have a very specific and highly identified level of spatial resolution. Once you have decided to represent something by a raster, the raster size determines the spatial resolution. The problem I would like to look at here is that of measuring the size of a patch from a raster representation. If you like, a typical application of this is measuring the area of a field from a pixel representation from remote sensing. This is one area where I think I can really say that the problem is now very well understood. We have good methods for assessing the accuracy with which a particular pixel size represents a zone. We can tell, with a fair degree of precision, what the level of accuracy will be in an estimate of area, given the pixel size and given information on the degree of contortion of the outline.

The fourth is the Digital Terrain Model problem. This is something that David Mark and Jim Little talked about. Here I would like to present one problem which illustrates spatial resolution from another point of view. Suppose we have a rectangular Digital Terrain Model of terrain, such as output from a Gestalt photomapper. The problem here is contouring from the spot heights which are a representation of the terrain. This is somewhat paradoxical problem too. If the terrain is very rugged, there is no particular problem. But consider what happens as the ground gets flatter and flatter. As we get towards a very flat surface, the very

small errors in the Digital Terrain Model have the property of tending to produce an extremely contorted contour, with spurious islands and spurious depressions. The problem gets worse and worse as the terrain gets flatter and flatter. It is a very real problem which David Mark and I are working on for Energy, Mines and Resources at the moment. We would like to find a way of replacing a contorted contour with something that is smoother and more acceptable cartographically. Again this is a problem of using statistical methods to discriminate between real and spurious features.

These are four examples of ways in which spatial resolution becomes particularly critical in geographical data processing. I would like to end with a few summary points which come out of the spatial resolution problem. First, a geoprocessing system is probably more accurate than the data that goes into it, so that the problem of accuracy within the system is really not very important. We can probably build a system which is at least as accurate, and probably more accurate than the data which goes into it. What geoprocessing tends to do is to expose the inadequacies in the raw data in a way that manual cartography doesn't. It forces us to deal explicitly with them.

The second general point is that geoprocessing tends to require that many of the subjective processes that go on in conventional, manual cartography, like line generalization, become objective. It requires that we write down algorithms for line generalization, whereas the cartographer, of course, works from his own intuition. This is a particularly acute problem, because many manual processes would give extremely complex algorithms if they were written down in any effective way. There is a temptation then to replace complex manual procedures with over-simplified algorithms. One example of that is in automated contouring, where we might have a system of contours produced by a digital system and then, superimposed on that, a digital representation of a streamline. But the streamline persistently misses the kinks in the contours. A cartographer would immediately correct the problem by placing the stream in the kink, but we would require quite a complex algorithm to detect and correct the problem automatically in a digital system. The point then is that the digital versions of conventional generalization methods are very much more complex than the generalization algorithms we currently use.

The third point is that in talking about accuracy in geoprocessing we are forced to treat accuracy as a parameter, as a measurable quantity, instead of treating it as something that a cartographer can take care of. A manager at some point along the line has to determine the accuracy which is required from the system. That's a very awkward thing to do, because with few exceptions we really don't have much in the way of suitable measures for describing accuracy in space, and we know very little about how accuracy can be paid off against the cost and benefits of a geoprocessing system.

A final point: Talking about accuracy really tends to expose our lack of understanding of spatial variability and spatial statistics. The real answers lie in that area, and these problems force us to pay a lot more attention to the question of variation in space.

DISCUSSION:

YAN: I'm not a cartographer. I come from computer science. Now when you say that cartographic generalization is professionally done, I wouldn't question that, but I would question repeatability, given several professional expert cartographers. Would that be any more repeatable than several digitizations? Are you sure that it would be more repeatable?

GOODCHILD: You mean would a manual generalization be more repeatable? No, I don't think it would. I think every cartographer would do a different generalization. But, I am inclined to think that each of those might be more acceptable than a digital generalization. Digital generalization would be repeatable yes, but at a cost. So the repeatability isn't that important.

YAN: I thought repeatability was an important aspect of accuracy.

GOODCHILD: A statistician would believe that very deeply, yes.

YAN: You get an artistically better product, but not necessarily a scientific one.

GOODCHILD: Yes, that's it.

GOLD: What we're after is to convey information. We as humans have an ability to make interpretations based on very many subtle things that computers aren't smart enought to handle. So I, as a geologist, for example, can tell a lot from the inflections of contours, which are not immediately obvious to others. I have to perceive the third dimension from some very complex and subtle hints. Only another geologist who is preparing that map is able to convey those impressions which he has obtained.

GOODCHILD: Yes, and the digital version of that would be extremely complex. My point is that the current versions of generalization algorithms, because they are not that complex, tend to produce results which are somewhat less acceptable.

MARK: Another, by way of a comment, which you may wish to take off on, is the concept of acceptability of output, which is interesting. We found that people didn't like it

when we contoured triangles with straight lines across the triangles which had corners where you crossed from one triangle to another, even though that was the least biased estimate of the contour. I guess it's ten years ago now that David Douglas was talking about rounding the corners off and putting spurious wiggles in, just to make contours look better because people knew that contours were never straight. This fractal introduction of detail, spurious detail, of the right character, is another example of that.

GOODCHILD: Yes, we do get to the point of saying that putting in spurious detail has serious purpose.

BOYLE: I understand that the geologists think that automatic methods destroy the anomalies, and they are looking for these very anomalies.

MEGGITT: Your example about taking out spurious contour loops: doesn't one get that by smoothing with spline functions?

GOODCHILD: Yes, but I would want the parameters of that smoothing to vary across the map, because smoothing should be greatest in flat areas.

MEGGITT: I wasn't thinking of contours which were derived from points. I was thinking of features which are lines in the first place, such as rural boundaries and that type of thing. But even rural boundaries relate to the topography. Although there are certain other things which are perhaps not that way, such as the flight line of an aircraft, most are. I am particularly concerned about the fact that in these two days no one has really come to grips with the question of geodetic precision and where it is important and where it's not important, and what sort of numbers you are going to put on that which is important. In your example about the soil type "A" and the forest type "B", there was no homogeneity in the arbitrary polygons you chose to call soil "A" and forest "B", so there are apparent errors there, and all that is accepted. That often comes with lots of data, especially in urban areas. But in other cases you may want to have a really well nailed down control point here or there, because of the propagation of errors. Is it not highly desirable to make an effort to obtain really good precision. You have got to have in mind the most exacting use to which people may put the data. Of course, legal boundaries are a typical example, but they are not the only ones.

GOODCHILD: Not everything needs to be at the highest common denominator of precision. It's difficult, of course, to generalize for different kinds of lines. For legal descriptions, which tend to be straight line segments with relatively few points, vector representation is very efficient, and we can afford to have large numbers of significant digits. A resource management information system with a representation of a wiggly river is a very different matter. I don't think it is something one can generalize. I think every system has to have its level of precision, and in some systems the required level of precision is extremely high. One can afford to make it high because the volume of data is not particularly great.

It's very hard to throw away data. It's like giving up part of yourself. I think it is unfortunate that some systems which have been developed in the past have not allowed unneeded data to be discarded. The data in resource management systems does not justify the level of precision the system has, and yet the system is paying to carry that precision.