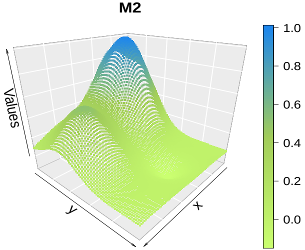
Statistical methods for spatial data analysis

Lecture 2: Introduction to spatial data (1)

Joaquin Cavieres

Geoinformatics, Bayreuth University



Outline

1. Introduction

Are you agree with this sentence?

Spatial data are everywhere!

Are you agree with this sentence?

Spatial data are everywhere!

- Television
- Newspapers
- Route planners
- Computer screens
- Plain paper maps
- etc..

Could spatial data help us to answer..

Could spatial data help us to answer..

- Could spatial data give us patterns of a disease incidence? could they are grouped by clusters? are the clusters found related to factors such as age, relative poverty, or pollution sources?

Could spatial data help us to answer..

- Could spatial data give us patterns of a disease incidence? could they are grouped by clusters? are the clusters found related to factors such as age, relative poverty, or pollution sources?
- Considering some soil samples, which part of a study area is polluted?

Could spatial data help us to answer..

- Could spatial data give us patterns of a disease incidence? could they are grouped by clusters? are the clusters found related to factors such as age, relative poverty, or pollution sources?
- Considering some soil samples, which part of a study area is polluted?
- Given scattered air quality measurements, how many people are exposed to high levels of air pollution where do they live?

The main idea behind of to use spatial data is:

1. Understand the problems that we confront.
2. Propose a solution for them.

For the above we will cover the theoretical background of methods and models for data analysis, emphasizing the use of R for this purpose.

For the above we will cover the theoretical background of methods and models for data analysis, emphasizing the use of R for this purpose.

Why do we use R?

- R is a multiplatform, open source language and environment for statistical computing and graphics
- With a wide range of packages, R also supports advanced geospatial statistics, modeling and visualization
- R is flexible to create interactive maps

2. Types of spatial data

Spatial data have spatial reference: they have coordinate values and a system of reference for these coordinates.

For example, consider the locations of volcano peaks on the Earth. Specifically, we are interested in the volcanoes that have shown activity between 1980 and 2000, according to some agreed seismic registration system (only locations).

If we want to create a map (flat) of those locations, we are faced with the problem of projection: we have to translate from the spherical longitude/latitude system to a new, non-spherical coordinate system.

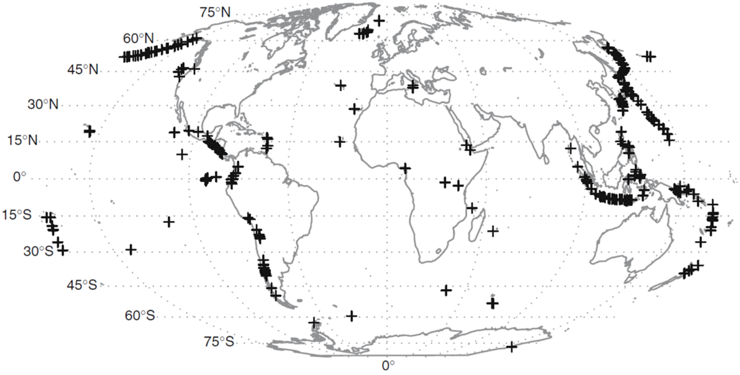


Figure 1: Volcanoes of the world, with last known eruption 1964 or later (+). Source [1]

If we also have the date and time of the last observed eruption at the volcano, this information is called an *attribute*: it is non-spatial in itself, but this attribute information is believed to exist for each spatial entity (volcano). However, in the previous figure we represent the purely spatial information (locations).

We also can distinguish between different data models:

- *Points*: A single point location, such as a GPS reading or a geocoded address
- *Line*: A set of ordered points, connected by straight line segments
- *Polygon*: An area, marked by one or more enclosing lines, possibly containing holes
- *Grid*: A collection of points or rectangular cells, organised in a regular lattice

Points, lines and polygons are vector data models, while a grid is a raster data model.

Locations (x, y coordinates) is sufficient to establish its position relative to other volcanoes on the Earth, but for describing a single volcano we can use more information, for example, its topography.

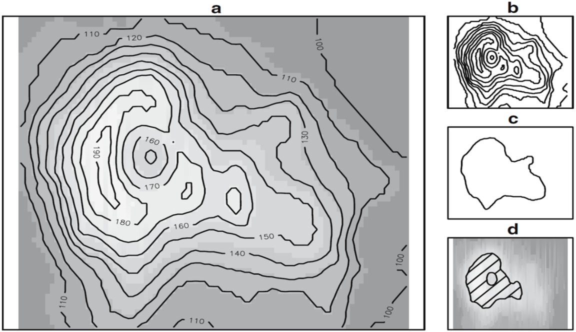


Figure 2: a) Topographic information (altitude, m) for Maunga Whau Volcano on a 10×10m2 grid, (b) contour lines, (c) 140m contour line: a closed polygon, (d) area above 160m (hashed): a polygon with a hole. Source [1]

We can also represent spatial data in a rectangular grid: the values in each grid cell may represent an average over the area of the cell or the value at the midpoint of the cell. This representation is important because:

- Cameras and remote sensing instruments register data on a regular grid
- Computer screens and projectors show data on a grid
- Many spatial or spatio-temporal models, such as climate models, discretise space by using a regular grid.

3. Attributes and Support

Attributes refer to the properties of features ("things") that do not describe the feature's geometry. They can be derived geometry but also can refer to non derived properties, as for example:

- The name of a street or a county
- People living in a country
- Type of a road
- The soil type in a polygon from a soil map
- The opening hours of a shop
- NO_2 concentration measured at an air quality monitoring station

When we have time properties, e.g. measures of NO_2 by hours, the time is handled with the geometry (e.g. in a data cube).

5.1. Attribute-geometry relationships and support

Changing the feature geometry without changing the feature attributes does change the feature, since the feature is characterised by the combination of geometry and attributes. However, is necessary to give an explanation of attribute-geometry relationship (AGR):

1. Constant: the attribute value is valid everywhere in or over the geometry
2. Aggregate: the attribute is an aggregate, a summary value over the geometry

For polygon data, typical examples of constant AGR (point support) variables are:

- Land use for a land use polygon
- Rock units or geologic strata in a geological map
- Soil type in a soil map
- Climate zone in a climate zone map

The geometry of those variables follows from mapping the variable observed.

Examples for the aggregate AGR (block support) variables are:

- Number of persons or as population density
- Socio-economic variables, summarised by area
- Total emission of pollutants by region
- Average reflectance over a remote sensing pixel

The properties of those variables come from legislation, devices or statistical analysis, not from the variable observed directly.

A third type of AGR is called **identity**. We call an attribute an identity variable when the associated geometry uniquely identifies the variable's value, for example a county name, however, an arbitrary point (or region) inside a county is still part of the county and must have the same value for county name, but it does not longer identify the (entire) geometry corresponding to that county

The spatial information that we can get from identities are:

- Fields: every location corresponds to a single value in a continuous space
- Objects: can represents a set of discrete locations, for example, houses or persons
- Aggregates: can be sums, totals, averages of fields, counts or densities of objects, associated with lines or regions

5.2. Aggregating and summarising

We also can aggregate records (rows) in a table considering:

- Grouping records based on a grouping predicate
- Applying an aggregation function to the attribute values of a group to summarise them into a single number.

Thanks!...

Pebesma, E., & Bivand, R. (2023). Spatial Data Science: With Applications in R. CRC Press.

Lovelace, R., Nowosad, J., & Muenchow, J. (2019). Geocomputation with R. CRC Press.

Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V., & Pebesma, E. J. (2008). Applied spatial data analysis with R (Vol. 747248717, pp. 237-268). New York: Springer.

Bureau International des Poids et Mesures. 2006. The International System of Units (SI), 8th Edition. Organisation Intergouvernementale de la Convention du Mètre. https://www.bipm.org/en/publications/si-brochure/download.html.

Lott, Roger. 2015. "Geographic Information-Well-Known Text Representation of Coordinate Reference Systems." Open Geospatial Consortium. http://docs.opengeospatial.org/is/12-063r5/12-063r5.html.

Evenden, Gerald I. 1990. Cartographic Projection Procedures for the UNIX Environment — a User's Manual. http://download.osgeo.org/proj/OF90-284.pdf.