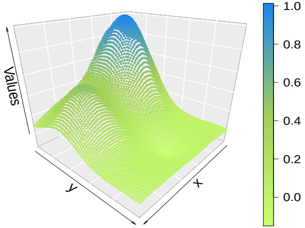
Statistical methods for spatial data analysis

Joaquin Cavieres

Geoinformatics, Bayreuth University



Outline

About me

Joaquin Cavieres, Dr. in Statistics (Universidad de Valparaíso, Chile).

Currently I’m a Postdoctoral Fellow in the Geoinformatics-Big Spatial Data group lead by professor Dr. Meng Lu.

My research interest is related with:

- Spatial statistics
- Bayesian inference
- Computational methods

Personal web page: `http://www.joaquincavieres.com`

About the course:

- An introduction to important topics/methods in spatial data analysis.
- R programming for spatial data analysis.
- Develop your own project.

Schedule:

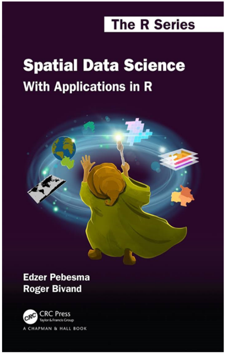
Wednesday from 2:00 pm to 6:00 pm

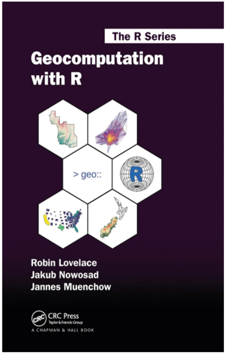
Grading:

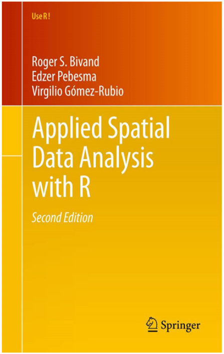
- Computer lab + assignments (30%)
- Project (70%)

Material of study:

- Spatial Data Science with applications in R [1]
- Geocomputation with R [2]
- Applied Spatial Data Analysis with R [3]







Project:

About

Statistical methods (including statistics and machine learning) for Geo-scientific data analysis, method development, improvement, application, etc.

Themes

- Spatial or spatiotemporal prediction in spatial data
- Quantification of the uncertainty
- Time series change detection

Project (cont.):

Submission

A project report (pdf) and the scripts reproducing the results.

Presentation

- Proposal presentation and project presentation.
- A project will be created on e-learning for more information.

Structure of the report (general):

1. Introduction
   - Motivation, literature review, challenges, objectives.
2. Data and study area
3. Method
4. Implementation
5. Results and discussion
   - Result analysis, limitations, improvements, etc.
6. Conclusion

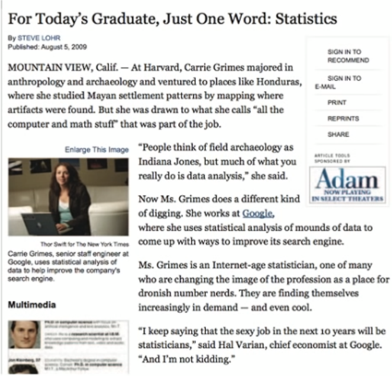
1. Introduction

How do we analyse Geospatial data?

How do we analyse Geospatial data?

Statistical methods: Statistics and Machine learning

"Modern Statistics" includes machine learning, thus a new field called "statistical learning" born..


For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Ms. Grimes is an Internet-age statistician, one of many who are changing the image of the profession as a place for dronish number nerds. They are finding themselves increasingly in demand — and even cool.

"I keep saying that the sexy job in the next 10 years will be statisticians," said Hal Varian, chief economist at Google. "And I'm not kidding."

Enlarge This Image

Thor Swift for The New York Times
Carrie Grimes, senior staff engineer at Google, uses statistical analysis of data to help improve the company's search engine.

Multimedia

SIGN IN TO RECOMMEND
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Adam
NOW PLAYING IN SELECT THEATERS

QUOTE OF THE DAY, NEW YORK TIMES, AUGUST 5, 2009

"I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding." — HAL VARIAN, chief economist at Google.

Figure 1: Description of the importance of Statistics today. Source: Internet

Statistics modelling

Statistical methods are commonly interested on model interpretation and uncertainty assessment (how certain we are about my model and predictions).

Machine learning

Commonly it's more interested on predictions

Machine learning

Commonly it's more interested on predictions

However, there is no clear boundaries between Statistics, Machine learning and Statistical learning!!

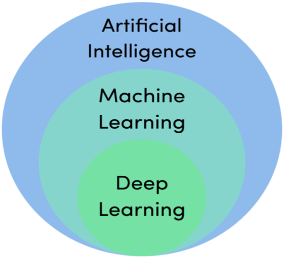


Figure 2: Areas of predictive modelling using Statistics and data modelling

2. Motivation

Geospatial data

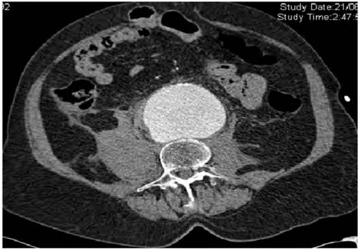
Geospatial data is data about objects, events, or phenomena that have a location on the surface of the earth.

The location may be static in the short-term (e.g., the location of a road, an earthquake event, children living in poverty), or dynamic (e.g., a moving vehicle or pedestrian, the spread of an infectious disease).

Geospatial data combines location information (usually coordinates on the earth), attribute information (the characteristics of the object, event, or phenomena concerned), and often also temporal information (the time or life span at which the location and attributes exist). So,

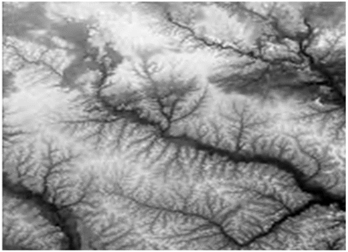
geospatial data = spatial data?

Spatial data, but not Geographical



Medical imaging, brain scan

Geospatial data



Elevation

Figure 3: Types of "spatial" data

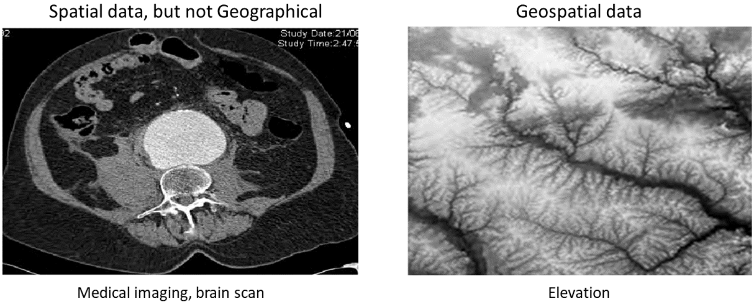
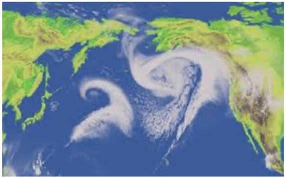


Figure 3: Types of "spatial" data

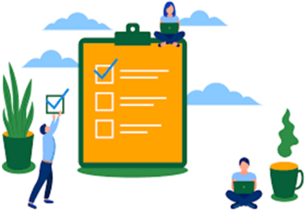
In Geosciences, “spatial” and “Geospatial” are used interchangeably. But you should be aware when you talk to people from other disciplines.

Large and diverse spatial data

Citizen science data



Numerical model simulations



Survey data

Figure 4: Spatial data from different sources (1). Source: Internet

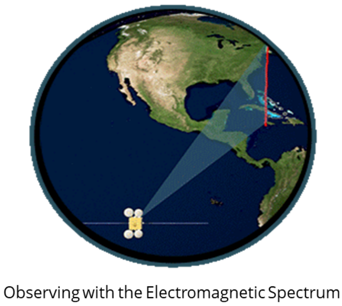
Large and diverse spatial data

Mobile measurements

Social media

Figure 5: Spatial data from different sources (2). Source: Internet

Remote sensing and satellite missions



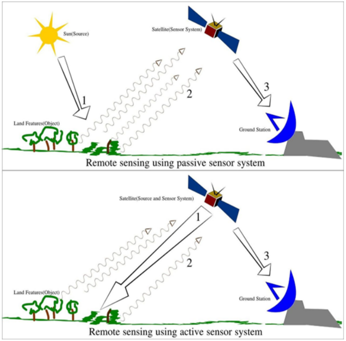
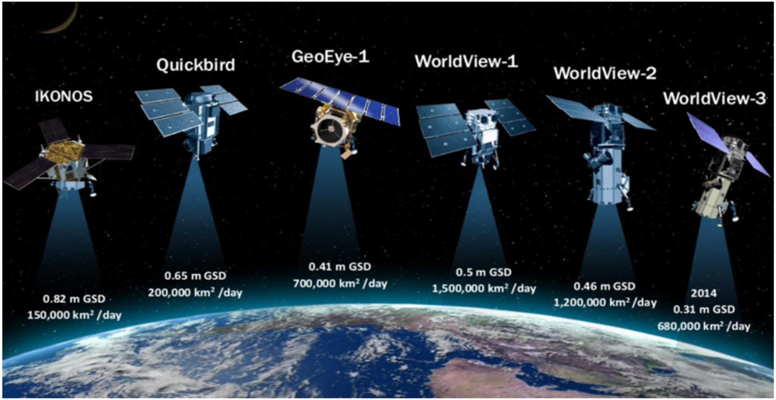


Figure 6: Spatial data from different sources (3). Source: `https://earthdata.nasa.gov/learn/backgrounders/remote-sensing`

Satellite missions: urban environment



The ground sampling distance (GSD), in mm/px, is the distance between the centers of two adjacent pixels, measured on the observed object. A GSD of **1mm/px** means that one pixel on the image represents 1 mm in the real world.

Figure 7: Spatial data from different sources (4). Source: Internet

3. Spatial data

3.1. Coordinates

For spatial data, the location of observations are characterised by coordinates, and coordinates are defined in a coordinate system.

For the above, different coordinate systems can be used, over a 2-dimensional or 3-dimensional space referenced to orthogonal axes (Cartesian coordinates), or using distance and directions (polar coordinates, spherical and ellipsoidal coordinates).

3.1.1. Quantities, units, datum

A measurement system consist of base units for base quantities, and derived units for derived quantities. For instance, the SI system of units [4] consist of seven base units: length (metre, m), mass (kilogram, kg), time (second, s), electric current (ampere, A), thermodynamic temperature (Kelvin, K), amount of substance (mole, mol), and luminous intensity (candela, cd).

For many quantities, the natural origin of values is zero but, for locations and times, differences have a natural zero interpretation: distance and duration. Absolute location (position) and time need a fixed origin, from which we can meaningfully measure other absolute space-time points: we call this a datum.

For space, a datum involves more than one dimension. The combination of a datum and a measurement unit (scale) is a *reference system*. In this way, we can express spatial locations as ellipsoidal or Cartesian coordinates.

3.1.2. Ellipsoidal coordinates

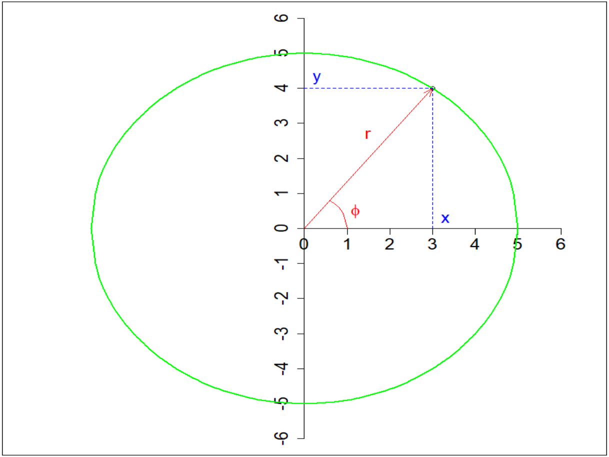


Figure 8: Two-dimensional polar (red) and Cartesian (blue) coordinates. Source: [3]

Polar and Cartesian coordinates are showed in the figure 8, where:

- Cartesian coordinates, the point shown is $(x, y) = (3, 4)$.
- Polar coordinates, the point shown is $(r, \phi) = (5, \text{arctan}(4/3))$.

Converting back and forth between Cartesian and Polar,

$$
\begin{aligned}
x &= r\ cos\phi, \;\; y = r\ sin\phi && (1) \\
r &= \sqrt{x^2 + y^2}, \;\; \phi = \text{atan2}(y, x) && (2)
\end{aligned}
$$

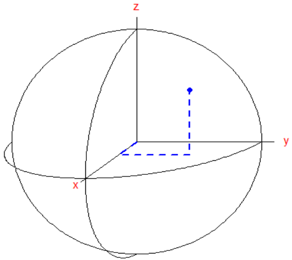
where atan2 is used in favour (y/x) of to take care of the right quadrant.

3.1.3. Spherical or ellipsoidal coordinates

For three dimensions, Cartesian coordinates are expressed commonly as (x, y, z). Spherical coordinates are the three dimensional equivalent of polar coordinates and can be expressed as (r, λ, ϕ), where:

- r is the raduous of the sphere
- λ is the longitude measured in the (x, y) plane
- ϕ is the latitude. the angle between the vector and the (x, y) plane

See next figure as example.



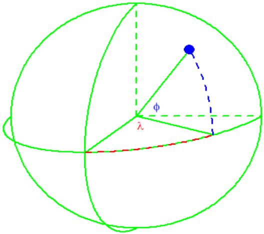


Figure 9: Cartesian geocentric coordinates (left) measure three distances, ellipsoidal coordinates (right) measure two angles, and possibly an ellipsoidal height. Source: [3]

We could add altitude or elevation to define points that are above or below the ellipsoid, and obtain a three dimensional space again, so we need to define:

- Where zero altitude is?
- Which direction is positive?
- Which direction is “straight up”?

The shape of the Earth is not a perfect ellipsoid. Thus, several ellipsoids with different shape parameters and bound to the Earth in different ways could be used. The name of those ellipsoids are called *datums*.

3.1.4. Projected coordinates, distances

Computing the locations in a two-dimensional space means that we work with projected coordinates. Projecting ellipsoidal coordinates means that shapes, directions, areas, or even all three, are distorted.

To calculate the distance between two points s_i and s_j in Cartesian coordinates is done commonly by the Euclidean distance:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{3}$$

where $s_i = (x_i, y_i)$.

In three dimensions:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{4}$$

where $s_i = (x_i, y_i, z_i)$.

3.1.4. Bounded and unbounded spaces

Two-dimensional (R^2) and three-dimensional (R^3) Euclidean spaces are not bounded.

On the other hand, spaces defines on a circle (S^1) or sphere (S^2) define a bounded set. That is, many points but the length and area of the circle and the radius, area and volume of a sphere are bound.

3.2. Coordinate Reference Systems

We will consider the following definitions [5]:

- Coordinate system: is a set of mathematical rules for specifying how coordinates are to be assigned to points,
- Datum: is a parameter or set of parameters that define the position of the origin, the scale, and the orientation of a coordinate system.
- Geodetic datum: is a datum describing the relationship of a two- or three-dimensional coordinate system to the Earth.
- Coordinate reference system: is a coordinate system that is related to an object by a datum; for geodetic and vertical datums, the object will be the Earth.

3.3. PROJ and mapping accuracy

PROJ [6] started in the 1970s as a Fortran project, and was released in 1985 as a C library for cartographic projections. In the early 2000s, PROJ was known as PROJ.4, after its never changing major version number.

PROJ definitions for coordinate reference systems would look like this:

```
+proj=utm +zone=33 +datum=WGS84 +units=m +no_defs
```

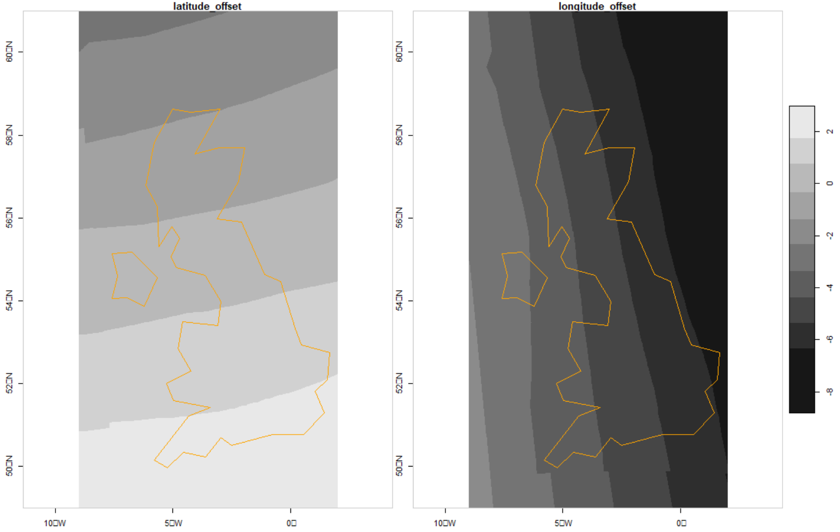


Figure 10: UK horizontal datum grid, from datum OSGB 1936 (EPSG:4277) to datum ETRS89 (EPSG:4258); units arc-seconds. Source: [3]

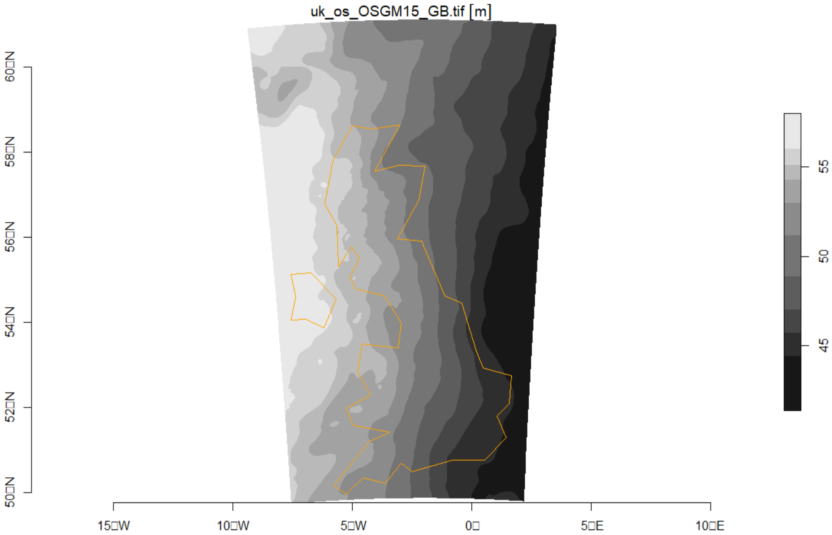


Figure 11: UK vertical datum grid, from ETRS89 (EPSG:4937) to ODN height (EPSG:5701), units m. Source: [3]

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.