Geographic Information Systems: Raster and Vector Systems ENV5188 FALL 2025

October 29th, 2025

Erich Seamon
Assistant Professor
Environmental Science
BSB C423
erich_seamon@baylor.edu
https://haclab.io

Baylor University

Preparation

- R and RStudio should be installed, with appropriate libraries
- Able to follow along with lessons from github or in Rstudio
  - https://github.com/Baylor-HACLab-Classwork/BAYLOR_5188_FALL2025
- Be prepared for technical snafus. It's going to happen. Let's try to roll with it.

Miscellaneous

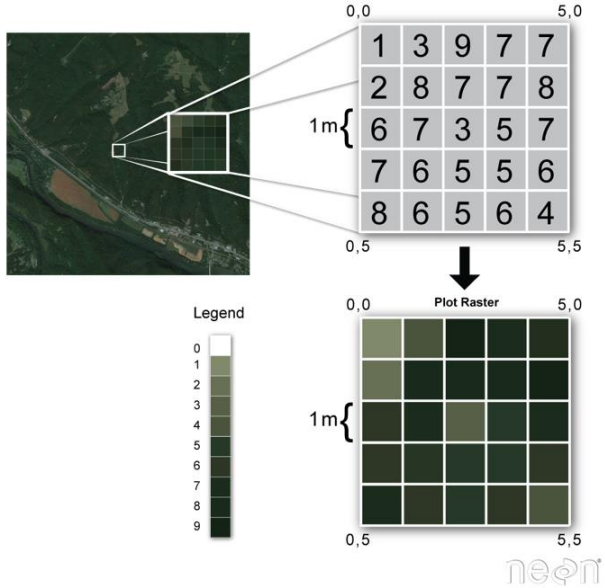
- Introductory exercises in basics of R in /src/introductory folder
- Extra exercises in differing geospatial topics in /src/additional_exercises
- Package install issues can be annoying. /src/day1-00-prep-librarytest.R
- Class website has links to the data carpentry exercises that are for day 1 and 2 (Tues, April 26 and Thurs April 28)
- Use Zoom for chat communications
- We are recording the workshop via zoom and will post it after completion

Miscellaneous

- We are skipping over the introductory R components, how to use Rstudio, how to use ggplot
- I have minimized, but have still included, components on manipulating raster and vector data

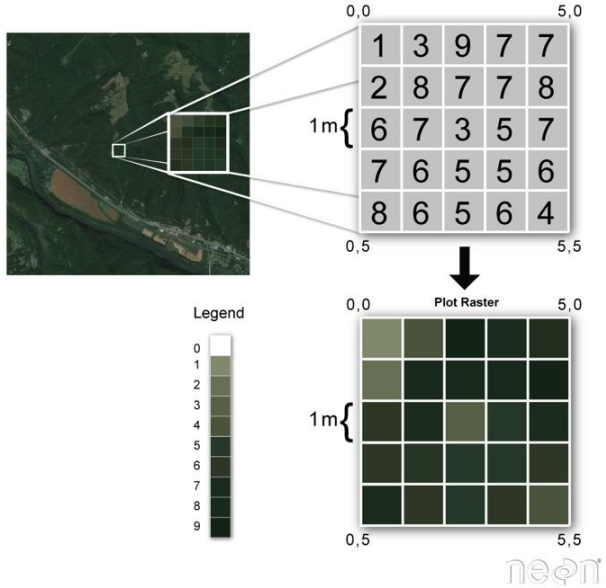
Introduction to Raster Data

- The two primary types of geospatial data are **raster** and **vector** data.
- Raster data is stored as a grid of values which are rendered on a map as pixels. Each pixel value represents an area on the Earth's surface.
- Vector data structures represent specific features on the Earth's surface and assign attributes to those features.



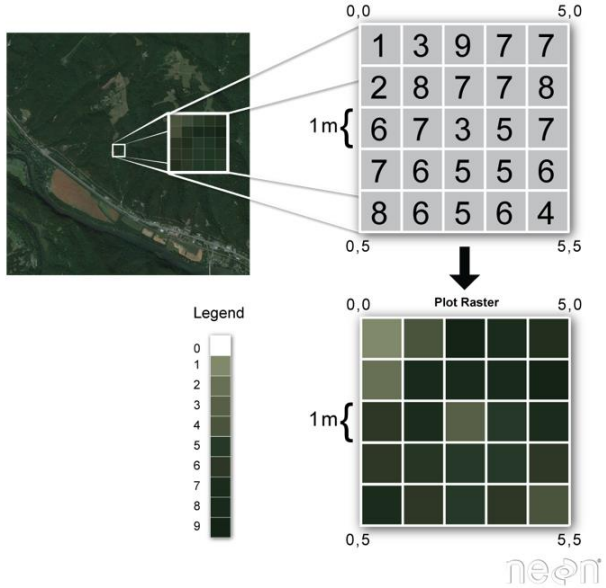
Introduction to Raster Data

- Raster data is any pixelated (or gridded) data where each pixel is associated with a specific geographical location.
- The value of a pixel can be continuous (e.g. elevation) or categorical (e.g. land use).

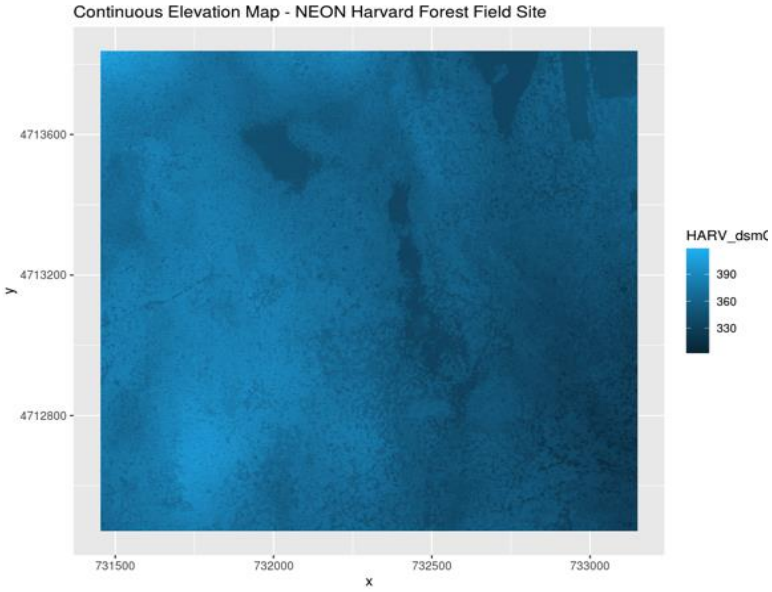


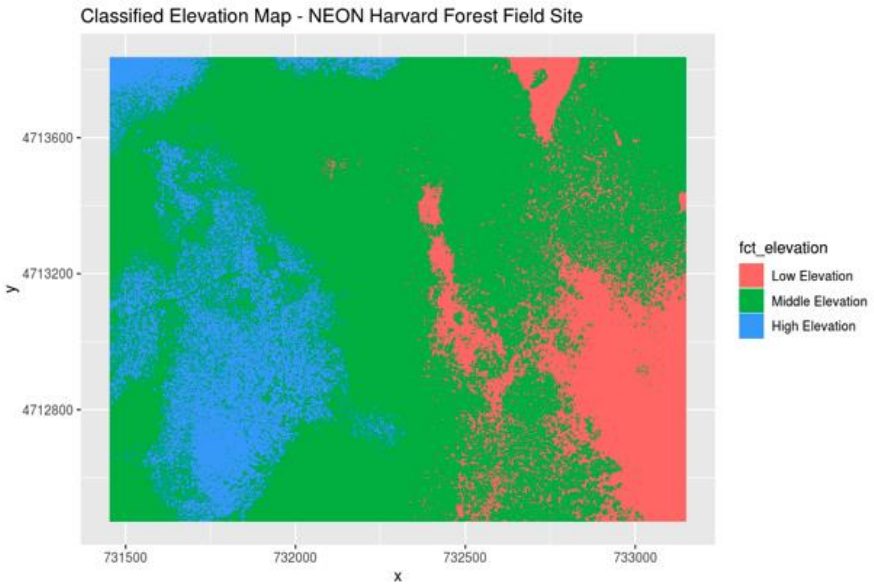
Introduction to Raster Data

- A geospatial raster is only different from a digital photo in that it is accompanied by spatial information that connects the data to a particular location.
- This includes items such as a raster's **extent** and **cell size**, the number of rows and columns, its **projection** and **coordinate reference system (or CRS),** as well as any associated attribute information.



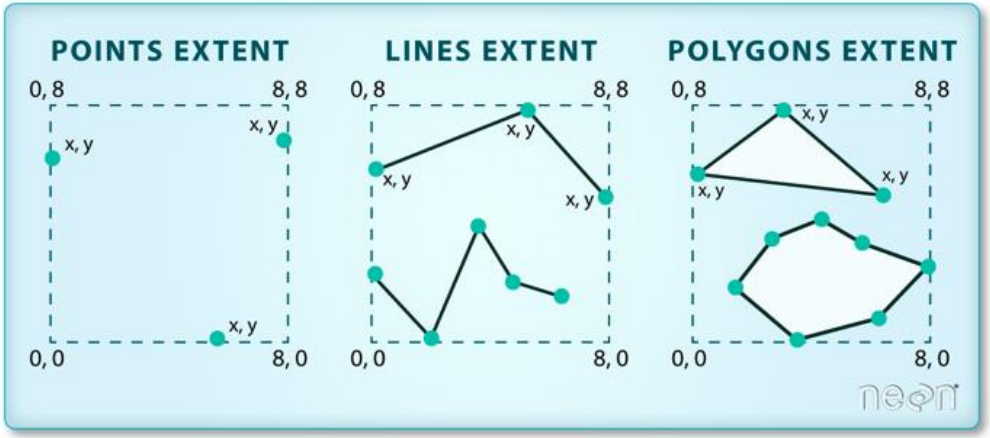
Continuous vs. Categorical





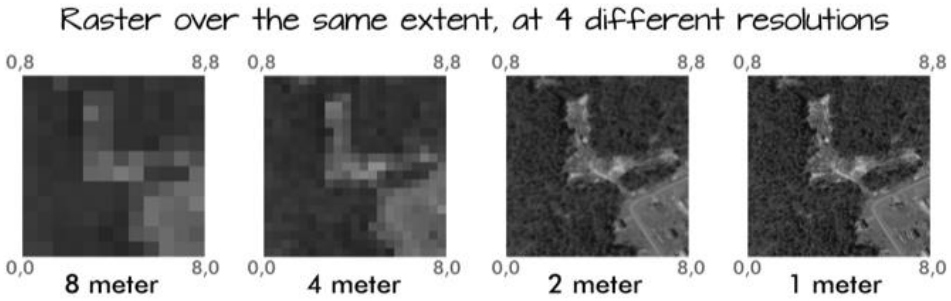
Important Aspects

- **Spatial extent**
- Resolution
  - We will focus on **spatial** resolution, but there are other forms of resolution, including **spectral**, **temporal**, and **radiometric**



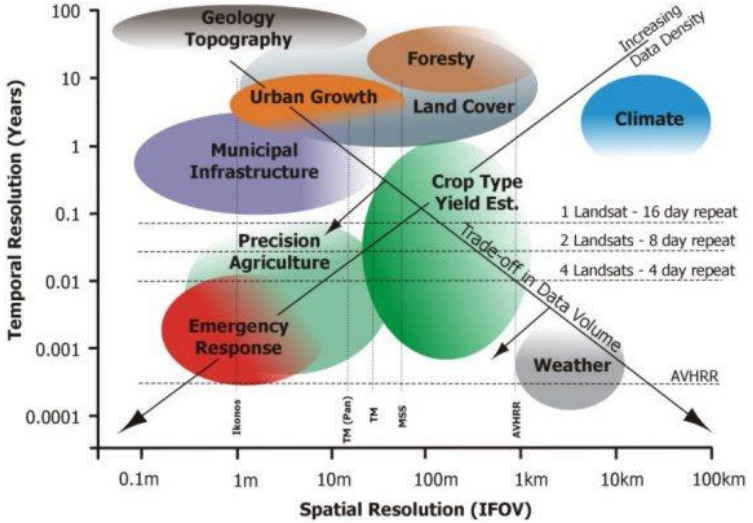
Important Aspects

- Spatial extent
- **Resolution**
  - We will focus on spatial resolution, but there are other forms of resolution, including **spectral**, **temporal**, and **radiometric**



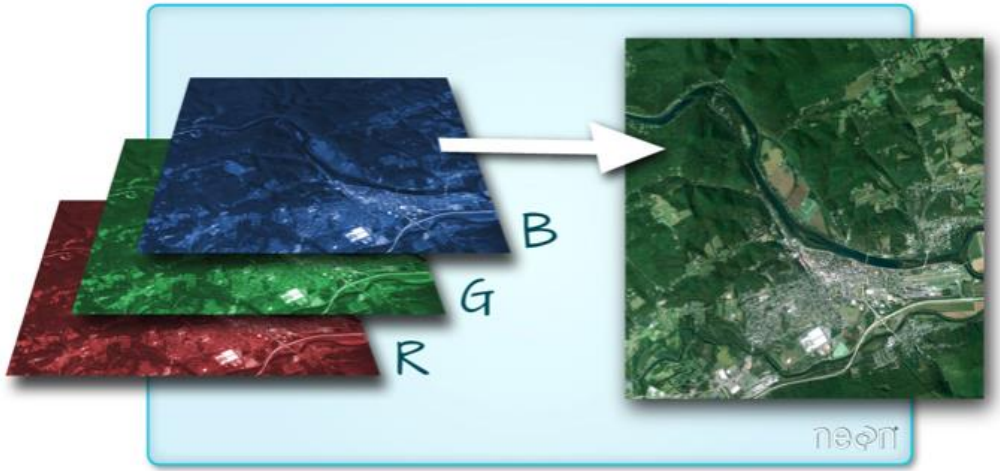
Important Aspects

- Spatial extent
- **Resolution**
  - We will focus on spatial resolution, but there are other forms of resolution, including **spectral**, **temporal**, and **radiometric**



Multi-Band Data

- A raster can contain one or more bands.
- One type of multi-band raster dataset that is familiar to many of us is a color image.
- A basic color image consists of three bands: red, green, and blue. Each band represents light reflected from the red, green or blue portions of the electromagnetic spectrum.



Coordinate Reference Systems (CRS)

- The CRS associated with a dataset tells your mapping software (for example R) where the raster/vector is located in geographic space.
- It also tells the mapping software what method (projection) should be used to flatten or project the raster in geographic space.

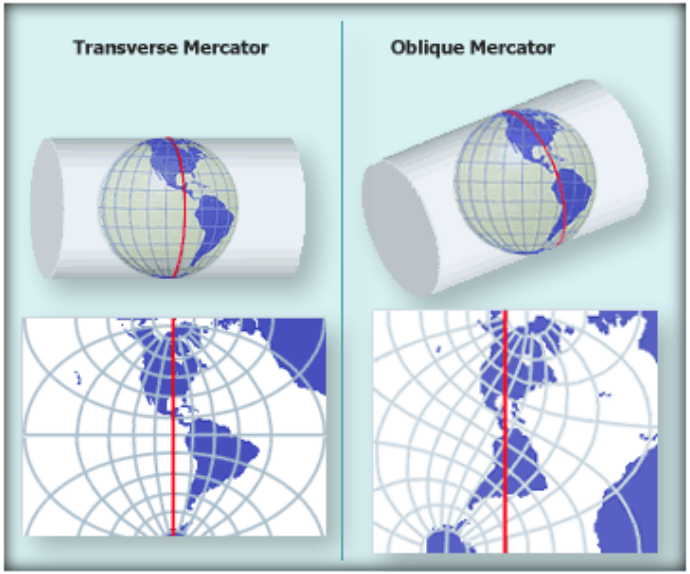
Coordinate Reference Systems (CRS)

- Key components of a CRS are:
  - Datum – a model of the shape of the earth (ex. WGS84, NAD83)
  - Projection – mathematical transform of angular measurements from spheroidal to flat. May include a zonal information if UTM
  - Ellipsoid - mathematical surface obtained by revolving an ellipse about the earth's polar axis. Selected to give a good fit to the geoid

Coordinate Reference Systems (CRS)

Map Projections: to convert geodetic positions of a portion of the earth's surface to plane rectangular coordinates, points are projected mathematically from the ellipsoid to some imaginary developable surface - plane that can be rolled out flat

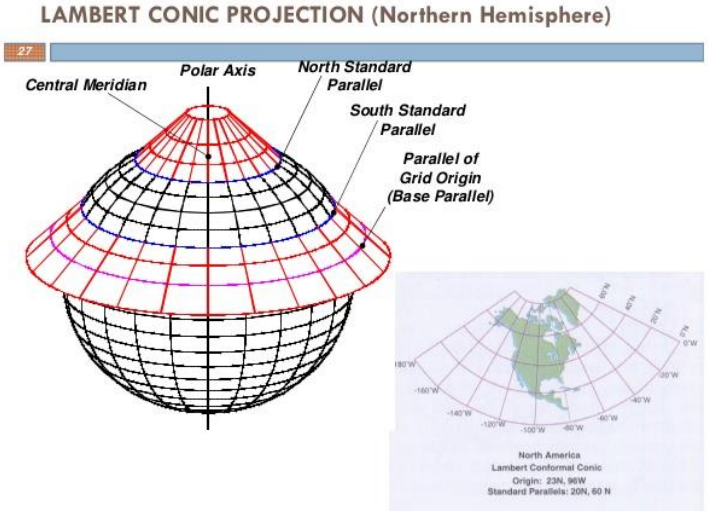
Coordinate Reference Systems: quantitative coordinate systems - based on mathematical projection models, often a cartesian coordinate system (i.e. x, y axes) representing relative positions within a particular map projection



Coordinate Reference Systems (CRS)

Map Projections: to convert geodetic positions of a portion of the earth's surface to plane rectangular coordinates, points are projected mathematically from the ellipsoid to some imaginary developable surface - plane that can be rolled out flat

Coordinate Reference Systems: quantitative coordinate systems - based on mathematical projection models, often a cartesian coordinate system (i.e. x, y axes) representing relative positions within a particular map projection



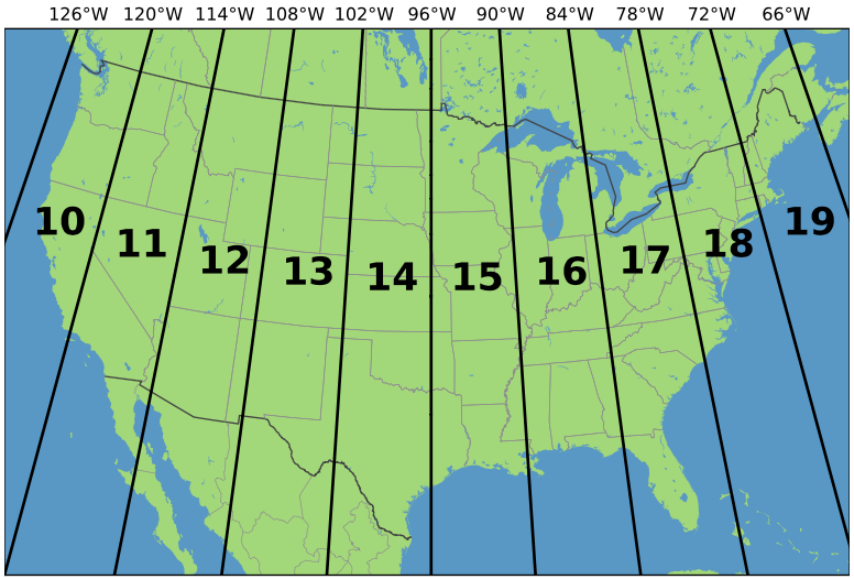
Coordinate Reference Systems (CRS)

PROJ is an open-source library for storing, representing and transforming CRS information. PROJ.5 has been recently released, but PROJ.4 was in use for 25 years so you will still mostly see PROJ referred to as PROJ.4. PROJ represents CRS information as a text string of key-value pairs, which makes it easy to customize (and with a little practice, easy to read and interpret).

A PROJ4 string includes the following information:

proj=: the projection of the data
zone=: the zone of the data (this is specific to the UTM projection)
datum=: the datum use
units=: the units for the coordinates of the data
ellps=: the ellipsoid (how the earth's roundness is calculated) for the data
Note that the zone is unique to the UTM projection. Not all CRSs will have a zone.

Coordinate Reference Systems (CRS)



UTM Zones

+proj=utm +zone=18 +datum=WGS84 +units=m +no_defs +ellps=WGS84 +towgs84=0,0,0

Coordinate Reference Systems (CRS)

+proj=utm +zone=18 +datum=WGS84 +units=m +no_defs +ellps=WGS84 +towgs84=0,0,0

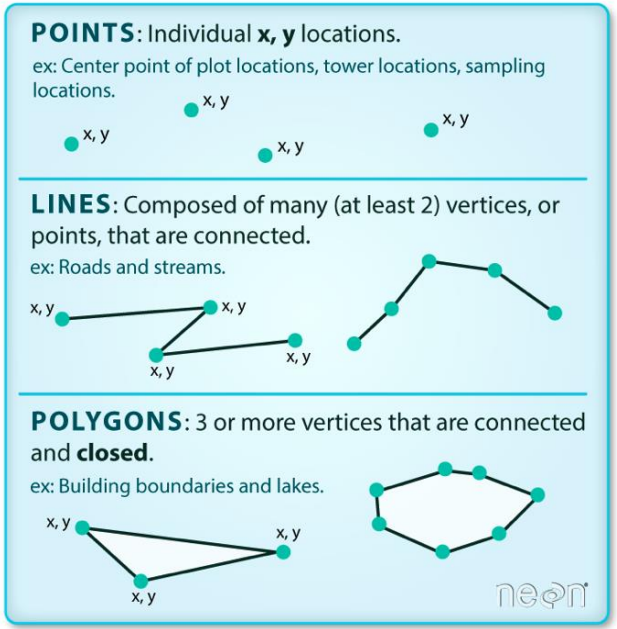
Coordinate Reference Systems (CRS)

- GDAL is a set of software tools that translate between almost any geospatial format in common use today (and some not so common ones).
- GDAL also contains tools for editing and manipulating both raster and vector files, including reprojecting data to different CRSs.

http://gdal.org

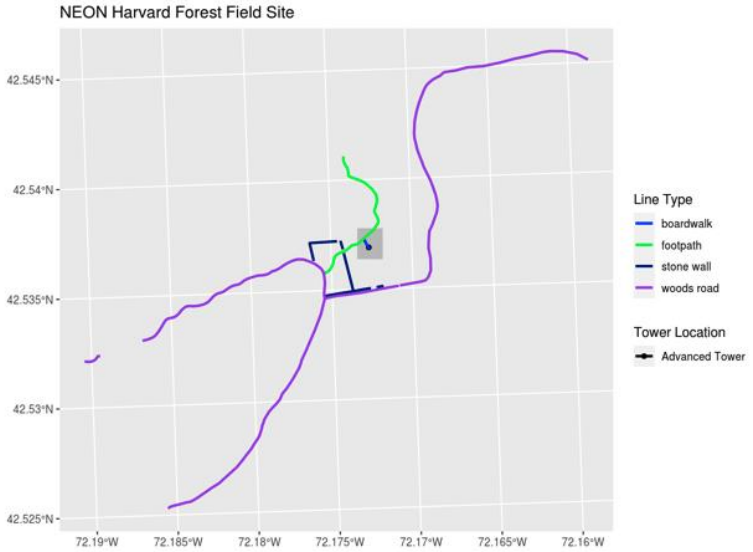
Intro to vector data

- Vector data structures represent specific features on the Earth's surface and assign attributes to those features.
- Vectors are composed of discrete geometric locations (x, y values) known as vertices that define the shape of the spatial object.
- The organization of the vertices determines the type of vector that we are working with: point, line or polygon.



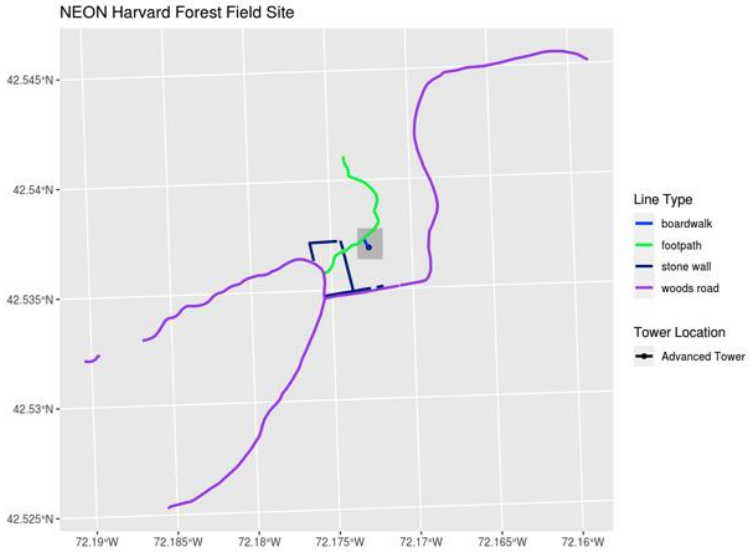
Intro to vector data

- Vector data has some important advantages:
- The geometry itself contains information about what the dataset creator thought was important
- Each geometry feature can carry multiple attributes instead of just one, e.g. a database of cities can have attributes for name, country, population, etc.
- Data storage can be very efficient compared to rasters



Intro to vector data

- The downsides of vector data include:
  - potential loss of detail compared to raster
  - potential bias in datasets - what didn't get recorded?
  - Calculations involving multiple vector layers need to do math on the geometry as well as the attributes, so can be slow compared to raster math.



Intro to vector data

- Like raster data, vector data can also come in many different formats. For this workshop, we will use the **Shapefile** format which has the extension .shp. A .shp file stores the geographic coordinates of each vertice in the vector, as well as metadata including:
- **Extent, Object type,** and the **Coordinate reference system (CRS),** and **Other attributes**: for example, a line shapefile that contains the locations of streams, might contain the name of each stream.
- Shapefiles are an atomic collection (multiple files)

Geospatial Landscape

Commercial

- ESRI
- MAPINFO
- Manifold
- Smallworld

Cloud

- Google Earth Engine
- ArcGIS Online

Open source

- QGIS
- GRASS
- GDAL
- PostGIS/Postgres
- R/Python
- Cesium

Geospatial Landscape

Geospatial Libraries: R: sf, sp, gdal, spplot, leaflet, spacetime, ncdf4

Geospatial Libraries: Python: shapely, geopandas, rasterio, gdal, rasterstats

Modeling: R: caret, mlr

Modeling: Python: tensorflow, keras, sci-kit learn, numpy/scipy, Pytorch

Visualization: R: ggplot2, ggpubr, seaborn, plotly

Visualization: Python: matplotlib

https://cran.r-project.org/web/views/Spatial.html

Advanced Spatial Modeling

- Spatial Autocorrelation
- Kriging
- Spatially Weighted Regression

Spatial Randomness

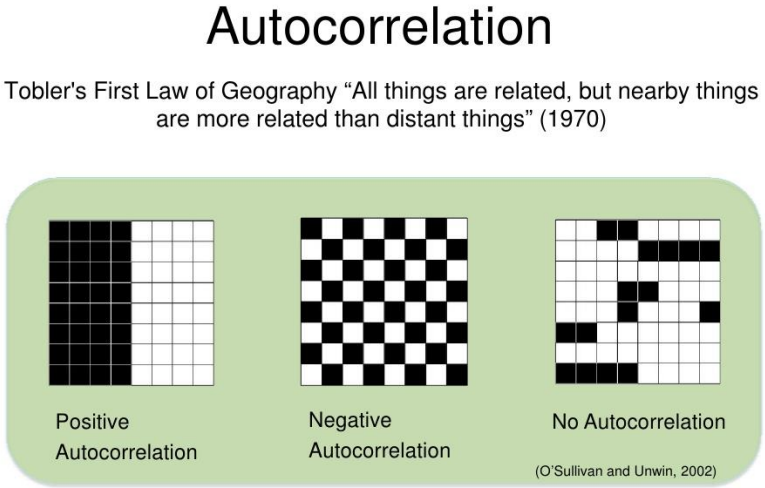
- **Spatial randomness** – no pattern
- If spatial randomness is rejected, then there is a spatial structure
- Value at one location does not depend on values at other neighboring locations

Spatial Autocorrelation

A measure of the degree to which a set of **spatial** features and their associated data values tend to be clustered together in space (positive **spatial autocorrelation**) or dispersed (negative **spatial autocorrelation**). Cliff and Ord 1973, 1981

Random **– no autocorrelation – Null Hypothesis**
If Null is rejected, the alternative hypotheses are:

Clustered **– above zero Alternative Hypothesis**
Dispersed **– below zero Alternative Hypothesis**



Quantifying SAC

- Statistical testing (using a test statistic)
- How likely is the test statistic value if it had occurred under the null hypothesis (spatial randomness)
- When unlikely – the null is rejected (low p value)
- For SAC – we are most interested in capturing/combining
  - **Attribute similarity –** summary of similarity/dissimilarity of observations of a variable at differing locations $f(x_i, x_j)$
  - **Locational similarity –** formalizing the notion of neighbors. Construction of spatial weights w_{ij}

$$\sum_{ij} f(x_i\ xj)\ \ wij$$

$f(x_i x_j)$ is attribute similarity between *i* and *j* for x

w_{ij} is spatial weight between *i* and *j*

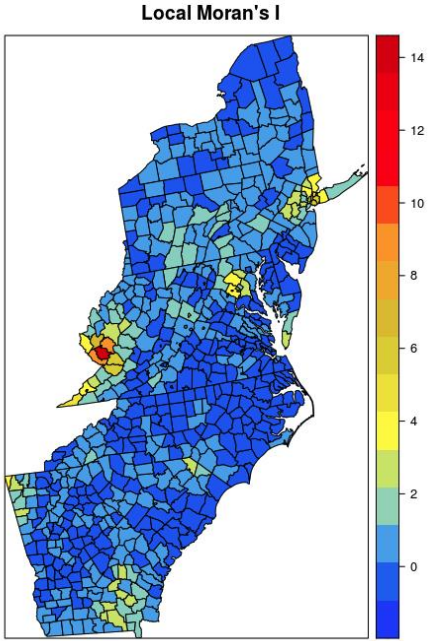
Morans I

Moran's I is an inferential statistic, which means that the results of the analysis are always interpreted within the context of its null hypothesis.

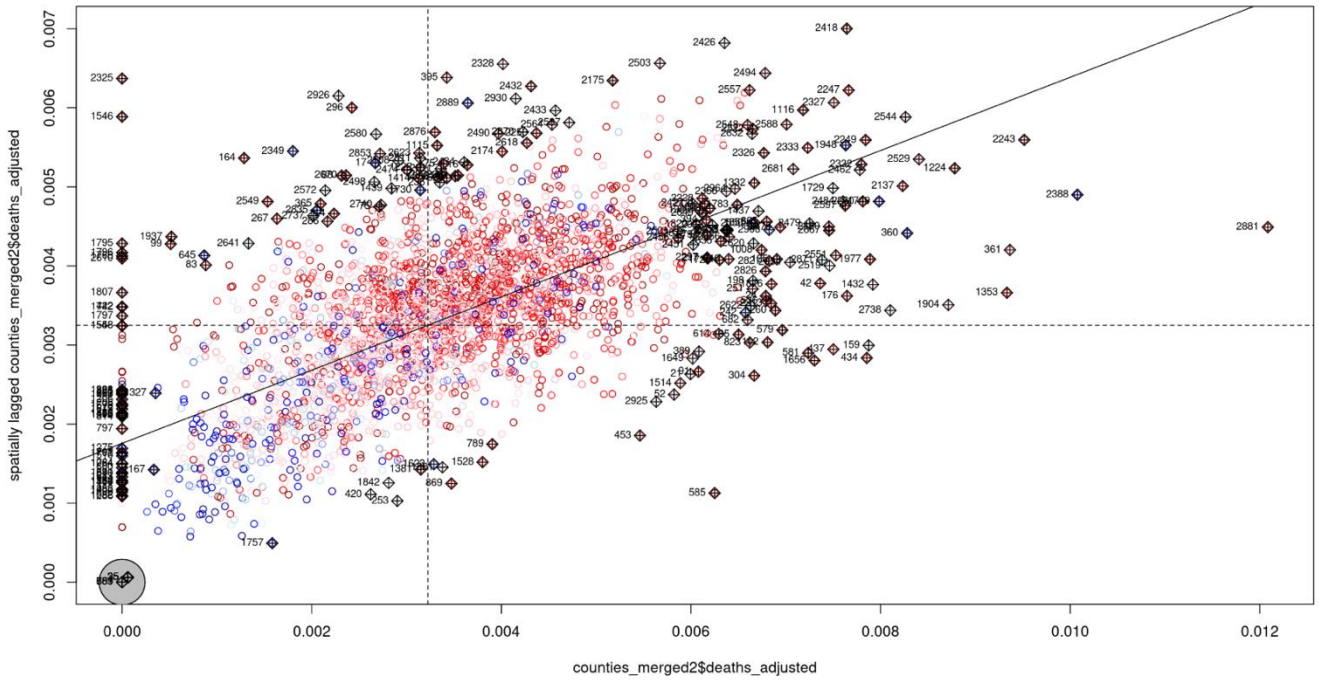
Like a correlation coefficient, values of Moran's I range from +1 meaning strong positive spatial autocorrelation to 0 meaning a random pattern to -1 indicating strong negative spatial autocorrelation.

Global Moran's I provides a one single value, which is the average across the dimensional space.

Local Moran's I statistic was suggested in Anselin (1995) as a way to identify local clusters and spatial outliers.



United States: Morans I: population adjusted deaths



Gearys C

Gearys C determines if adjacent observations of the same attributes are correlated in multi- or bi-directional ways.

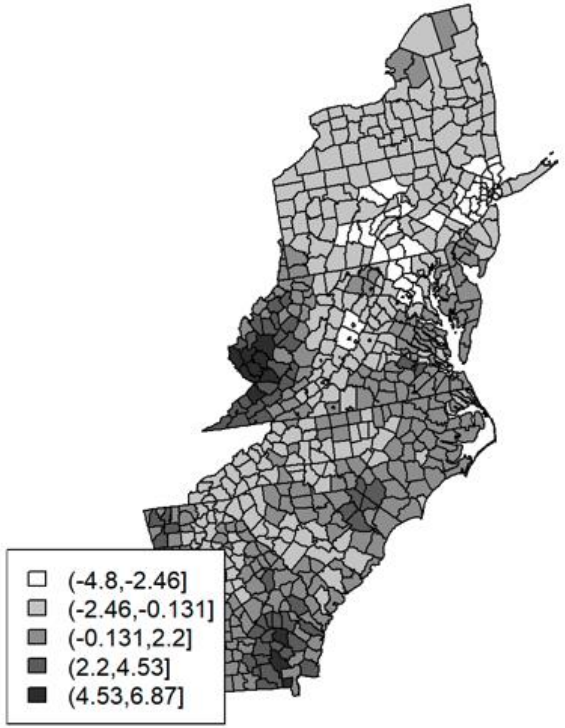
The value of Geary's C lies between 0 and some general value greater than 1. Values significantly lower than one demonstrate increasing positive spatial autocorrelation, while values significantly higher than one illustrate increasing negative spatial autocorrelation.

Geary's C is inversely related to Moran's I, but it is not identical. Moran's I is a measure of global spatial autocorrelation, while Geary's C is more sensitive to local spatial autocorrelation.

Geary's test for spatial autocorrelation using a spatial weights matrix in weights list form. The assumptions underlying the test are sensitive to the form of the graph of neighbour relationships and other factors, and results may be checked against those of the geary.mc permutation

Getis-Ord Gi*

Hotspot analysis using *Getis-Ord Gi** statistic (sometimes referred to as GI-star) uses spatial vectors to identify the locations of statistically significant hot spots and cold spots in data. The z-scores and p-values indicates where features with either high or low values cluster spatially.

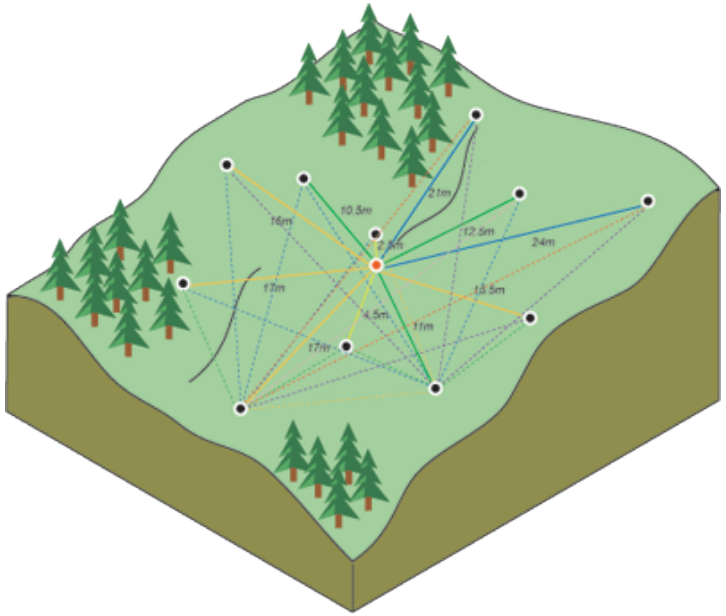


Introduction to Kriging

Kriging is a group of geostatistical techniques to interpolate the value of a random field at an un-sampled location from known observations of its value at nearby locations.

The main statistical assumption behind kriging is one of **stationarity** which means that statistical properties (such as **mean and variance**) do not depend on the exact spatial locations, so the mean and variance of a variable at one location is equal to the mean and variance at another location.

Variogram/Semivariogram

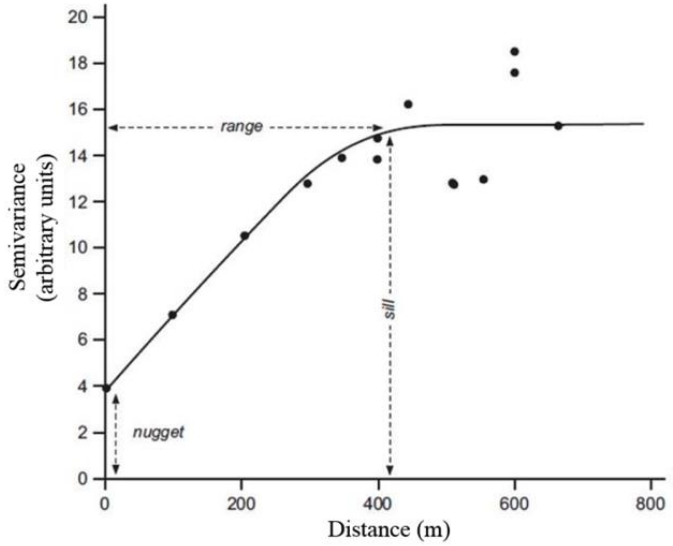


Fitting a model, or spatial modeling, is also known as structural analysis, or variography. In spatial modeling of the structure of the measured points, you begin with a graph of the empirical semivariogram, computed with the following equation for all pairs of locations separated by distance **h**:

$$\text{Semivariogram}(\text{distance}_h) = 0.5 * \text{average}((\text{value}_i - \text{value}_j)^2)$$

The formula involves calculating the difference squared between the values of the paired locations.

Variogram/Semivariogram



The semivariogram depicts the spatial autocorrelation of the measured sample points. Once each pair of locations is plotted, a model is fit through them. There are certain characteristics that are commonly used to describe these models.

The Semivariogram and covariance both measure the strength of statistical correlation as a function of distance.

Kriging

- Kriging predicts the value at a given point by computing a weighted average of the known values of the function in the neighborhood of the point.
- Unlike other deterministic interpolation methods such as inverse distance weighting (IDW) & Splining, kriging is based on the statistical relationships among the measured points to interpolate the values in the spatial field.

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i)$$

where:

$Z(s_i)$ = the measured value at the ith location

λ_i = an unknown weight for the measured value at the ith location

s_0 = the prediction location

N = the number of measured values

Kriging

- Kriging produces a prediction surface with uncertainty. Although **stationarity (constant mean and variance) and isotropy (uniformity in all directions)** are the two main assumptions for kriging to provide best linear unbiased prediction, there is flexibility of these assumptions for various forms and methods of kriging

$$\hat{Z}(s_0) = \sum_{i=1}^{N} \lambda_i Z(s_i)$$

where:

$Z(s_i)$ = the measured value at the ith location

λ_i = an unknown weight for the measured value at the ith location

s_0 = the prediction location

N = the number of measured values

In IDW, the weight depends solely on the distance to the prediction location. However, with the kriging method, the weights are based not only on the distance between the measured points and the prediction location but also on the overall spatial arrangement of the measured points

Geographically Weighted Regression

Geographically weighted regression (GWR) is a useful tool for exploring spatial heterogeneity in the relationships between variables where non-stationarity is taking place on the space, that is where locally weighted regression coefficients move away from their global values.

It allows us to understand changes in importance of different variables over space. First In GWR, the appropriate bandwidth needs to be selected for an isotropic spatial weights kernel (typically a Gaussian kernel), with a fixed bandwidth chosen by leave-one-out cross-validation.

$$y_i = \beta_{i0} + \sum_{k=1}^{m} \beta_{ik} x_{ik} + \epsilon_i$$

Where:
yi is the dependent variable as location *i*;
xik is the value of the kth independent variable @ location *i*;
m is the number of independent variables;
Bio is the intercept at location *i*;
Bik is the local regression coefficient for the kth independent variable at location *i*
Ei is the random error at location *i*

Geographically Weighted Regression

The traditional regressions like ordinary least square (OLS) tend to ignore the spatial dependence:

$$y = X\beta + \epsilon$$

The estimates may be:

- Biased (expectation of estimates not equals to the true parameter)
- Inconsistent (estimates not converging to the true parameters as data points increases)
- Inefficient (variance of estimator not to the minimum)

It could be diagnosed that the dependent variable and/or the error term are spatially autocorrelated.

Geographically Weighted Regression

There are two general forms in which spatial autocorrelation enters the regression equation: the spatial lag form and the spatial error form (also called the spatial moving average form).

Spatial Lag Model:

$$y = X\beta + \rho Wy + \epsilon$$

Where W is a spatial weights matrix with elements $w_{ij} = 1$ indicating spatial units i and j are neighbors and $w_{ij} = 0$ otherwise, ρ is the partial regression coefficient for the spatial lag variable. Maximum likelihood estimation is employed to produce estimates.

Geographically Weighted Regression

Spatial autocorrelation in the error terms result from measurement error, or from absent spatially autocorrelated variables influencing variables in the model. Spatial Error Model is formed as:

$$y = X\beta + U$$
$$U = \rho W u + \epsilon$$

Where U is a composite error term including $\rho W u$, spatially autocorrelated errors, and ϵ, the normal error term.

After the spatial lag/error model, the residuals can be tested if there is remaining spatial autocorrelation. To select between the spatial lag/error model, Lagrange Multiplier test is used.

Readings

Cliff, A. D., & Ord, J. K. (1981). Spatial processes: models & applications. Taylor & Francis.

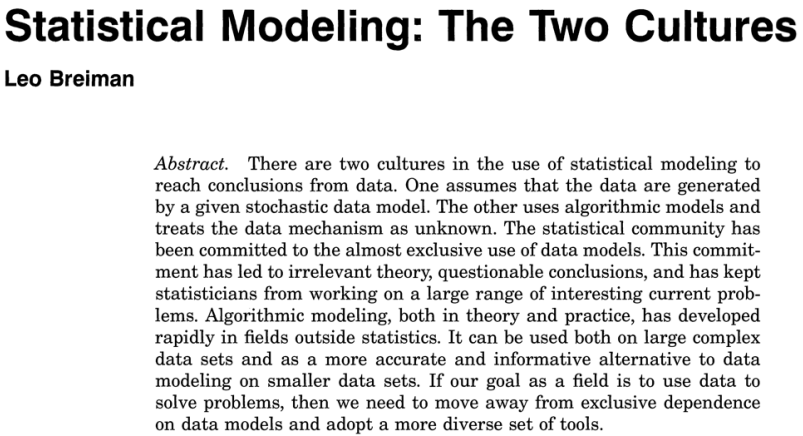
Anselin, L. (1988). Spatial Econometrics: Methods and Models (Vol. 4). Springer Science & Business Media.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2003). Geographically weighted regression: the analysis of spatially varying relationships. John Wiley & Sons.

Expanded Spatial Modeling Approaches

- Random Forest

- http://staff.pubhealth.ku.dk/~tag/Teaching/share/material/Breiman-two-cultures.pdf



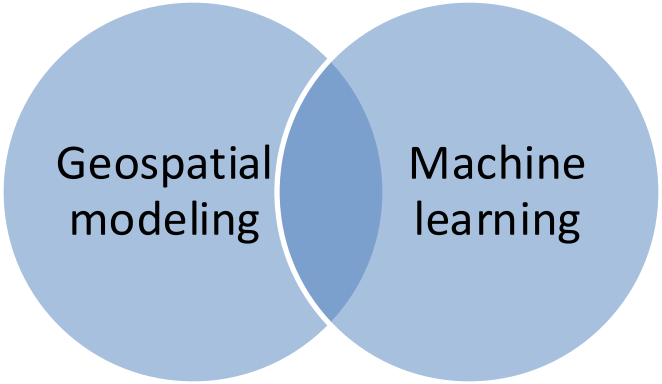
Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

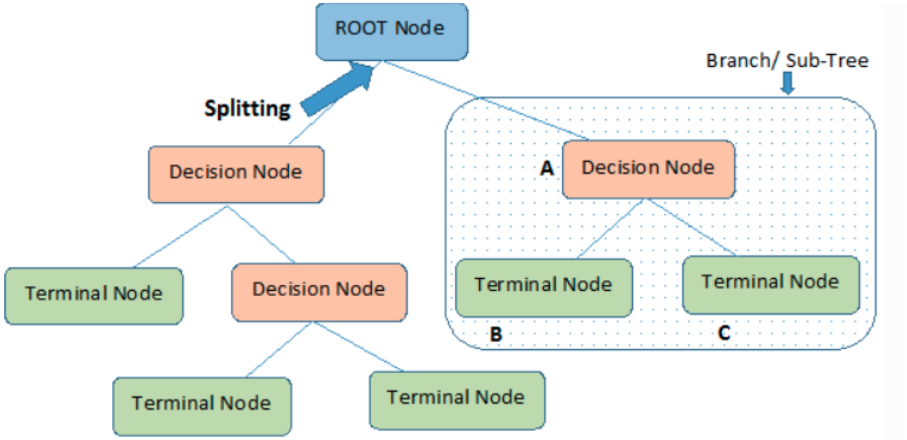
Terms

- Train – validation – test models
- Cross validation
- Ensembling
- Bootstrap aggregation ("bagging")



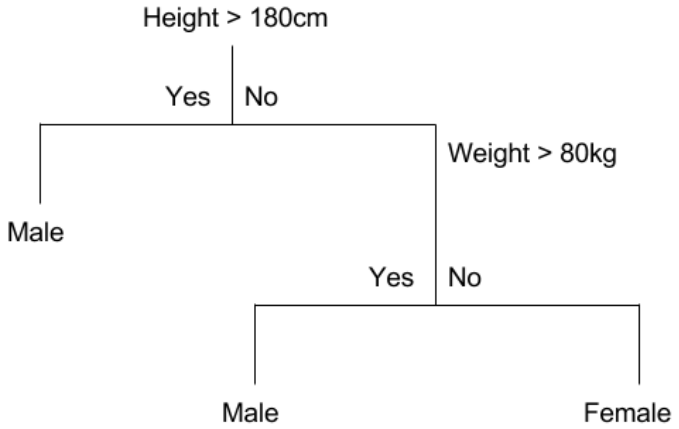
Decision Trees

- Random Forest model is an **ensemble** of single decision trees
- **Let's initially talk about a decision tree**
- **DT based on recursive partitioning**



Decision Trees

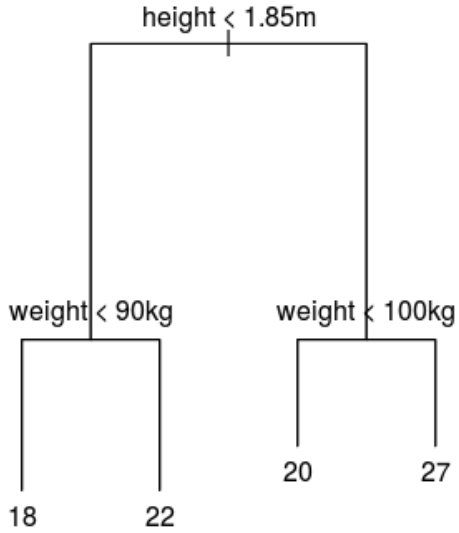
- Random Forest model is an **ensemble** of single decision trees
- **Let's initially talk about a decision tree**
- **DT based on recursive partitioning**



classification

Decision Trees

- Random Forest model is an **ensemble** of single decision trees
- **Let's initially talk about a decision tree**
- **DT based on recursive partitioning**



regression

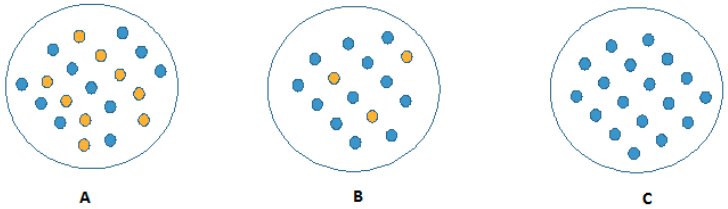
Classification: splitting

- **Gini index**. Assists in determining which split provides the most homogeneous sub nodes. Can only perform binary splits. The lower the Gini value, the higher the homogeneity
- **Information Gain**: Measures disorganization (entropy)

$$\text{Gini} = 1 - \sum_{i=1}^{n} (p_i)^2$$

Classification: splitting

- **Gini index**. Assists in determining which split provides the most homogeneous sub nodes. Can only perform binary splits. The lower the Gini value, the higher the homogeneity
- **Information Gain**: Measures disorganization (entropy)



$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Regression: splitting

- **Reduction in Variance:** This algorithm uses the standard formula of variance to choose the best split. The split with lower variance is selected as the criteria to split the population

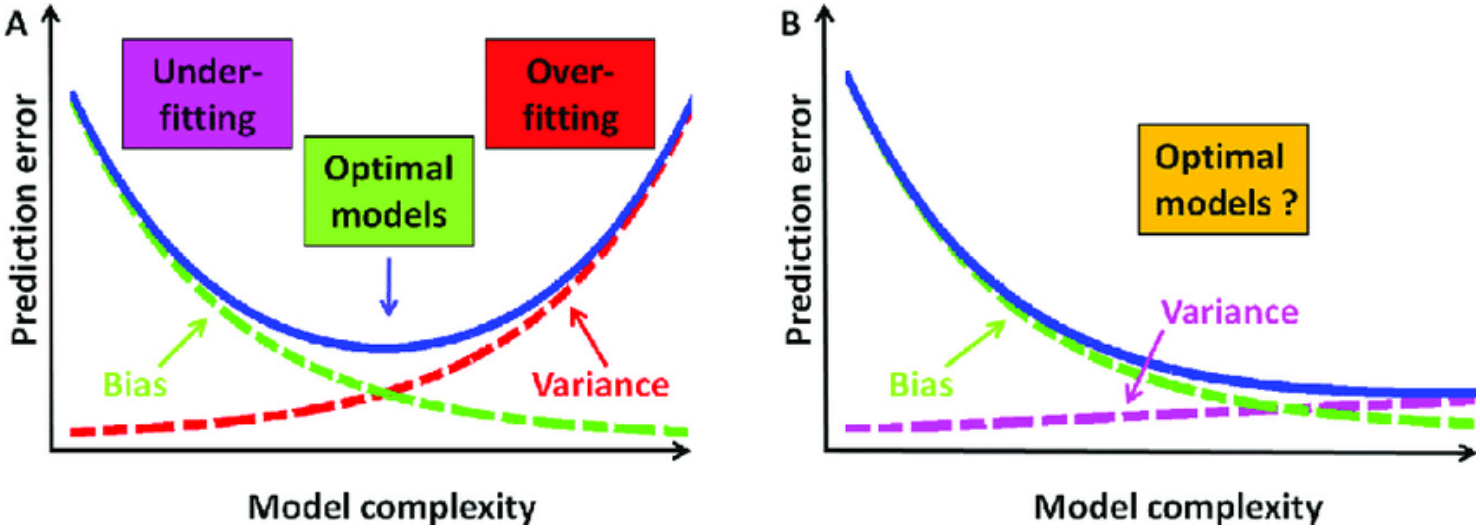
$$\text{Variance} = \frac{\sum(X-\bar{X})^2}{n}$$

- Calculate variance for each node.
- Calculate variance for each split as weighted average of each node variance.

Steps to a Decision Tree

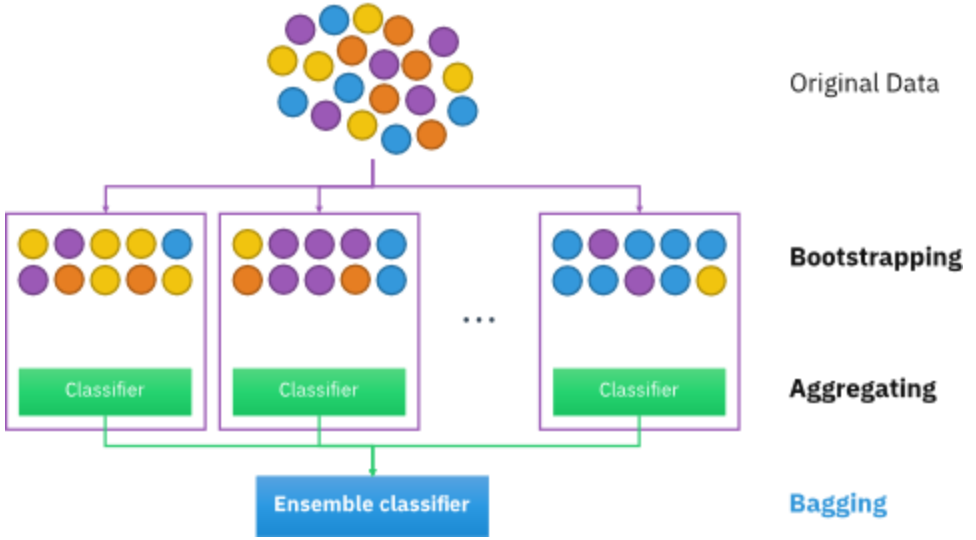
- We start at the tree root and split the data on the feature that results in the **smallest GINI** or the **largest information gain (**reduction in uncertainty towards the final decision).
- We can then repeat this splitting procedure at each child node **until the leaves are pure**. This means that the samples at each leaf node all belong to the same class.
- In practice, we may set a **limit on the depth of the tree to prevent overfitting**. We compromise on purity here somewhat as the final leaves may still have some impurity.

Bias/Variance Tradeoff



Ensembled Decision Trees

- Random Forest model is an **ensemble** of single decision trees (bagged decision trees)
- Leo Breiman - Statistical Modeling: The Two Cultures

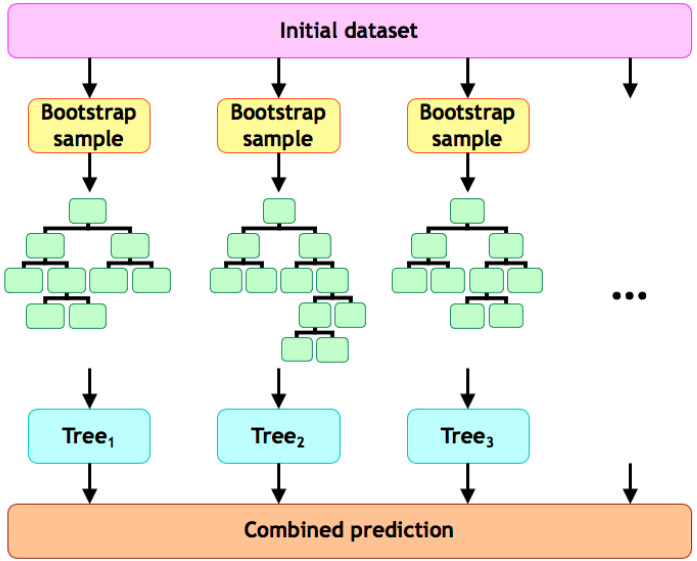


Bagged = bootstrapped aggregation

Ensembled Decision Trees

- Random Forest model is an **ensemble** of single decision trees (bagged decision trees)

- Leo Breiman - Statistical Modeling: The Two Cultures



Advantages

- RF models are robust to overfitting/bias issues
- There is no need in pre-selection of variables.
- RF has its own reliable procedure for estimation of predictive ability of model.
- RF allows for estimation of variable importance
- RF method is very fast and effective in working with large datasets

Geographically Weighted RF

- Geographical Random Forest (GRF) is a spatial analysis method using a local version of the Random Forest Regression Model.
- It allows for the investigation of the existence of spatial non-stationarity, in the relationship between a dependent and a set of independent variables. The latter is possible by fitting a sub-model for each observation in space, taking into account the neighbouring observations. This technique adopts the idea of the Geographically Weighted Regression, Kalogirou (2003).

Geographically Weighted RF

- The main difference between a tradition (linear) GWR and GRF is that we can model non-stationarity coupled with a flexible non-linear model which is very hard to overfit due to its bootstrapping nature, thus relaxing the assumptions of traditional Gaussian statistics.
- Essentially it was designed to be a bridge between machine learning and geographical models, combining inferential and explanatory power. Additionally, it is suited for datasets with numerous predictors, due to the robust nature of the random forest algorithm in high dimensionality. (Fotheringham et al. 2003)