

GIS – The Aerial Perspective

Why we are drowning in data

Stanton H. Moll

GIS/RS Program Manager

When I first started in the GIS business it was possible to know and understand how to deploy every function of one's favorite GIS or CAD software package. The domain of knowledge required to be an adequate technician was not so great that a reasonably intelligent person couldn't master it. Unfortunately (perhaps) it is not in the nature of technology to rest on its laurels, to say "that's good enough" and to move on to something else. There is always a greater problem to solve, a more sophisticated way to do things, an exception to the way we've always done things, an urge to build an edifice larger and grander, more data to process.

And therein lies the problem: there is always more data to process. The data fluxes – the volumes of data being created – are growing faster than our ability to render it into a useable form. Raw data is a fine thing – ask a data provider – but it requires reduction or some type of context to make it much more than a pretty picture. Data is pretty much useless until rendered into information. Software is increasing in complexity and permits us to make ever more subtle characterizations, but in almost all cases a human is required to perform the "value-added" conversion of data into information.

Take for example an aerial photograph. A typical 9" air photo contains about 2.5Gb worth of information. An air photo is nice to look at – depending on scale you can see your yard, the swimming pool in your neighbor's yard, his impeccably manicured lawn and flowering crabapple tree. You can probably tell those items for what they are, but to a computer they are merely pixels, a collection of three or four numbers describing the brightness of the component colors. The computer can't tell a red rose from a red Corvette.

So in order to make that red object something useful you generally have to park a human in front of the computer to evaluate the imagery and derive *information* from the data. A photogrammetric compiler does this all day long, describing the boundaries of vegetation features, the edges of buildings and roads, and in general the geographic location of a universe of objects visible on the imagery. Even so, the context of the object in the image is terribly important – a single pixel is pretty useless.

In fact, I would argue that basically an image by itself is pretty useless. Most geographic applications of imagery, setting aside the orthoimagery subset for a moment, involve human interpretation to create *vector information*, which is a form of information much more readily useable for further human analysis. Think of almost any domain of human knowledge and the most useful information is in a vector form: geology, wetlands, urban land use, forest stand inventories, road networks, utility infrastructure, hydrology. An archeologist will receive an air photo or a satellite image and immediately begin drawing on it.

An orthoimage might seem to be different – it has, after all, been geometrically improved – but it too is most often used as a backdrop for vector information. It is the contextual basis for further interpretation and information extraction. It is becoming increasingly common for clients to reject the option of buying planimetric information, opting instead to purchase an orthoimage and extract the planimetry themselves.

Topography is one major, and more complicated, exception to this thesis. Elevation information is generally more analytically useful in an image, or *raster*, format. Good algorithms exist for working in surface analysis such as hydrology, bathymetry, viewshed, etc. Visualization is generally far more rewarding using data with elevation than with the data alone. However, contours, i.e. vector topography, are also very useful for many purposes. There is a large variety of ways in which topography can be collected, stored, and utilized.

Which brings me back to my main point – we are suffering from a flood of raw geographic *data* that will overwhelm us if we cannot figure out how to extract the *information* we need from it. Traditionally we have been using human labor to collect and extract information from data – a photogrammetric compiler, or a surveyor. Computers were a boon to our industry because they helped increase productivity many fold – who wants to go back to using scribe coats instead of CAD systems, or doing traverse calculations by slide rule? But the growth in computing power is threatening to be overtaken by the growth in data storage and acquisition capacity.

Everyone by now has heard of Moore's Law, which states that the densities of transistors on a computer chip will double every 18 months. This has been popularly interpreted to mean that computing power will double in that period. According to Intel this has proven true since Gordon Moore coined the Law in 1965, although in reality the period has been closer to 24 months per doubling.

However, according to Scientific American, Kryder's Law claims that disk storage capacity is doubling every 18 months *or faster* (doesn't everyone own an iPod by now?). In 2000 the 6Gb drives were common; by 2004 we were using 120Gb drives, and now you can exceed 500 Gb in a single unit. This not-insignificant rate of divergence means we will be storing ever-larger amounts of data faster than we can process it. According to Wikipedia there is a corollary of Parkinson's Law that says "data expands to fill the space available for storage".

We are now seeing this growth in the rate of raw data collection. Digital cameras, satellite imagery, and especially LIDAR units are beginning to overwhelm our capacity to extract information from them. We ourselves in many cases went right past DVDs to external hard drives to store the massive amounts of data collected by our LIDAR system.

The implications of this are significant. If we continue to employ human labor to extract the information in these increasing volumes of data we will quickly outpace our capacity; we will need to train more analysts, outsource the work, or cut back on the amount of work we do.

I view the solution to managing and harnessing this growth of data fluxes to be with software. We need software to do more than just manipulate data faster, although that

will help – it needs to fulfill the promise of artificial intelligence and extract information from data. We are beginning to see success in a variety of fields relevant to the spatial sciences: automated terrain extraction; automated aerial triangulation; segment classifiers; pattern and object recognition.

So if you feel like you're drowning in data, think of it as having your glass half full and rising. Data are the first steppingstone to human knowledge; we need new tools to harness that data into the second steppingstone, information. Support your local toolsmith, the university.