CHAPTER
16

INTRODUCTION TO GEOSTATISTICS

Sreenath K. R.

Marine Biodiversity Division
ICAR-Central Marine Fisheries Research Institute

Geographical Information System (GIS) is a technological tool used to describe and characterize spatially referenced geographical information for the purpose of visualizing, querying and analyzing. The tool enables capturing, storing, analyzing, sharing, displaying and modelling of spatial data maintained with in single database. Making decision based on geography is basics to human thinking and spatial analysis using GIS enable people to combine information from many independent sources and derive entirely new layers of information that are more accurate and reliable in decision making. Spatial analysis involves study of phenomenon that varies with time and space. Geostatistics is a branch of statistics used for analysis of spatial or spatiotemporal data set by applying sophisticated set of various statistical and probabilistic models. Here, we estimate the value of phenomenon from unknown location where no measurements are available with the help of direct measurements derived from known locations. GIS has emerged as powerful tool in the recent years by providing various geospatial solutions in urban planning, mining, natural resource evaluation and management, pollution estimation, risk assessment and large scale mapping thus becoming integral tool in our day today life.

1. Data for GIS

Data for GIS are obtained from various sources like aerial & satellite imageries, digital data, conventional maps, census, meteorological department, field data (surveys/GPS) *etc.* The information obtained can be classified into types of database: spatial data which describe the location (where the object is?) and attribute data which characterize the location (what the object is? or how much the object is?)

1.1. Spatial Database

Spatial data is representation of complex real world in a simplified manner. Here the geographical features are represented by three basic types- points, lines and area. *Points* represent dimensionless features such as wells, post box, tube well *etc.* that are very small and their location can be explained by only coordinate's values. Lines depict features with length, such as roads, railways stations, administrative and international boundaries *etc.* and are two-dimensional. Area or polygons are used to represent three-dimensional objects that have height, width and length such as agriculture lands, water bodies, forest areas and administrative areas.

1.2. Attribute database

Attribute data depicts various characteristics of different object/ features on earth surface. This can of qualitative types like land use type, soil type, name of the city/ river *etc.*) or of quantitative types (like elevation, temperature, pressure of a particular place, crop yield per acre *etc.*). Thus, attribute data can be both numeric and textual.

2. Representation of Database

The way that location is represented in a geodatabase can be either a raster or a vector position.

2.1. Raster data

A raster based format uses imaginary grid of cells or matrix to displays, locates and stores graphical data. The fundamental unit of raster system is pixel. Here, the whole study area is divided into uniform rows and columns and each cell or pixel is used for storing point, line or area entities. Here, points are represented by individual column/ row entities, lines are depicted by connecting the adjacent cells or pixels and areas are stored as set of contiguous cells defining the interior. The accuracy of raster data formats depends on pixel or grid size and may vary from submeter to kilometres. Layers are functionally related map features that are used to represent different two-dimensional features on map. Different layers are used to in GIS for storing various unique information such as forest cover, soil types, land use pattern and wetlands. Satellite images, Digital terrain models (DTM) and digital elevation models (DEM) are examples of raster data (Koeln *et al.,* 1994 and Huxhold 1991). Raster data formats require less processing over vector formats but they consumes more computer space for storing of data.

2.2. Vector data

In vector maps, world is represented by points, lines and polygons. The fundamental unit of vector system is point. Lines are set of mathematically connected points and area are represented by set of mathematically connected coordinates or lines joined together to form polygons which define the boundary of area. Unlike raster images, vector images can be of high resolution. Vector data requires less computer storage space and maintaining topological relationships is easier in this system (Koeln *et al.,* 1994; and Huxhold 1991).

3. Projections

Once the spatial data have been collected, the data needs to be in same coordinate system for display and analysis. As earth surface is ellipsoidal therefore set of set of systematic mathematical transformation is needed to display earth's latitude and longitude onto a plane.

Projection is a method by which curved surface of the earth is portrayed on a flat surface. Initially the earth was thought to be flat surface but later on it was proven that earth is an ellipsoidal/spheroid, the circumference of the earth is about 1/300th smaller around the poles *vs* equator. This difference in distance around the poles and equator use to cause error in the readings and to rectify the errors different projection systems were created. These are just different measurements of the "flattening" at the poles. The different projection systems are helpful in measuring and preserving one or more properties such as area, shape, direction or distance over commonly used latitude longitude (x, y, which measures in degree and not in distance) coordinate system.

3.1 Different types of projections

Azimuthal or planner projection: Projection surface laid flat against the earth.

Conic- Cone is placed on or through the surface of earth.

Cylindrical- projection surface wrapped around the earth.

Coordinate system: A reference framework consisting of set of points that are used to define its position in space either in two or three dimensions.

Cartesian Coordinate system: Two dimensional, planner coordinates system in which the horizontal distance is measures along the x axis and vertical distance is measures along the y axis. Each point ids are defined by x, y coordinate.

Datum: Set of coordinates that measures the position on a surface using x,y coordinates (horizontal) and height above or below the surface (vertical datum).

Geocentric datum: A horizontal geodetic datum based on a ellipsoidal that has its origin at the earth centre's mass and measures coordinate of every point on Earth using latitude longitude and height above its surface. Ex. World Geodetic system of 1984. (WSG84).

3.2 Common GIS projections

***Mercator* :** It is cylindrical projection tangent to the equator of earth. Preserves the local shapes and display accurate compass bearing for sea travel.

Transverse Mercator: It is also a type of cylindrical projection similar to Mercator except the cylinder is tangent along a meridian instead of the equator. It minimizes the distortion along north-south line, but does not maintain true direction.

Universal Transverse Mercator (UTM): *UTM is* based on transverse Mercator projection and divides the whole world in 60 north south zones, each zone having a width of 6^{o} longitude. Each zone is numbered consecutively beginning with zone one covering longitude 180^{o} to 174^{o} West and progressing east word to zone 60, between 174^{o} to 180^{o} East longitude.

Lambert Conformal Conic – A conic, conformal projection typically intersecting parallels of latitude, standard parallels, in the northern hemisphere. This projection is one of the best for middle latitudes because distortion is lowest in the band between the standard parallels. It portrays shape more accurately than area.

Most commonly used projection in GIS is UTM or the preference may change depending on the area of interest.

4. Interpolation

Spatial data is important in making important decisions in natural resource management. Collection of spatially continuous data is often difficult and expensive. Most of the data collected by field surveys will be typically from point sources. But scientists and managers requires accurate spatial continuous data to make justified interpretations.

Spatial continuous data of environmental variables are in demand in the geographic information systems (GIS) and modelling techniques for studying the ecology and biological conservation (Collins and Bolstad, 1996; Hartkamp *et al.*, 1999). Thus, spatial interpolation methods have overarching importance in converting point data in to spatially continuous data. Interpolation methods can fall under two categories 1. Global methods and 2. Local methods. Global methods use all available data of the region of interest to derive the estimation and capture the general trend. Local methods operate within a small area around the point being estimated (*i.e.*, use samples within a search window) and capture the local or short-range variation (Burrough and McDonnell, 1998).

4.1 Global interpolators

4.1.1 Regression Models

Regression interpolation is using a linear regression model (LM) as interpolator and assumes that the data are independent of each other, normally distributed and homogeneous in variance. Regression methods explore a possible functional relationship between the primary variable and explanatory variables that are easy to measure (Burrough and McDonnell, 1998). The final model can be selected by a thorough understanding of the relationships between the primary variable and secondary variables and/or by Akaike information criteria (AIC) or Bayesian information criteria (BIC) methods.

4.1.2 Trend surface models

An inexact method, trend surface analysis approximates points with known values with a polynomial equation. This is similar to the regression model but uses only geographic coordinates as indirect variables for prediction of the primary response variable (Collins and Bolstead, 1996).

4.2 Local interpolators

4.2.1 Nearest Neighbours (NN)

The nearest neighbours (NN) method draws perpendicular bisectors between sample points (n), predicting the values at the unsampled regions. The resultant polygons are called as Thiessen or Voronoi polygons. All the area inside each polygon will have same value, which is the value of the midpoint of the polygon.

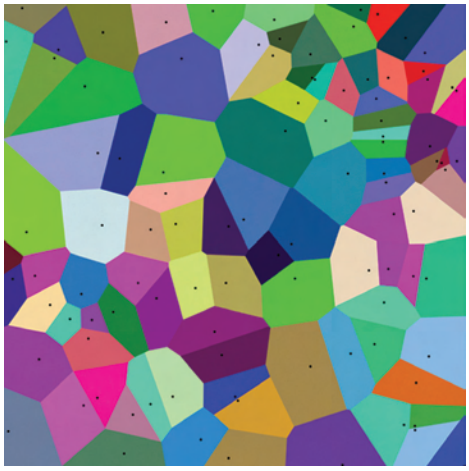

Fig. A Voronoi Polygon map

4.2.2 Triangular Irregular Network

In the triangular irregular network (TIN), all sampled points are joined into a series of triangles based on a Delauney's triangulation. It forms a different basis for making estimates in comparison with those used in NN. The value of the regions falling in a triangle is estimated by linear or cubic polynomial interpolation. Peuker *et al.,* (1978) developed the method for digital elevation modelling (DEM) that avoids repetitions of the altitude matrix in the grid system.

4.2.3 Natural Neighbours

The natural neighbours (NaN) method combines many characters of NN and TIN. The method was developed by Sibson (1981). The first step is a triangulation of the data by Delauney's method, in which the apices of the triangles are the sample points in adjacent Thiessen polygons. This triangulation is unique Spatial Interpolation Methods 7 except where the data are on a regular rectangular grid. To estimate the value of a point, it is inserted into the tessellation and then its value is determined by sample points within its bounding polygons. For each neighbour, the area of the portion of its original polygon that became incorporated in the tile of the new point is calculated. These areas are scaled to sum to 1 and are used as weights for the corresponding samples (Webster and Oliver, 2001).

4.2.4 Inverse Distance Weighting

The inverse distance weighting (IDW) method estimates the values of the unsampled points using a linear combination of values of the sampled points weighted by an inverse function of the distance from the said point to the sampled points. The weight diminishes by an inverse factor and sampled points will have more influence on nearby points. The rate of diminishing value depends on the factor (Isaaks and Srivastava, 1989).

The weights can be expressed as:

$$\lambda_i = \frac{1/d_i^p}{\sum_{i=1}^{n} 1/d_i^p}$$

where d_i is the distance between point of interest x_0 and sampled point x_i, p is a power parameter, and n stands for the number of sampled points used for the estimation. The power parameter is arbitrary decided by the validation at field. Most popular value for p is 2 and then the IDW is called as Inverse square or inverse distance squared (IDS) method.

4.2.5 Splines

This is an inexact, gradual interpolation which uses piecewise polynomial equation as interpolator. The polynomials describe pieces of a line or surface (*i.e.*, they are fitted to a small number of data points exactly) and are fitted together so that they join smoothly (Burrough and McDonnell, 1998; Webster and Oliver, 2001). For degree p = 1, 2, or 3, a spline is called linear, quadratic or cubic respectively. Typically, the splines are of degree 3 and they are cubic splines (Webster and Oliver, 2001).

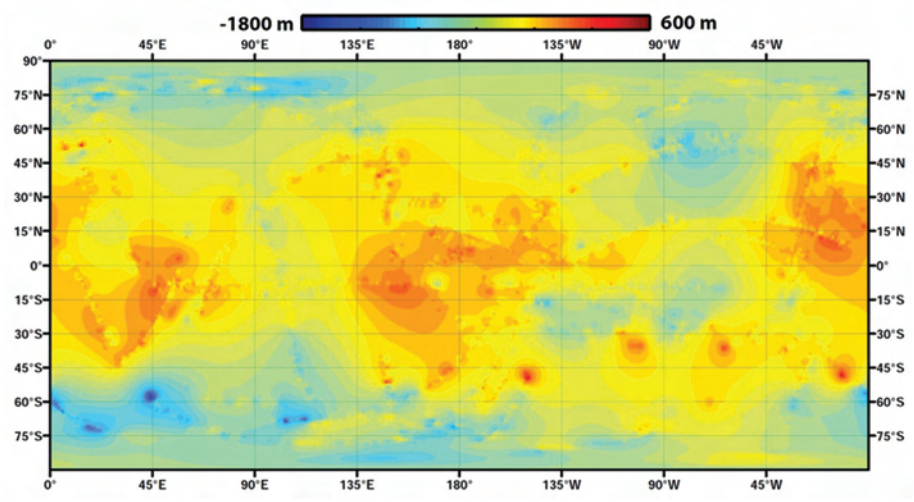

Fig. NASA's Cassini spacecraft gridded elevation data has been splined create the first global topographic map of Saturn's moon Titan [Image credit: NASA/JPL-Caltech/ASI/JHUAPL/Cornell/Weizmann]

4.2.6 Krigging

Basic concept of Geostatistics is that variables of a specific geographic region tend to have a particular structure. Though this particular domain of spatial interpolation has its origin in 1910s in agronomy (Webster and Oliver, 2001), this is mostly developed in the works of geology and mining by Krige (1951). Geostatistics includes several methods that

use kriging algorithms for estimating continuous attributes. Kriging is a generic name for a family of generalised least-squares regression algorithms, used in recognition of the pioneering work of Danie Krige (1951). Li and Heap (2008) gives a good review of all the available interpolation methods. In Krigging interpolation is performed by modelling a Gaussian process which considers method of interpolation for which the interpolated values are modeled by a Gaussian process governed by prior assumptions and gives the best unbiased estimate of the unsampled values.