Field Techniques Manual: GIS, GPS and Remote Sensing

- Section C: Techniques

Chapter 7: GISci Analysis

7 GISci Analysis

7.1 Introduction

A GIS is not simply a tool for storing data from different sources (images, maps, points), nor just a method for relating points on a map to a database. The real advantage of a GIS is that it allows the display, analysis and output of spatially-related data. This is made possible by the geographical framework under-pinning all of the data layers. GISci analysis is when all the hard work of integrating spatially-referenced data starts to pay off! In this chapter we start by looking at straightforward GIS functions: *mapping* layers of data and graphically *symbolizing* their attributes so that patterns may become apparent. *Spatial selections and queries* allow features from one layer to be extracted from an area defined by another layer. Similarly, *spatial overlay functions* provide a useful way of manipulating data, using the extent of one layer to add or subtract from another or to combine the two. *Buffering* and *proximity analysis* are useful when looking at phenomena within (or beyond) a certain distance from another. More advanced techniques include *correlation*, to look at the relationship between different layers.

Several techniques are used when a 'third dimension' (such as height, rainfall, temperature) is being considered. *Interpolation* is used to estimate values between known data points, while *contouring* is used to draw lines around existing values, as on a normal height contour map. *Triangular irregular networks* provide an efficient way to represent and analyse complex '3D' surfaces. Zukowskyj and Teeuw (2002) compare the performances of various GIS software in carrying out interpolation along a river valley.

This chapter also looks at data *topology,* which defines the relationships between points, lines and areas in geographic data. This seemingly obscure topic becomes relevant when, for example, mapping roads or analysing area data such as vegetation types. When data contained in several layers determine the pattern in another layer, for example, vegetation and soil determining the distribution of a plant species, then *modelling* techniques can help to predict or understand the distribution of that plant. Finally, there is a description of statistical techniques for characterising and analysing data, useful for both spatial and non-spatial data sets.

Most of the techniques described here relate mainly to vector data - points, lines and areas. The next chapter looks at the analysis of raster data and digital image processing. However, some techniques here concern 'surfaces' of values being calculated or predicted, and the raster model proves useful in these cases. A summary of suitable software for all of these functions is given in Chapter 1.

7.2 Mapping and symbolizing

A visual assessment of field data can be made using different symbolisation techniques: Figure 7-1 shows a variety of styles available in the MapInfo GIS package.

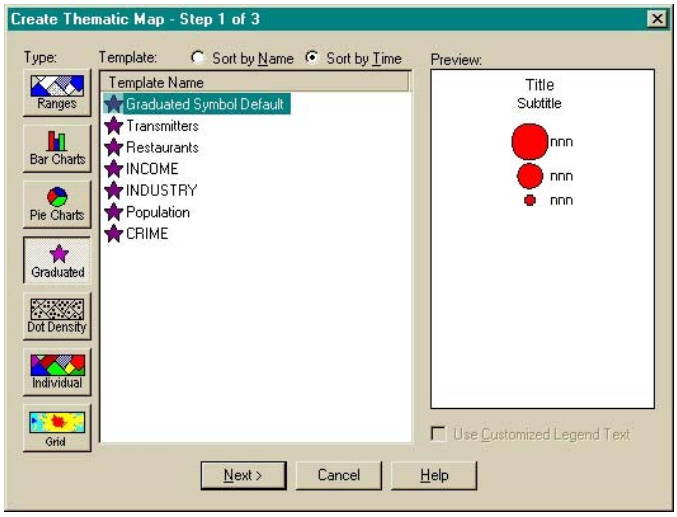


Figure 7-1 Defining plot types based on attributes.

For example, records of zebra sightings in the Mkomazi Game Reserve, Tanzania, were mapped out using circles whose diameter is proportional to the size of each group. This is shown below in Figure 7-2. Simple mapping techniques like this can summarise a large quantity of field data in a readily accessible way, as well as providing a useful overview prior to further analysis.

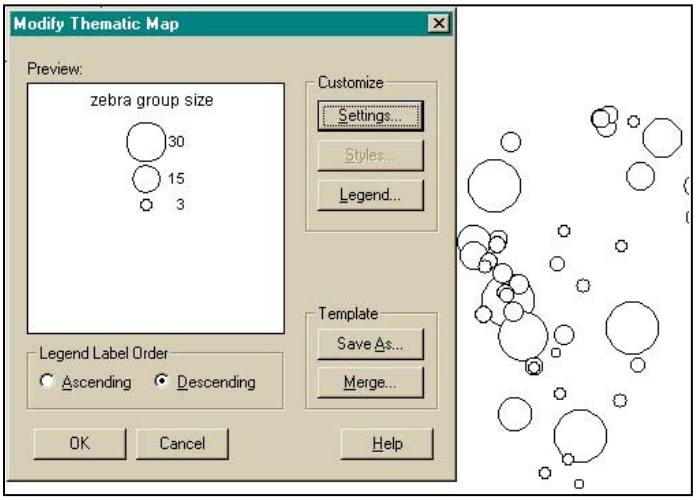


Figure 7-2 Defining actual plot sizes.

7.3 Spatial selections and queries

7.3.1 Simple data queries

A simple 'conventional' database query might select records from a GIS dataset based on some attribute or value within the data. For example, in the Bogda Shan set of GPS readings, we might select all point with an altitude above 1,000 m. Making such a selection is a function available in most GIS programmes, or in database terms a function that can be entered with a Structured Query Language (SQL) statement (see Chapter 4). The resulting selection can be mapped, as in Plate 6.

However, a key benefit of using a GIS is being able to query data not only according to its attributes, but according to its geographic position. This might take one of several forms:

- Selecting all features in one layer that lie within another layer (see *Figure 7-3*), e.g. all zebra sightings made within an area of grassland, or all occurrences of a plant within a certain country.
- Selecting all features that lie within (or beyond) a certain distance of another feature e.g. all woodland that lies within 1 km of waterholes, or all houses within 10 km of a volcano (See Plate 7).
- Selecting all features within an area drawn on the map – this can be a regular area such as a square or circle, or an irregular shape drawn with a 'lasso'.

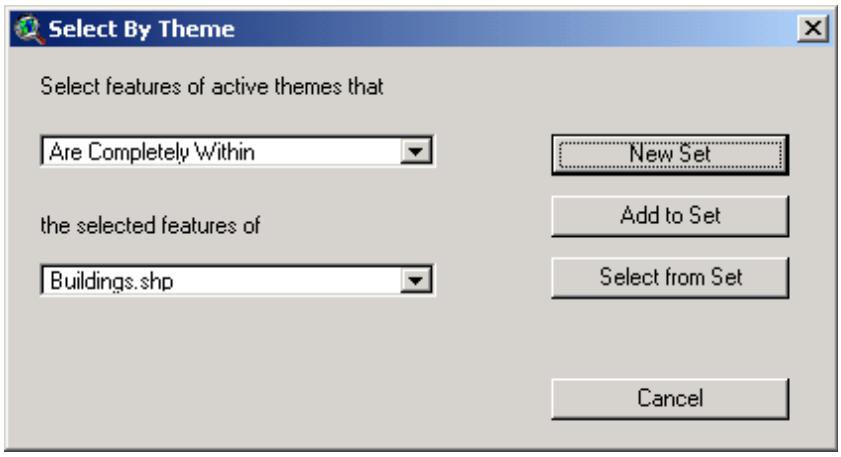


Figure 7-3 Using a 'Select By Theme' query in ArcView 3.2 to select the points in one data layer that lie within areas covered by buildings.

In all of these cases, point, line or area features can be selected. Such spatial and attribute queries can be combined: for example, selecting all trees within 1 km of a waterhole (spatial query) and above 5 m in height (attribute query).

7.4 Creating 'buffers'

GIS can create new layers comprising areas within a specified distance of existing features – these are known as buffers. *Figure 7-4* gives an example, where different levels of error are shown as buffers around a series of GPS readings. In fact, in this case three buffers have been created, showing the distribution of likely GPS positional errors within one, two and three standard deviations respectively. In this case, buffers have been used to visualise error analysis but there are many other expedition applications: areas prone to flooding within a given distance of a river; an 'exclusion zone' around breeding bird colonies; or areas of radio coverage, for example.

.

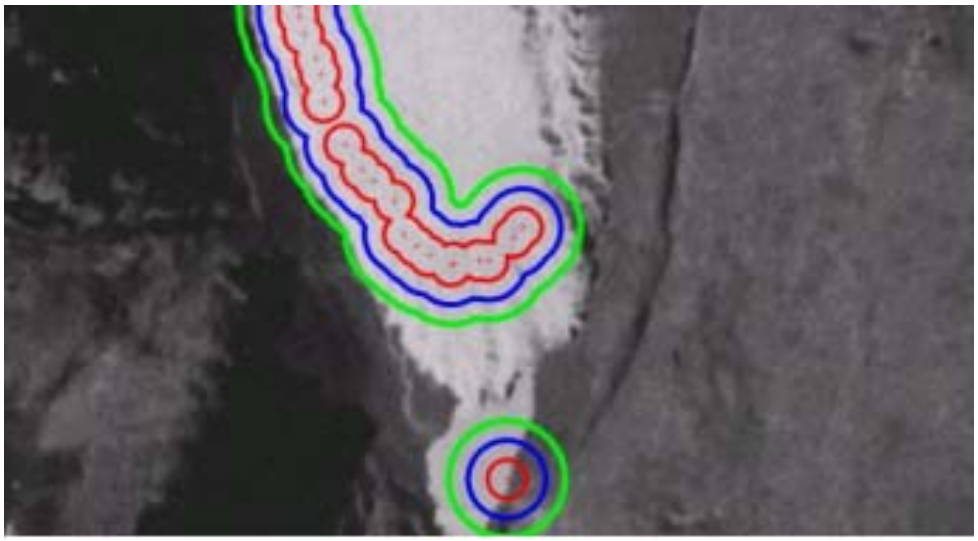


Figure 7-4 Buffers around GPS points, at 15 metre intervals, indicating the likely positional error.

7.5 Testing relationships between data layers

In some cases the relationships between data will not be easily measurable and some level of uncertainty may exist. This section touches upon testing hypotheses and the confidence that can be placed in an analysis. It is common for an expedition's project to involve testing some theory or hypotheses. In a GISci context, this typically involves an investigation of whether the distribution of one phenomena (e.g. vegetation) is significantly associated with another (e.g. elephant distribution). There may or may not be a causative relationship between the two: does the vegetation determine where elephants are seen; do elephants determine the vegetation through their feeding; or does some other factor influence them both at the same time?

This manual does not aim to describe the statistical methods needed for such tests. Rogerson's *Statistical Methods for Geography* (2001) is recommended, as it presents statistical techniques and issues with a strong spatial emphasis, as well as setting them in a GIS context: "GIS do not reach their full potential without the ability to carry out methods of statistical and spatial analysis".

The chi-squared test is an example. It measures the probability that the differences between two datasets have occurred by chance – or put another way, it indicates the similarity between two datasets. It is often used in GIS and geographical studies, as it makes few assumptions about the data being used: they do not need to be normally distributed (see the final section of this chapter), small sample sizes can be sufficient, and categorical data (such as soil types) and ranked data (such as a list of villages ordered by population size) can be used as well as numerical data. Many other methods are used to test whether sets of variables are associated and to quantify the relationships between them. The chi-squared test considers just two sets of data, but other methods exist for the analysis of many sets of variables – see Rogerson (2001) or other statistics textbooks.

7.6 Spatial overlay operations

We have seen how one data layer can be used to *select* data from another. We now look at overlay operations, where one layer is used to *change* another. The general principles are

shown in *Figure 7-5*: layers can be combined together (top), overlapped (middle), or cut (bottom). Different GIS programmes use different terminology for these operations; for example in ArcView and ArcInfo, these three functions would be called union, intersect and erase. Generally, one layer must be a polygon (for example, vegetation type), while other layers can be point, line or area/polygon (for example, animal sightings, roads or protected areas).

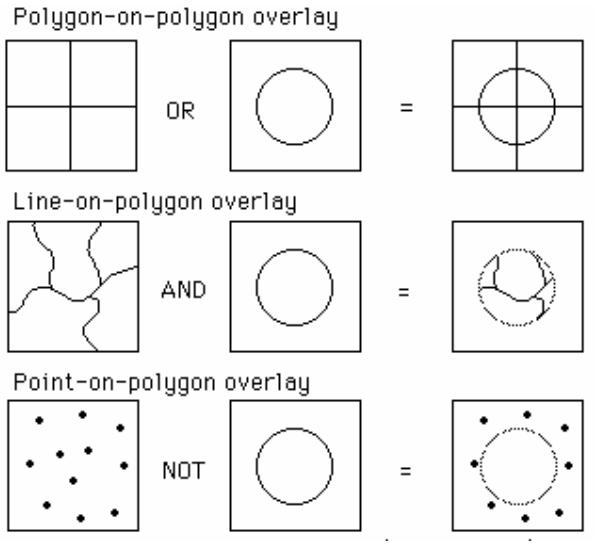


Figure 7-5 The principles of some spatial overlays using 'or', 'and' and 'not' logic. In these examples, point, line and area/polygon data types are all used. These three functions correspond generally to ArcView's 'union', 'intersect' and 'erase'.

These operations are commonly used in expedition GIS. In describing an animal habitat we might wish to find all rivers (represented as lines) flowing through open grassland habitats (polygons): this is an 'intersect' operation. We could also find all grassland lying more than 2 km from roads by first creating a buffer around the roads and using it for an 'erase' operation on the grassland layer. And we might use 'union' to combine two habitats, for example to measure their total surface area or to map a site for protection.

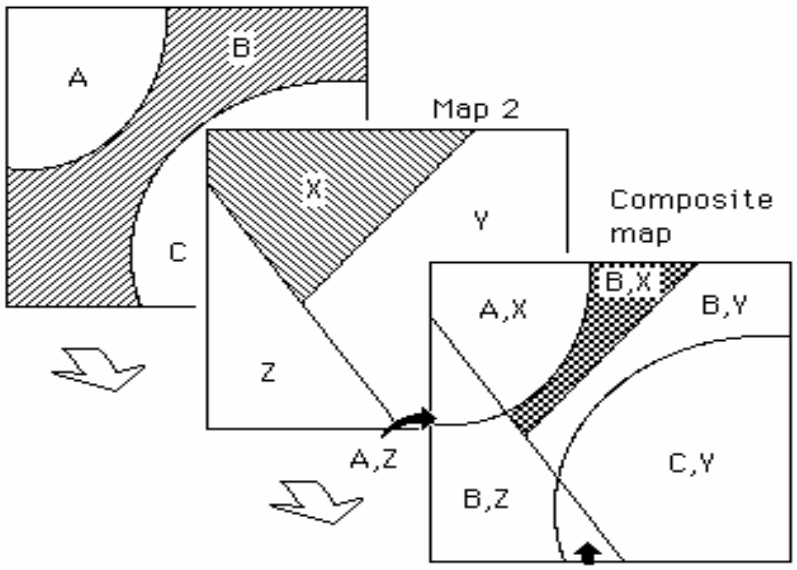


Figure 7-6 Example of the union of two polygon layers. Note that the attributes from each of the two original layers are all retained in the resulting output layer.

7.7 Spatial analysis

GIS analysis can be used to combine various datasets and query them effectively. Unfortunately, GIS are often seen as 'bins' for depositing lots of layers of information without delving into and manipulating / modelling the data. It is not until the team starts querying the relationships of these themes the true benefits of a GIS will be realised. This section of the manual uses two case studies to show how layers can be built up and used together to predict results. In the first example, the location of semi-precious minerals are modelled using limited field data and remote sensing techniques. In the second example similar techniques are used to predict the nesting sites of animals based on certain criteria.

In the first example from Moore *et al*. 2000, a GIS was used to determine the most likely occurrence of diamonds and jade from the Kunlun Shan, NW China. Alluvial deposits had been found in the Yurunkash River and a larger source lithology was logically located upstream. The project used Landsat TM data, previous field observations and hardcopy topographic and geological maps to identify the source rock for the minerals. The work concentrated on locating large source localities further back in the mountains.

The comparatively coarse pixel size of TM data (28.5 m) meant small-scale lithological changes or detrital jade localities were unresolvable. Also, there was not enough information to determine exactly where jade would occur, only enough to locate lithologies that had similar metamorphic histories. Only knowing the general metamorphic units from the maps and the Landsat data revealed hundreds of square kilometres of potential source rocks. This was too great an area to ground truth for jade and additional queries were required in the GIS to find more definite locations. The geological map was digitised by hand into a GIS database for use in querying the relationships between rock units and potential mineral distribution. Landsat data was used in an effort to corroborate the geological map and locate areas of medium grade metamorphic units. The image was enhanced to remove the dominant brightness of the red and blue bands and then processed to make a lithological map. The techniques for removing colour bias and processing images for specific minerals using band ratios is described in more detail in Chapter 1.

The exploration model used in the preliminary stages of this investigation was to define fluvial activity that crossed known jade localities. A vectorised overlay of the rivers was produced and projected onto the digitised geological map as shown in Plate 8. Once data is in a GIS it can be extracted and analysed in many ways. Certain parts of a theme can be extracted based on their relationship with another theme or based on their own attribute data. The Hotien case study required both of these techniques. To find the source of the jade deposits required examination of the area's drainage systems. The rivers in the area were digitised by hand from the satellite data and then imported into the GIS. From the stream data, vector overlays of the drainage patterns could be constructed. From those overlays, streams could be identified that intersect lithologies of potential economic significance. The GIS was then used to locate lithologies most likely to have deposited the jade fragments.

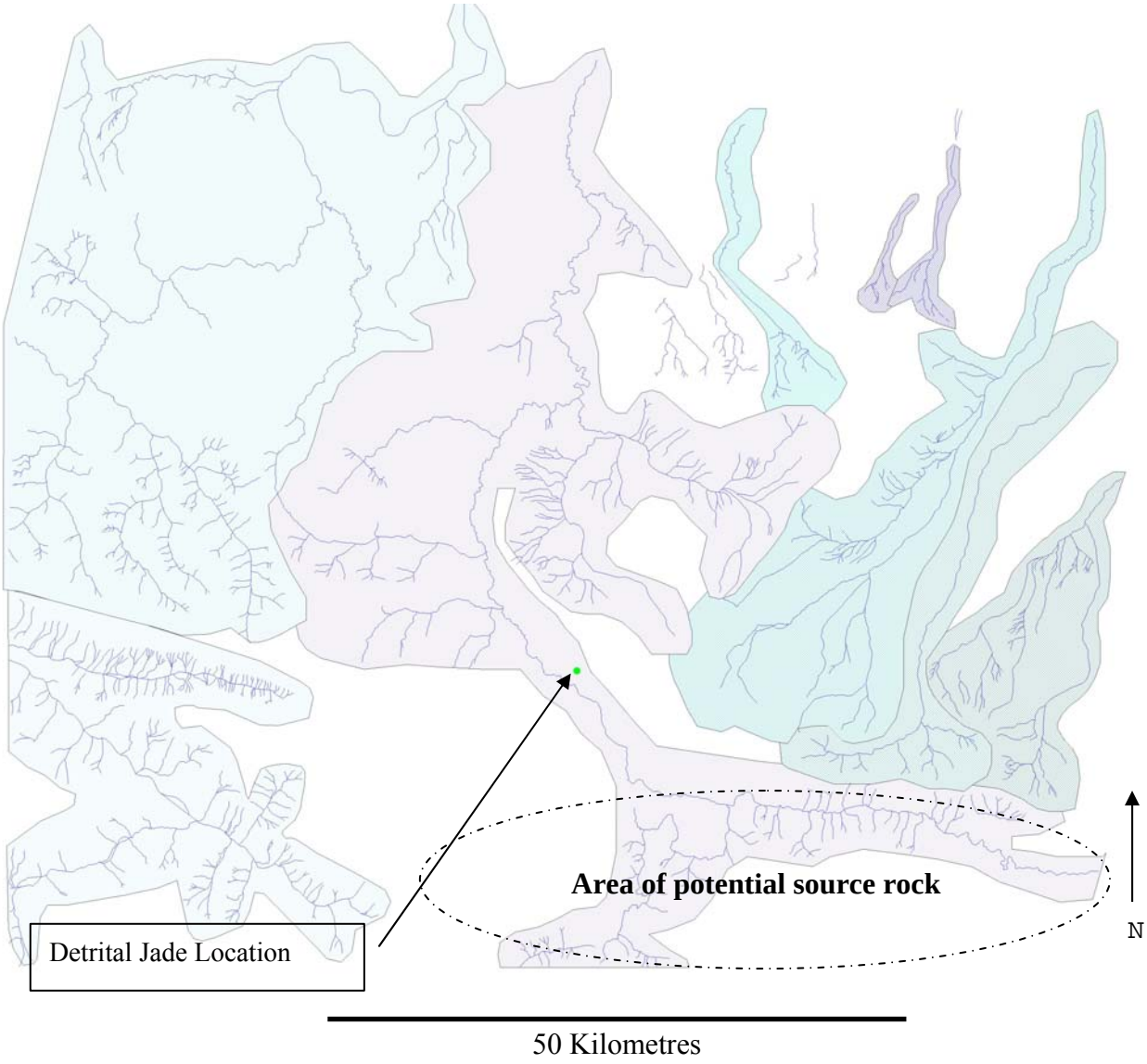
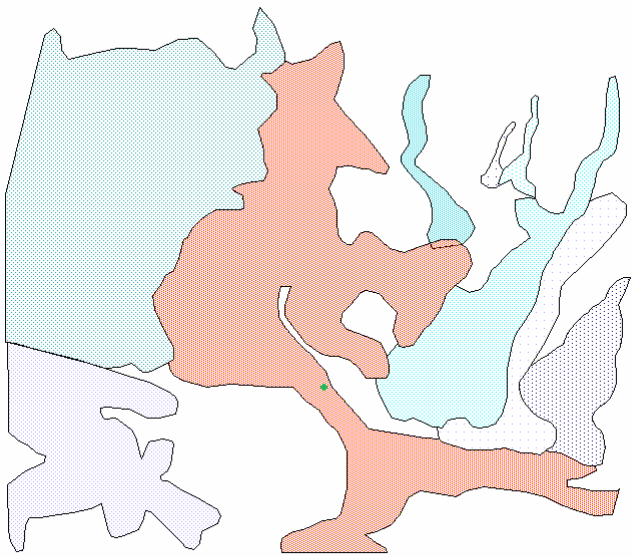
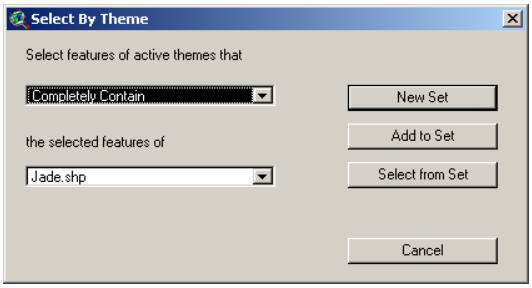


Figure 7-7 Extracted drainage patterns from GIS.

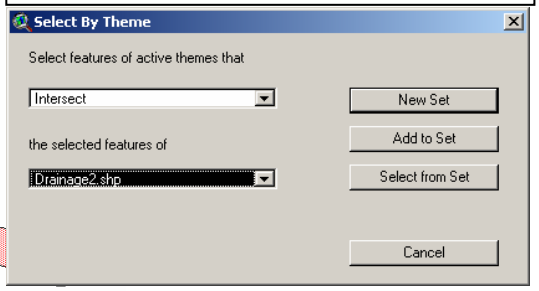
Using the GIS we can merge drainage patterns that intersect the known jade localities and exclude those where no jade was found. We can then highlight other potential units further back in the mountains to reveal larger sources of the semi-precious minerals. This step-by-step procedure is shown in the series of figures over the page.

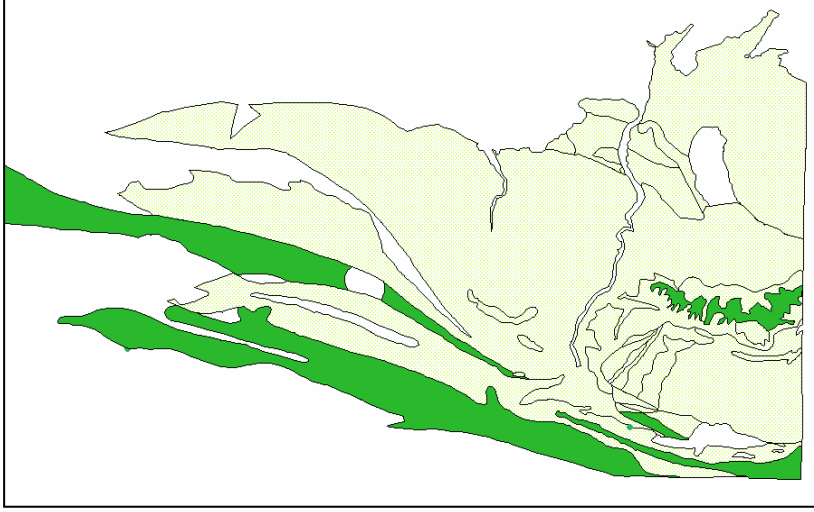


The rivers have been simplified into catchment areas. The relevant catchment areas are selected using a spatial Select by Theme query. This highlights the drainage areas that cut through jade bearing lithologies.

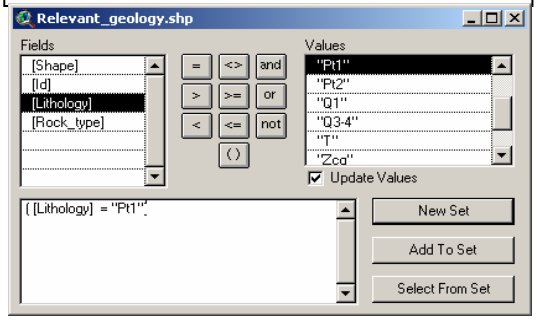


The selected catchment area can be used to filter the lithologies that are underneath it. Only rock units that intersect the river are highlighted.





Each lithology has a set of attributes including a code and other information. Knowing the potential source units of jade (i.e. code 'Pt1' from key in Plate 8) we can write a standard query to extract the relevant source materials.



Using a GIS to quickly interrogate data using spatial queries and the relationship of different themes restricts the need for prolonged work in the field and maximises the chances of getting the best results from a project. In the Hotien project, hundreds of square kilometres of potential jade bearing rocks were reduced to small areas believed to have jade bearing lithologies Plate 9 shows the project conclusions and attempts to localise the jade source rocks.

This sort of analysis can be used in many other disciplines, not just in geological sciences. Selecting polygon layers based on other themes is a common tool in habitat mapping, biogeography and various social/political geography studies. To visualise how these techniques might be used in a habitat mapping exercise, a simplified example is shown below using a similar methodology to the above geological mapping model.

In this example, potential habitats for squirrels are predicted using ASTER commercial satellite data, KVR declassified spy-satellite data and GPS track points. This example is a theoretical study used as a training exercise at the RGS-IBG EAC Mapping Unit training weekends. It is overly simplified with very basic criteria, but the principles can be scaled up to apply to even large scale projects. The example considers parts of Hyde Park in London but the reader should be able to see how these techniques can be extended to large, complex mapping exercises. We can input data into the GIS define certain criteria and get the GIS to output the most likely result. The schematic flow is shown over the page in Figure 7-8.

The data sets supplied for the exercise include high spatial resolution KVR panchromatic data and medium resolution ASTER multispectral data. When conducting any kind of analysis it is important to select the dataset that best suits the task. Both ASTER and KVR have strengths and weaknesses.

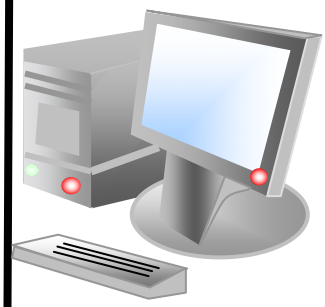
The criteria required for the mapping exercise are to find nesting grounds that are within 500 m of adequate water and within a forest, copse or lightly wooded area of at least 5000 m^2. An additional set of restraints are defined as follows; the habitat must not be within 50 m of a path or bridleway or 75 m of a main road. We can use the imagery to define these areas and then ask the GIS to display the most likely habitat.

Figure 7-8 Schematic Flow Showing Inputs and Outputs from a GISci Habitat Analysis Exercise.

1) INPUTS

Date Type	Resolution		Acquisition Date
KVR	2 m	Visible	1980
ASTER	15 m	NIR	2001

Field Observations & GPS (*current*)



GIS

GPS data downloaded via GPSU into a dbf file and imported into GIS

2) CRITERIA

Must be within:
500 m of adequate water & within a forest, copse or lightly wooded area of at least 5000 m^2.

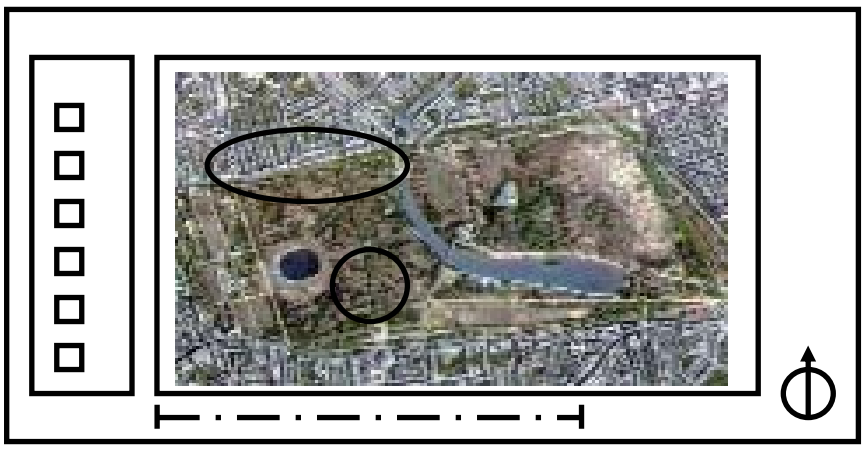
Must not be within:
50 m of a path or bridleway, 75 m of a main road.

3) ANALYSIS

GISci Analysis

- Digitising
- Buffering
- Cutting

4) RESULTS



Final output, presentation quality map with results marked clearly.

To begin with we will define the areas that are good potential habitat areas. Though the supplied KVR data is high resolution and could be used to determine tree locations, its age makes it unsuitable. It is before the storms of 1987 and many other factors may have affected tree growth over a 25-year period. Decisions on the quality and relevance of data to a given task must always be given as much thought as possible. In this example, what seems to be the best data for the task is not immediately suitable. The NIR data from ASTER is good at delineating areas of dense, healthy tree growth. These areas have been hand digitised using ArcGIS and can be seen below in Figure 7-9.

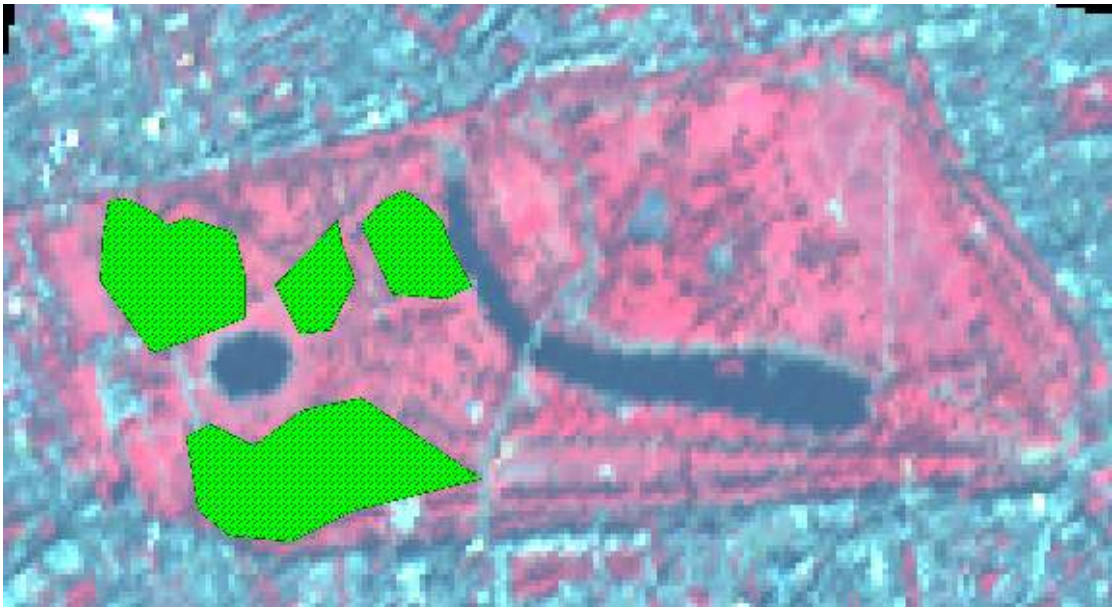
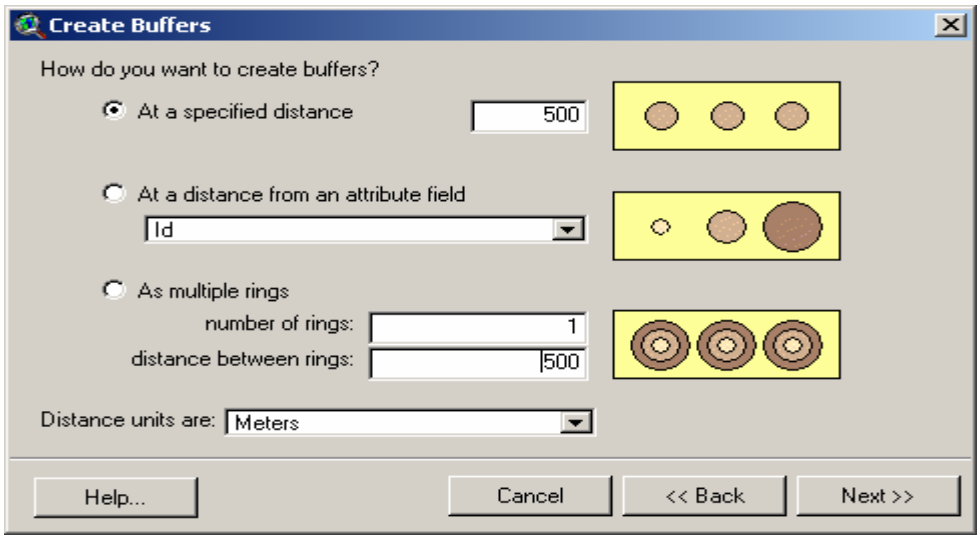


Figure 7-9 Areas of dense copses of trees in Hyde Park.

The squirrels are given an arbitrary requirement of needing a nesting site within 500 m of water. Considering just the areas in Kensington Gardens, the most notable water supply is the round pond. The pond is digitised from the KVR data because it is less likely a pond will have changed in 20 years and the higher spatial resolution of KVR data means the digitising will be more accurate. After digitising the pond we can project a buffer around it using ArcGIS's buffer wizard. This works in the same way as the GPS points were buffered around the glacial snout in the diagram above.



The process of buffering creates a ring around the outside of the pond. When buffering, the final step of the wizard asks whether to include areas within the original shapefile. In this case, this area must be excluded as the squirrel habitat can not be within the water. However, if this was a boundary around a forest or wildlife reserve then the area within the shapefile would be a valid habitat. In these cases the area would need to be included.

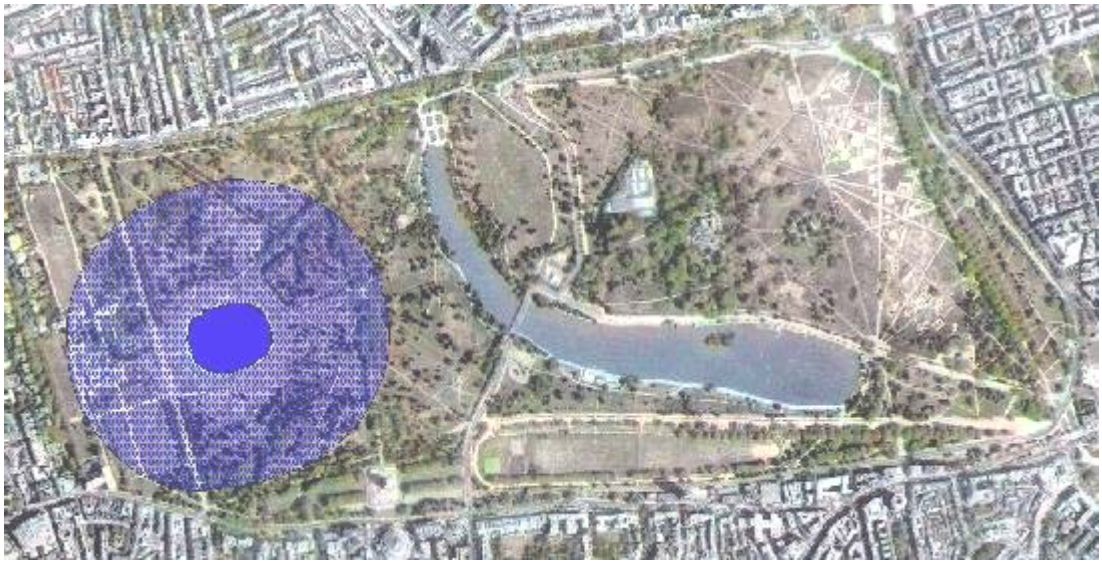


Figure 7-10 500 m buffer around round pond (excluding areas within the pond).

The pond buffer can then be used to 'cookie-cut' the forest polygons. In the Hotien lithology example in Plate 9 the whole lithological unit was selected if it intersected with a river. This was because any part of that lithology might contain jade and knowing the full extent of the jade bearing rocks would be critical to its extraction. In the habitat example we do not just want the forests that intersect the pond buffer, we specifically need the parts of that forest that are within 500 m of the water. This cuts off the outer extent of the forests and eliminates one polygon. The result of this query can be seen overlaid on the ASTER data in Figure 7-11.

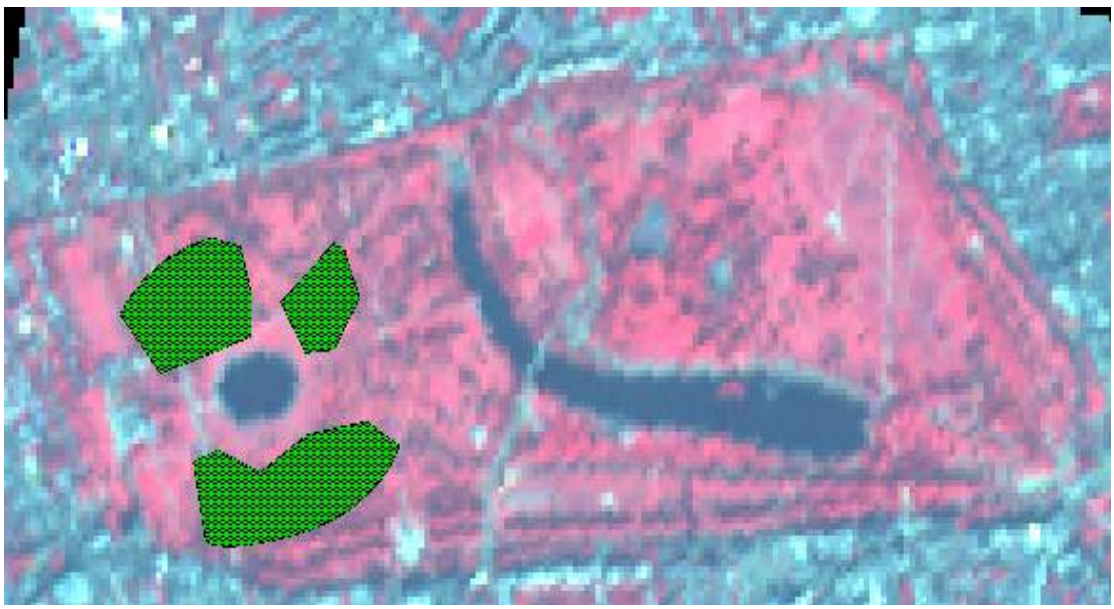


Figure 7-11 Selected relevant areas of forest.

Finally we want to eliminate any areas along footpaths where the squirrels might be disturbed. To do this we take point data from the GPS and buffer this by 50 m. This creates another polygon overlay that can cut the remaining three forest areas.

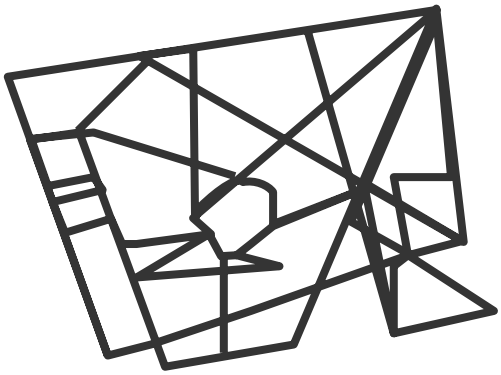


Figure 7-12 Shapefile from GPS tracklogs shown without background raster data.

We can also use GPS waypoints to map any tracks that don't appear in the raster data. We can bring those in as points and buffer them. This combination of good sites overlain with exclusion zones is shown in Figure 7-13. The GIS can union these features together creating one shapefile of all the exclusion zones. We can then 'cookie-cut' the forested zones to leave only the acceptable habitats.

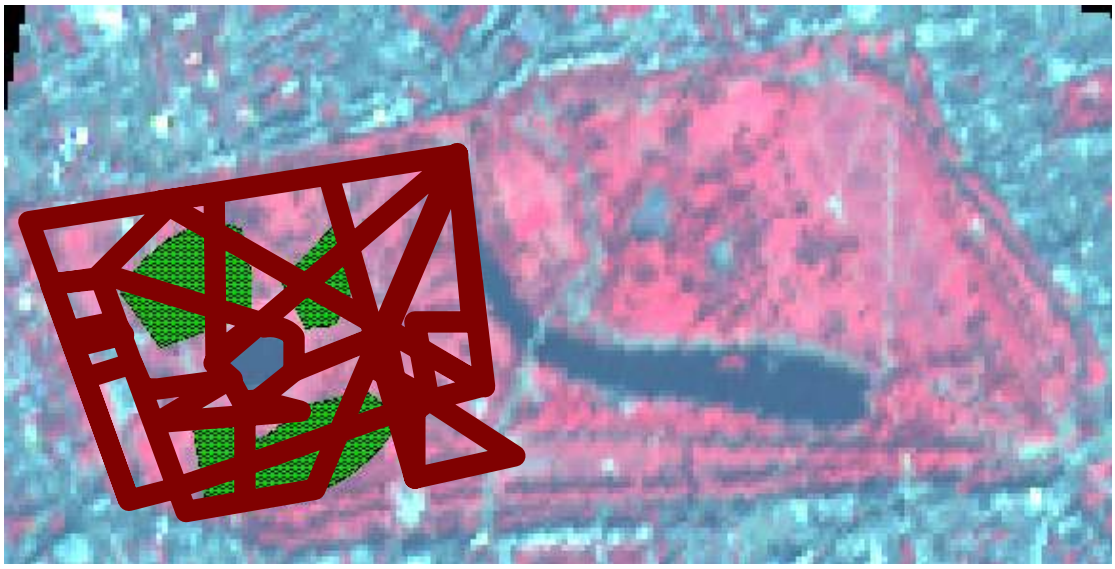


Figure 7-13 Potential habitat areas with exclusion zones.

There are approximately 10 potential sites. However, we might want to restrict our potential habitats to those above a certain size. For example, we might decide that the area must be greater than 5000 m^2. In ArcView a simple query can select the relevant polygons. Because our units are already defined in the GIS we can simply use:

[Area] > 5000.

The result is shown in Figure 7-14 as a yellow polygon.

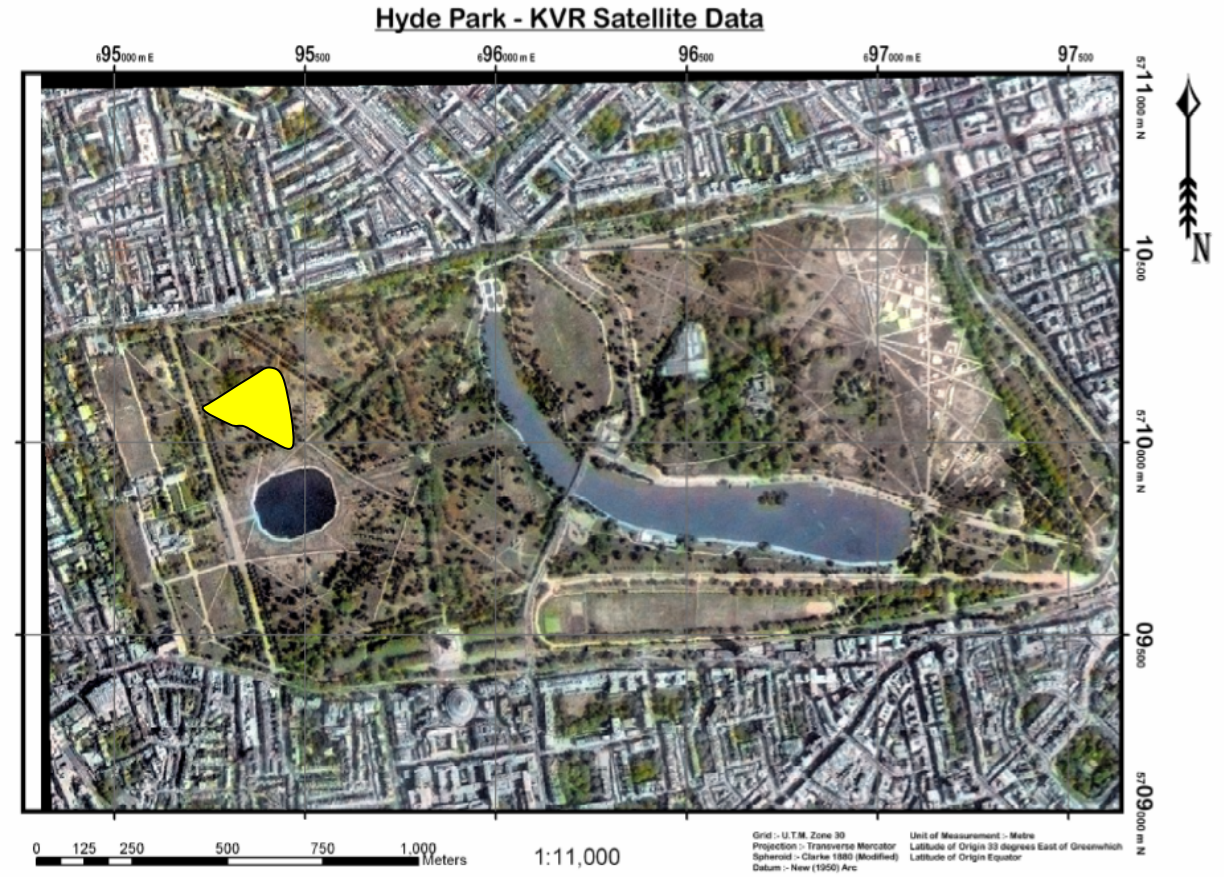


Figure 7-14 Most likely nesting site based on specified criteria.

This kind of analysis can be applied to many study types. The expedition might look at how many people are dependant on a source of water such as a well. It would be simple to query houses within a set distance of the well. If each of those houses had an attribute for number of people living there, it would be simple to total the number of people using the resource.

7.8 Digital elevation data

All GISci data requires two dimensions (X,Y) to be plotted in the GIS. Logically this data will require additional attribute information about the point. This information could be considered a third dimension. The most typical third dimension and the easiest to visualise is altitude. If a series of GPS points were available their height data could be plotted on a Z-axis. This can be seen in below in Figure 7-15.

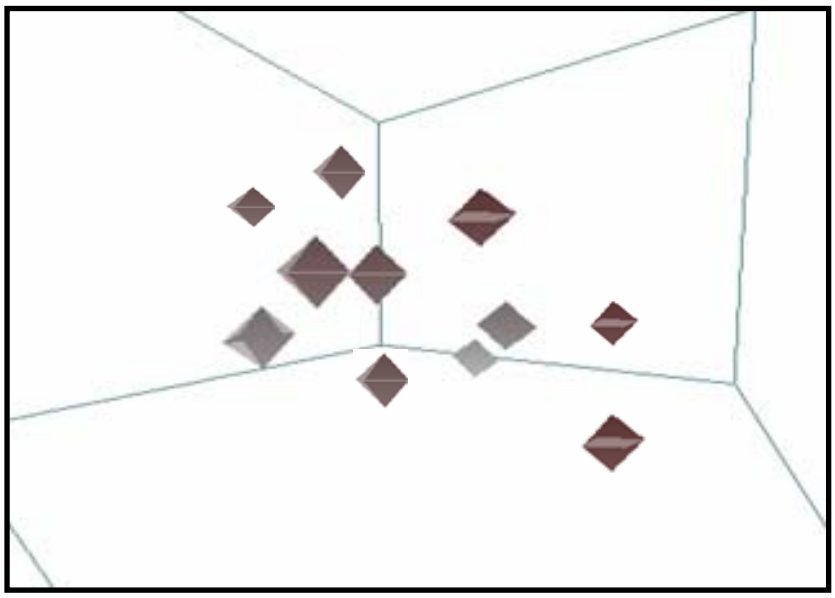


Figure 7-15 Individual GPS points plotted in 3D space.

There are three ways of inputting height data into the GIS. Height data can be random points as seen in Figure 7-15, or a regular grid with a height value at every intersection or a raster surface where the height is expressed as a DN Value. A raster image is essentially the same as a regular grid but expressed as pixels not discrete values. For the GIS to use it effectively the data must be extracted and converted to a grid.

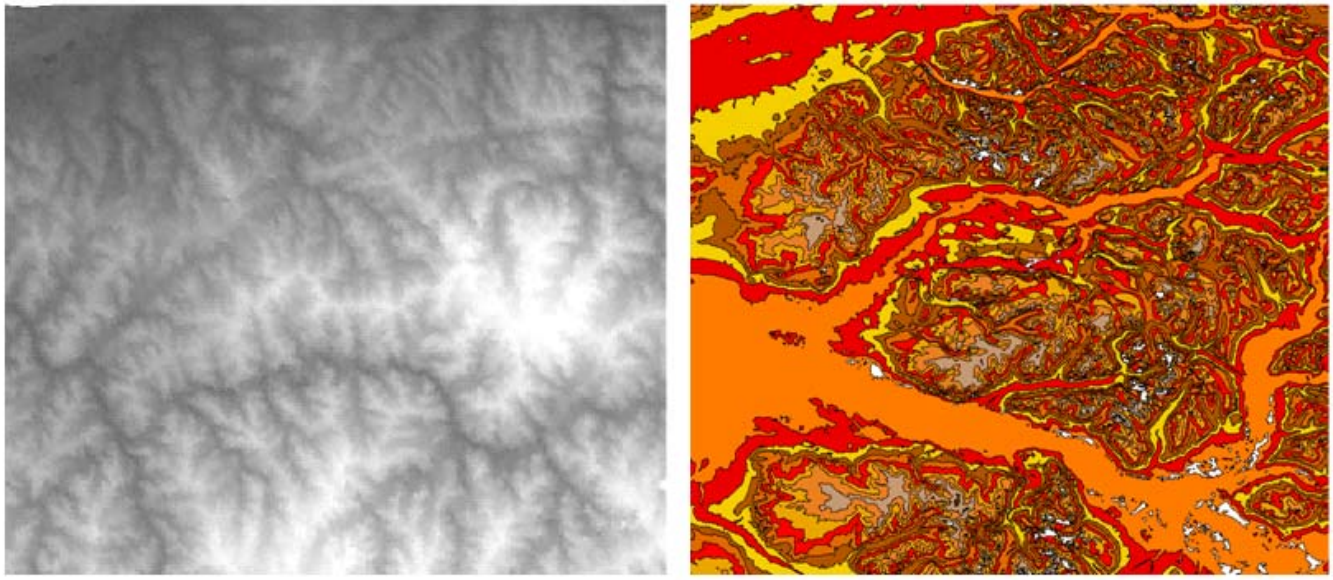
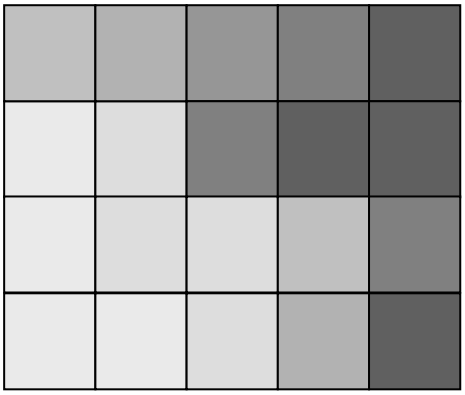
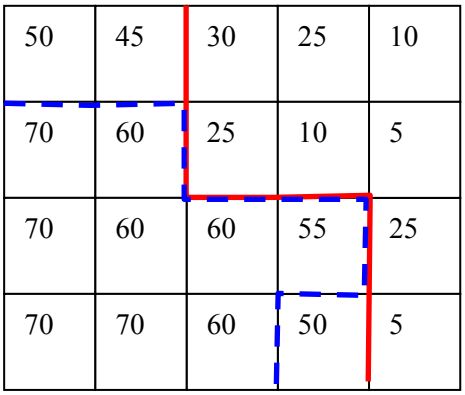


Figure 7-16 Examples of Raster Data (left) and Grid Data (right) (actual areas do not correspond).

When expressed as individual points either at random or in a regular grid some GIS analysis can be conducted. However, no matter how densely the points are plotted they are still individual points and the GIS needs to know more about their relationship. For example, one typical use of altitude data is to generate contours for a map. The weaknesses of individual points or raster data are shown below in Figure 7-17.



50	45	30	25	10
70	60	25	10	5
70	60	60	55	25
70	70	60	50	5

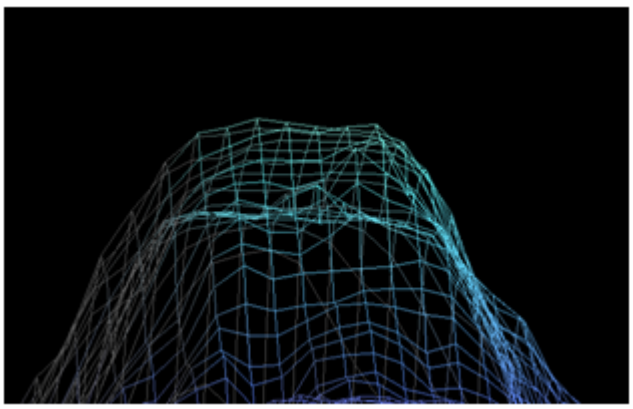


50	45	30	25	10
70	60	25	10	5
70	60	60	55	25
70	70	60	50	5

If we created a contour based on a regular grid of data the limitations become apparent. The solid line maps areas at 30 m and the dashed line maps areas at 50 m. It is clear there is a problem because for a considerable length of the middle section the lines lie on top of each other.

Figure 7-17 A raster is converted to a grid of points but the GIS can not understand where unknown / un-sampled points would lie.

The reason the contouring failed was because of the limitation of the data. The GIS was unable to calculate what happened in between the points. Cleary, in between two points such as 55 m and 25 m the topography must pass through 30 m and it is probable that the 30 m line would be closer to the 25 m point than the 55 m. The individual points can not tell us where that would be. The GIS is capable of turning these discrete points into a surface by interpolation. This process allows the GIS to predict the values of un-sampled points based on the surrounding points. The GIS interpolates the points into a series of triangles with the diagonals of the triangle used to calculate points that are not specifically mapped. These triangles are often referred to as a Triangular Irregular Network (TIN). A TIN allows the contours to be plotted more realistically as shown in Figure 7-18.



50	45	30	25	10
70	60	25	10	5
70	60	60	55	25
70	70	60	50	5

Figure 7-18 Grid data converted to a TIN and contours generated. Each node in the network is a point of value 'z' and from those nodes lines can be drawn to predict values at unknown points.

In the right hand image of Figure 7-18 the GIS is using a TIN to predict where the un-sampled 30 m and 50 m lines should be. The transition of raster data to a TIN allows for a lot more analysis. The TIN allows the user to see and model aspects of the geography. Areas at risk from flooding can easily be viewed by putting in a flood depth and seeing which areas would be above or below the water. It is also important to remember that the TIN need not be referring to actual elevation data. It could easily be expressing variables such as temperature or rainfall.

The example in Plate 10 shows a real life use of this data. By using stereo imagery first in an image processing suite and then in a GIS application we can see how the data can be used to calculate changes in volume. The example uses data from the ER Mapper 6.3 examples tutorial. Even though image processing software allows a view of the data it does not allow an analysis of the data. For this we need to transfer the data into a GIS such as ArcView. Unfortunately, when raster data comes into the GIS the software does not necessarily know any details about the image. It simply represents the z-axis as unknown DN values. These DN values could be height, rainfall or just standard reflectance returns. First of all the GIS needs to be told that the data should represent individual points. This creates a grid as shown below in Figure 7-19.

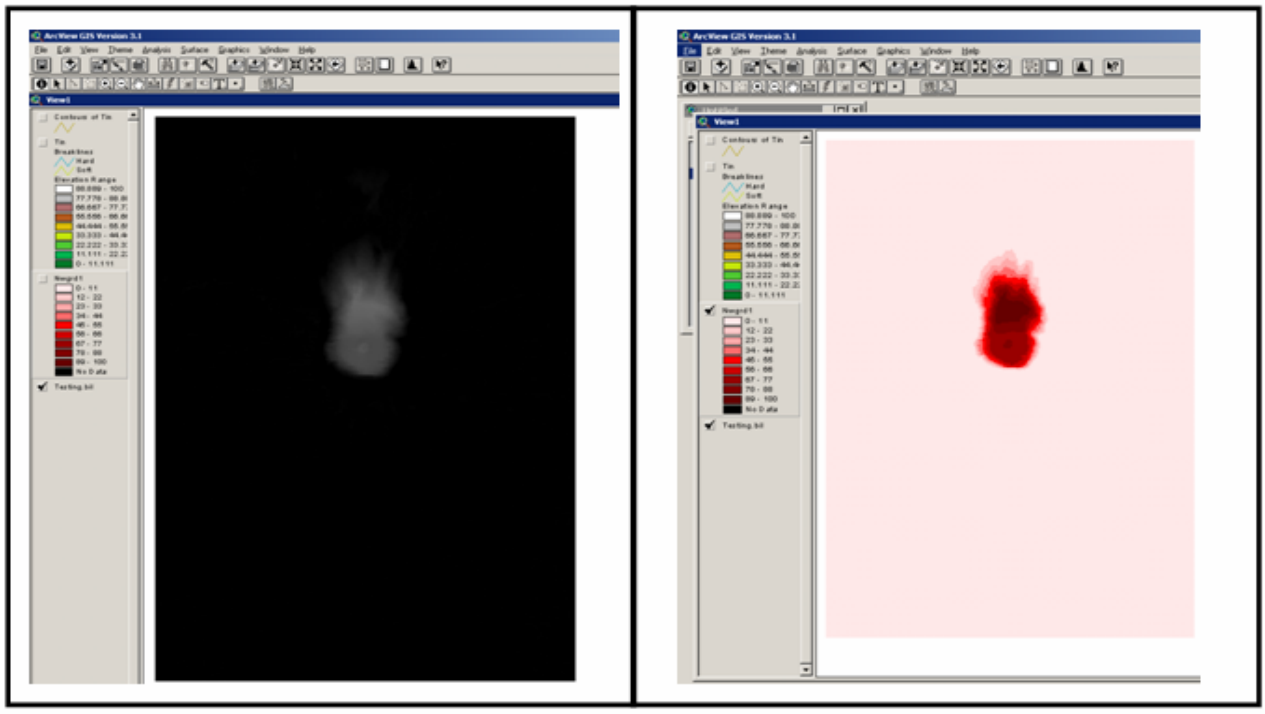


Figure 7-19 The raster image shown of the left is converted to a regular grid of points as shown on the right. This grid is then shaded at known intervals. This can then be used to generate a TIN for GIS analysis.

In the left-hand image the GIS is simply displaying a raster image of unknown DN values. By gridding the data we tell the GIS the data has intrinsic values that are important. A TIN can be built to connect the points and create a surface the GIS can query. We can ask a GIS to analyse a TIN in many ways. One interesting process in this example would be to find the volume of rock destroyed in the eruption. We can do that by using the area and volume statistics options from ArcGIS surface analysis tools. This generates the following dialogue box shown in Figure 7-20.

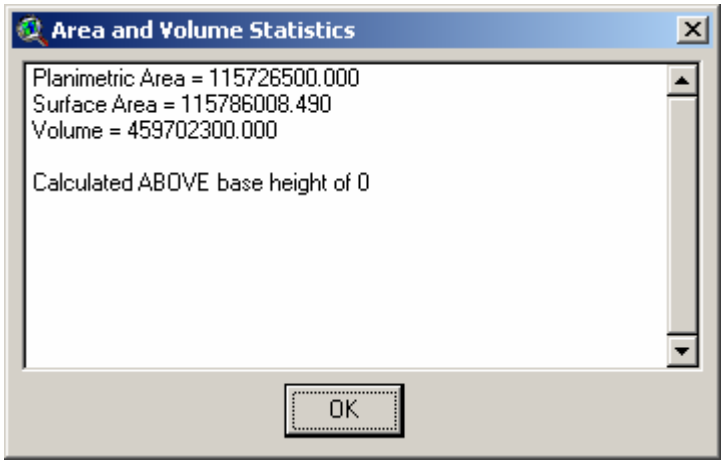


Figure 7-20 ArcView report based on Mt St Helens TIN.

The result of our analysis is 459,702,300 cubic metres of material or 0.46 cubic kilometres of rock. This shows the use of this type of data and hopefully the reader can see how these processes can be applied to many other applications.

Another similar application is measuring the volume of ice lost from a glacier. This can be conducted from a series of GPS points over time but is more easily done using stereo imagery. The glacial heights can be subtracted in a similar manner to the volcanic dome to give a volume of ice lost.

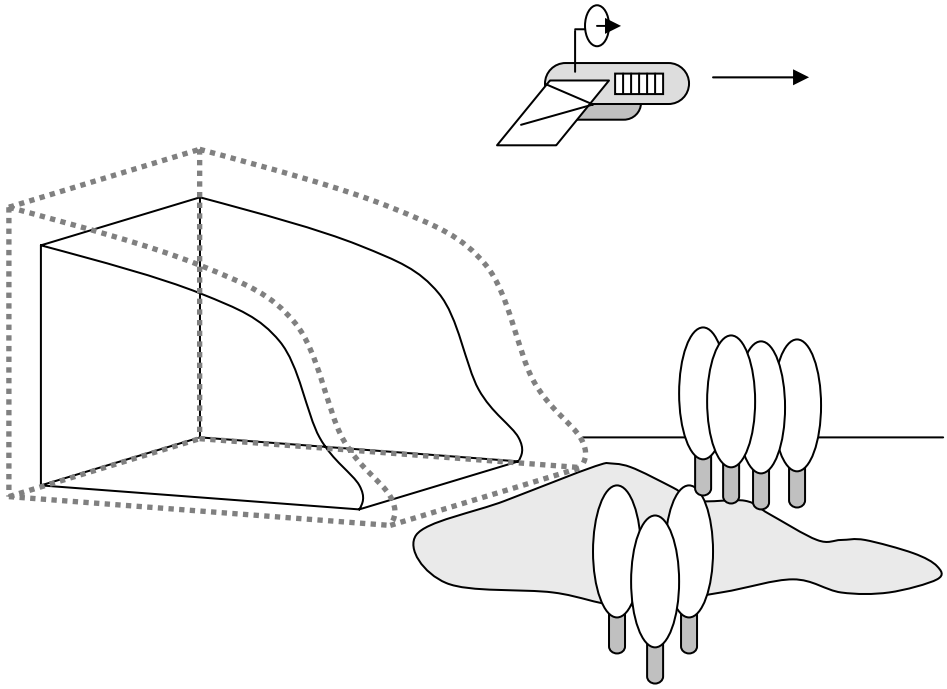


Figure 7-21 Schematic diagram showing glacial retreat in 3D.

The dotted snout in Figure 7-21 represents the level of the ice at some point in the past. The inner solid line represents the position of the ice today. The GIS can take points across the glacier, grid them and then calculate a TIN from points.

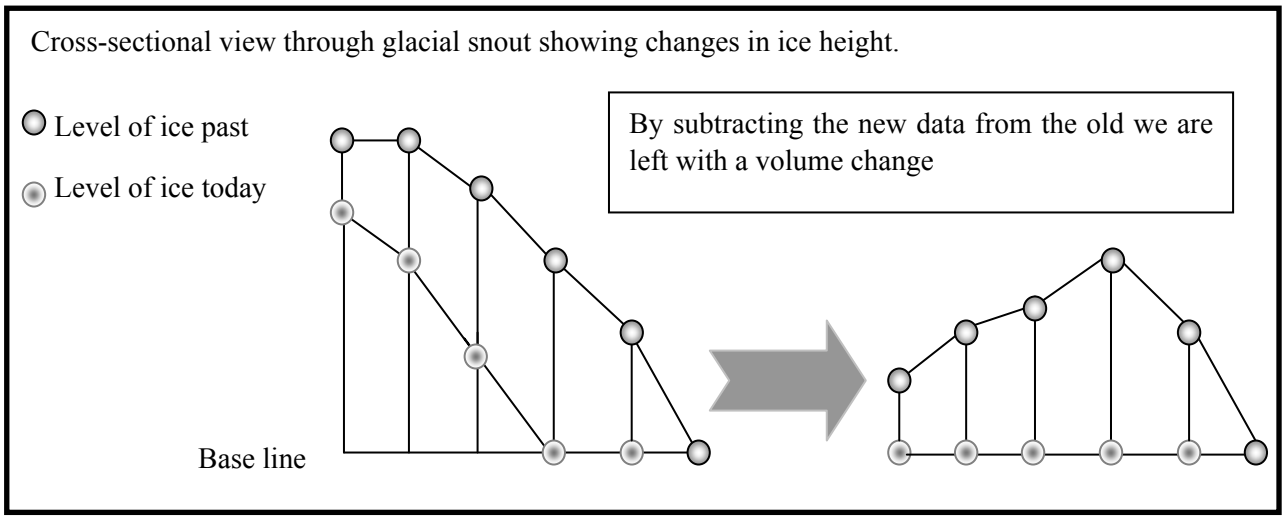


Figure 7-22 Subtracting new data from old calculates the volume change.

Some of the oldest available stereo satellite data is from the US Corona series (known as KH for Key Hole). High resolution KH4 data has a resolution better than 6 foot. This level of accuracy from the early sixties gives a temporal resolution of over 40 years and combined with modern GPS data or equally high quality stereo imagery such as IKONOS gives excellent results. Any stereo data can be used but using either ASTER 15 m pixels or SRTM 90 m data will degrade the results when compared to KH4 or aerial photography. An example of glacier data used in a study of this type is shown below in Figure 7-23.

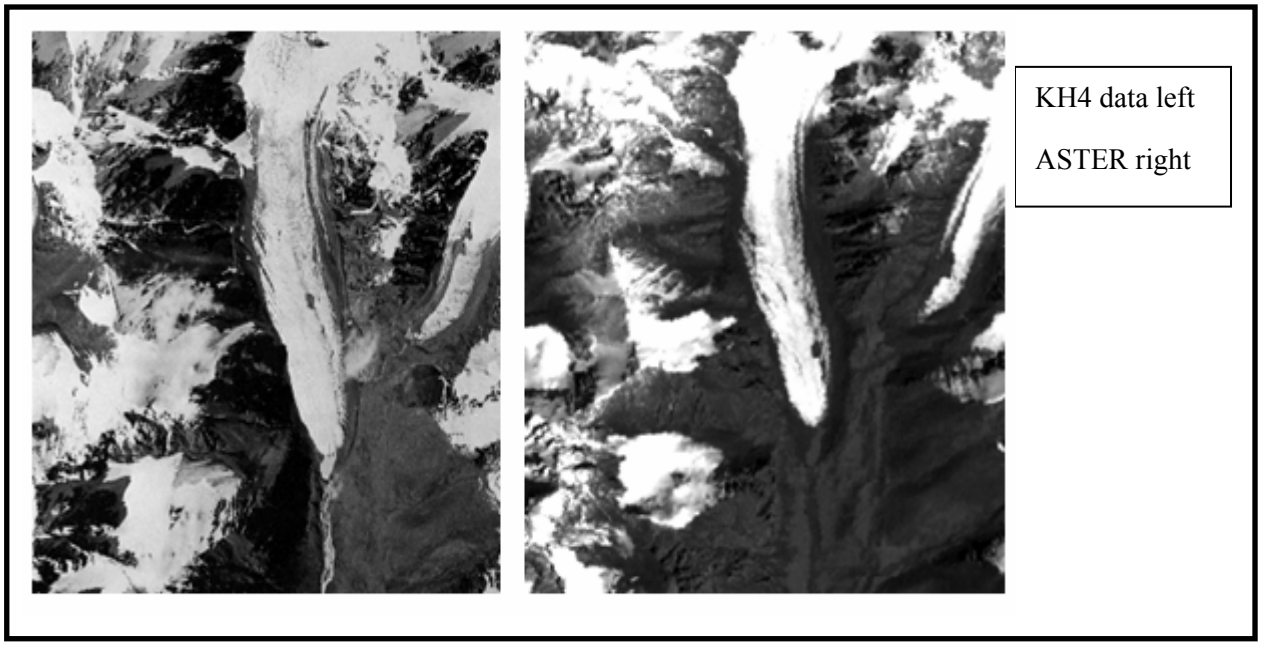


Figure 7-23 The imagery above clearly shows a change in the snout of the imagery. Example from Whiteside et al 2001 (b).

Using the highest possible resolution will always be beneficial to the expedition but the team should note that even theoretically accurate data like KH4 suffers from errors when rectified. Rectifying hardcopy data introduces considerable errors and the team should not count on getting results as accurate as modern IKONOS data from older hardcopy information.