**GIS AND SPATIAL ANALYSES**

LECTURE 1 – GIS concepts

GIS (Geographic information systems):

* Any system used for creating, storing, manipulating, analysing and presenting geographical information

## Geographic information:

* Any piece of data that can be located in space, using both
  + a set of coordinates
  + a known coordinate system.
* Set of coordinates can come from:
  + Vector data: coordinates of points/lines/polygons
    - A set of features containing either individual points or connected points forming lines/polygons
    - Needs a coord system. Cords are a precise location, may have precision/accuracy information
    - Features may have an attribute table – attributes of the data at the point
  + Raster data: eg. Grid data, satellite and aerial images
    - Image covering continuous surface, made up of individual pixels with a value. Can be **categorical** (classes eg land cover, species presence) or **continuous** (eg. Temperature, precipitation)
    - Needs origin and coord system so can locate individual pixels
    - Has a resolution (pixel size)

|  |  |
| --- | --- |
| **Raster** | **Vector** |
| Fixed grid with one value per pixel per band | Features with arbitrary shapes |
| Often multiple stacked bands, eg. Red, green, blue bands for an image |  |
| Attribute tables for **values** (VAT) | Attribute tables for **features** |

## Coordinate systems:

* Spherical coordinates: latitude and longitude.
  + Latitude is an angle above or below the equator. Angle between the same dist above and below equator are parallel. 1 degree of latitude at equator is a bit less than at poles.
  + Longitude is an angle around the equator, line is known as a meridian. Distance between 2 meridians changes: biggest at equator
  + Can include height. Example: 90.00E, 30.00N
* Ellipsoid coordinates
  + Earth is flatter at the poles. Diff coordinate systems have differing estimates for the radius at poles/equator, so important to know which you’re using
  + Airy 1830, Clarke 1866, International 1924, WGS 1984
  + Earth isn’t actually an ellipsoid either: surface is uneven and dynamic, and direction of gravity isn’t always exactly down. Geoid shows hypothetical surface of earth (approx. mean sea level) – perpendicular to the direction of gravity.
  + **WGS 1984** – combines data (ellipsoid measurements) and geoid to give coordinate system.
    - Uses modern satellite data to get data and geoid.
    - Used by GPS
    - Prime (Greenwich) meridian isn’t at 0 longitude – is slightly to the east based on geoid.

**Diagram

Description automatically generated**

## Different systems for different situations:

* Fit between geoid and datum varies in space – so countries adopt whichever coordinate system that fits the best locally
* Global system like WGS 1984 fit well on average
* **Eg Britain:** British National Grid uses **OSGB 36 datum** = shifted version of WGS 1984. Shift varies nationally
* Data shift from WGS 1984 varies depending on local system used by each country. Can be up to ~900m eg. Tokyo

## Projected coordinate systems

* Ellipsoidal geometry is iterative and slow
* Spherical geometry is exact and fast but globes aren’t convenient or easily scalable, or easily used on a flat surface
* Therefore need a flat representation of space
  + It is impossible to project a spherical surface onto a plane without introducing distortion
* Map projections can preserve (usually) only one of these things:
  + **Shape** – conformal maps
  + **Area** – equal-area maps
  + **Distance** – equi-distant maps
  + **Direction** – azimuthal maps
* Choice of map projection depends on the task: eg. For biodiversity comparison, equal area is likely most important to ensure fair comparison.
* Also, projection may only conserve from a focal point: eg. For distance
* **Tissot indicatrix:** circles on the surface of the Earth that show the distortion of ellipsoid surfaces on planar projections

**Examples of projected coordinate systems:**

* Often classified according to how they map from a sphere to a planar surface
* Equirectangular – one world maps use. Latitude and longitude as x and y coordinates. Shape and size is preserved well at equator and badly at poles
* Gnomonic - preserves bearings from a single central point. Good for looking at a very specific place of interest
* Cylindrical - preserves area, not shape**. Useful for biodiversity mapping**
* Mercator – preserves shape, not scale. Criticised as overestimates size of northern hem and underestimates Africa.
* Fuller Dymaxion – does local projections onto triangular planes. Arranged so majority of borders go through sea, and has no up and down. Tries to preserve both shape and scale, does it pretty well.

LECTURE 2: Remote sensing – a primary method of GIS data collection

* Used to have to map landscapes manually, which gives fine level of detail but its expensive, slow and inconsistent (hard to standardise)
* Remote sensing kicked off with aerial photography and then later satellite imagery

**Remote sensing:** acquisition of information about an object without making physical contact with it. More specifically, detecting and monitoring physical characteristics of an area by measuring reflected and emitted radiation

## Remote sensors:

* Passive – sense reflected solar radiation
* Active – emit radiation and sense reflection, eg. LIDAR (emits light) and RADAR (microwaves)
  + Detects alteration in reflected light
  + Trip time (time between sending out light and its return to sensor) gives heights

**Reflectance – relied on by both types**

* Albedo: proportion of radiation reflected from a surface, strongly affected by texture and angle of incidence
* Monochrome images: different objects have diff albedo (intensity of radiation). Can construct maps by looking at contrast, texture, edges
* Multispectral images: albedo of surfaces varies with wavelength. Compare bands recording reflectance in different wavelengths. Eg. RGB

## Satellites

* Bulk of this data is collected by satellite, although there are also LIDAR planes/drones
* Height of satellite orbit affects the data you get:
  + **High Earth** – geostationary orbits, meaning moves at same speed as earth rotation so looks down at same view the whole time
  + **Mid Earth** – used for navigation and communication – of less interest
  + **Low Earth** – often used for earth observation: looks at a small area that changes over time as earth rotates
  + **Sun synchronous** – whenever it’s looking at a particular place, it’s looking at the same time of day – incredibly useful for consistent data of a particular place

## Resolution

* Spatiotemporal
* Spectral

Spatiotemporal

* **Trade off between spatial and temporal resolution – closer to Earth means higher spatial but lower temporal resolution**
* Low earth orbits eg Pleiades:
  + Close to planet
  + High spatial resolution
  + Narrow path width, small scenes
  + Low temporal resolution – doesn’t look at particular place very often = less frequent images
    - Can get around this issue by using constellations of satellites – have multiple satellites on same orbit so will see same location at different times of the day
* High earth orbits eg Meteostat:
  + Far out in space
  + Low spatial resolution
  + No path width – always looks at same spot
  + Global scenes
  + Can take images constantly (but need to transmit them back to Earth)

Text

Description automatically generated with low confidenceSpectral

* Determined by the satellite’s mission
* Constrained by absorption of radiation by the atmosphere
* Light gathering sets the resolution and band width
* **Trade-off between spatial and spectral resolution:** eg. For ASTER satellite (diagram), in visible and near-infrared spectrum, has fewer wider bands = collects wider set of wavelengths = higher spatial resolution (15m) but lower spectral resolution. Bands 5-9 in short-wave infrared are opposite (30m)
* **See slides (lecture 2) for example satellites and their bands and resolutions**

## Using satellite images – steps required:

* Multiple steps needed before can use data:
  + Georeferencing – where is image
  + Orthorectification – remove perspective (eg might not be looking directly down) + terrain effects (may cause shading)
  + Calibration – convert sensor value to actual reflectance value (often a model does this)
  + Atmospheric correction – aerosols and water vaprou can impose spectral biases which vary on a daily basis (often model does this)

## Earth observation products

* Use satellite reflectance data to produce derived maps:
  + Use standard algorithms
  + Map land surface traits at global scale
  + Temporal scales: daily to annual
  + Resolution (spatial): 250m - >8km
  + Validation: many have pixel-by-pixel accuracy: how much trust to put in each individual pixel value, eg. If covered in clouds will have low validation

## Examples:

1. Vegetation indices
2. Digital elevation models
3. Fire signatures
4. Land cover
5. Productivity

### 1 – Vegetation indices

* Uses simple direct calc from sensor values (output of satellite)
* **Text

  Description automatically generated**Normalised Difference Vegetation Index (NVDI) uses near-infrared (NIR) and red (RED) bands:
* Enhanced Vegetation Index (EVI) adds more params to enhance vegetation mapping:

**A picture containing text

Description automatically generated**

### 2 – Digital elevation models

* 2 examples:
  + Shuttle Radar Topography Mission (SRTM)
  + ASTER Terra DEM
* Both near global with 30m resolution

### 3 – Fire signatures

* Live fires:
  + Spectral signature in infrared bands – look at reflection compared to usual, if higher then could be fire
  + Eg. MODIS, SPOT
* Burned area:
  + Added complexity – maps changes in burned area over time
  + Eg. MODIS

### 4 – Land cover

* Spectral signatures differ between different surfaces – each terrain type has a spectral signature that can be used to map land cover. Eg. Desert reflects strongly in all bands
* Spectral signatures differ over time – so need to measure how signatures differ for different habitats
* Can also do ground sampling, to measure reflectance of different habitats over time, which can be compared to satellite data to create overall land cover map
* Eg. MODIS, Global Forest Change, landcover.org/data

### 5 – Productivity

* V complex example
* Plants use light to store carbon. Can predict gross and net primary productivity if we know:
  + Amount of photosynthetically active light absorbed
  + Radiation conversion efficiency, given the temp and humidity
  + Respiration costs
* Need to create biological model of how light absorption is turned to biomass ie. Conversion efficiency, as well as for respiration

LECTURE 3 - Species Distribution Models

Different terms for SDMs:

* Species Distribution Modelling
* Climate Envelope Modelling
* Bioclimate Envelope Modelling
* Habitat Distribution Modelling
* Niche Modelling

Species Distribution Modelling definition:

* Interpolating biological survey data in space
* Creating quantitative predictive models of species/environment relationships
* Probabilistic model of how likely species is to occur in a particular area based on explanatory variables (eg. Temperature)
* Identifies areas in a landscape that have similar environments to localities where the species has been observed

## Overview of SDM process:

* Data on species occurrence in geographical space (presence and ideally absence data)
* Maps of environmental data – landscape properties (Eg elevation) and climate/weather properties (eg. Precipitation)
* Model to link occurrence to environmental explanatory variables (need to decide which env vars to use)
* GIS with which to produce map of predicted species occurrence
* Validation of predictions

## Why use an SDM

* To understand species distributions
* To predict occurrence of species for locations where survey data is lacking
* Can be used to:
  + Guide future field surveys
    - Eg Tibetan argali – used initial data to create SDM to infer stratified sampling method to try to improve data collection
  + Assess climate change impacts
    - Fate of diff species under diff climate scenarios
  + Predict invasive species spread
    - Eg predict where cane toad will spread next so can monitor these areas
  + Inform reserve selection
  + Inform our view of the past – examples slides 23-24
  + Compare drivers
    - Eg compare drivers of extinction and where they are most relevant and where they coincide
  + Predict future
    - Eg what might happen in arctic as climate change progresses

## Theoretical framework

* **Niche theory:** have abiotic, biotic and accessible niches, and core range is where these 3 overlap
* Individuals outside core range are sink populations – can only be maintained by those in core range migrating out (source pop)
* SDM aims to describe these patterns – in n dimensional parameter space (eg temp, precip…) wants to find the volume within which the species can survive + reproduce
* **Diagram

  Description automatically generated**Maps between geographical and environmental space.
* Lots of different model algorithms and approachesTable

  Description automatically generated

## Assessing predictive performance

* **Data type:** 
  + Presence/absence
  + Presence only
  + Presence/pseudo-absence
    - Pick locations where you think species isn’t
  + Presence/background
    - Probability of species occurring relative to it occurring elsewhere
* **Model evaluation – test accuracy:**
* Calibration data, then use the rest for test data to evaluate model and see how good your model is
* **Diagram

  Description automatically generated**Also project model: run in diff scenarios and see how well it translates
* **Model errors:**
  + Count up true positives and true negatives (where model agrees with data) and false positives and false negatives and put into **confusion matrix** (slide 35)
* **Probability to presence**
  + Probability of occurrence is what model outputs – if you want to translate this to predicted presence/absence, then need to decide on cut-off threshold of probability
  + Note don’t have to do this if not suited to the task

## Concerns and future directions

* Model choice matters:
  + Can project model into its occupied realm or onto whole world, and will likely give different answers
  + Some model approaches emphasise explaining, some predicting
    - Eg. If want to explain, then will only include important variables, simplest possible model
    - If want to predict, then include all variables, as gives maximum accuracy. Necessary for black box approaches eg. Neural networks where don’t know importance of each variable
  + Model performance varies
  + Model uncertainty – diff models give diff predictions
    - **Ensemble forecasting** aims to combat this by building many different models then superimposing them on each other to see what they agree with. Reduces uncertainties/inaccuracies
* Assumptions
  + Appropriate data exists, and at a relevant scale
  + Species are at an equilibrium with their environment ie. Assumes current range of species will persist long term and is stable
* Warnings
  + Garbage in, garbage out – data quality matters
  + Model transferability – feasibility of extrapolation across time/space
  + Most complex technology isn’t necessarily the correct one to use
* Future directions:
  + Incorporating dispersal
    - Simple models often assume either no dispersal or unlimited dispersal
  + Incorporating biotic interactions
    - Interactions between species
  + More mehanistic models
    - Shape of relationship between species presence and environment is constrained by model type, which more mechanistic models can address
    - Eg. If use thermal tolerance curves for temperature, see a sharp drop at higher temps that more simple models can’t predict
  + Connect to demography:
    - Ie connect to model that can predict if species can disperse to and persist – extinction risk - in the locations the SDM allocates as suitable
  + Community assembly
    - Use SDM as one component amongst other models in order to work out realised community assemblage
  + Bayes and spatial autocorrelation
    - More complex ways of fitting SDMs

LECTURE 4 – ASSESSING MODEL ACCURACY

## Confustion matrix

Table

Description automatically generated

## Accuracy

* All correct predictions / all predictions
* Bad measure to use
  + random models have 50% accuracy
  + Affected by prevalence (proportion of data that’s positive (eg. Presence) rather than negative (eg absence)

Table

Description automatically generated

## Rates of outcomes:

* Divide each outcome above by its respective observed count. Aka divide top tow ones by total number of positive observations

## Sensitivity

* Another name for true positive rate
* Proportion of correctly predicted positive observations

## Specificity

* Another name for true negative rate
* Proportion of correctly predicted negative observations

## Cohen’s kappa

* Measure of agreement that rescales accuracy to account for chance agreement
* Values: - infinity to 1, where 1 is perfect prediction and anything < 0 is worse than chance

**A picture containing text, clock

Description automatically generated**

* A = accuracy calculation
* Pe = chance agreement

**Calc Pe** – probability of positive outcome + prob of negative outcome where prob of an outcome is proportion of observed \* proportion of predicted (See below)

Table

Description automatically generated



A picture containing text

Description automatically generated



Graphical user interface, text, application, chat or text message

Description automatically generated

However, kappa is affected by prevalence, so not ideal.

## True skill statistic

**Text

Description automatically generated with medium confidence**

TSS is not affected by prevalence, but in specific use case of SDM, where low prevalence + large nums of observations, TSS converges on one or other of specificity or sensitivity, so less useful. See paper on slide 21 for alternatives

## Threshold model for probabilistic classification

* Probabilistic model predicts the probability of success/presence based on a variable (Eg temp)
* Chart

  Description automatically generatedNeed to set a threshold for success ie. Any value of the variable above the threshold is a presence and anything below is an absence

## ROC curve

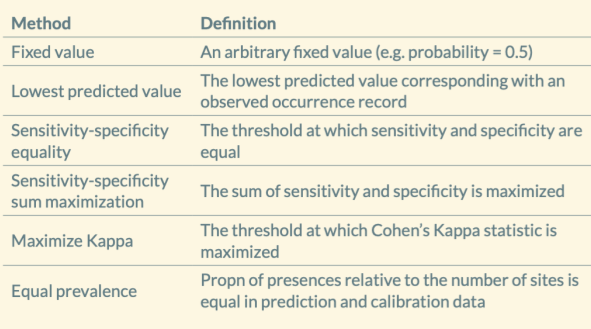
* Chart

  Description automatically generatedGives values of sensitivity and 1-specificity for different threshold values
* Chart, histogram

  Description automatically generated**Area under ROC (AUC)** – varies between 0 and 1, higher = better. Gives indication of overall model performance

## Threshold choices:

* Haven't covered calibration of model here - see prac.



<-common

Would be typical to do this using test data rather than calibration data

LECTURE 5 – Spatial modelling

= Incorporating spatial data into statistical models

Have spatial data on species richness and 3 explan vars (eg. Mean annual temp, elevation)

## Naïve models

* **Can fit naïve models using sp richness as response and explan vars as explan – these won’t predict the data well**
* Simple linear model
* Simple GAM (generalised additive model)
  + Replaces explan var with smoothed function of explan var

## Describing spatial autocorrelation

Spatial autocorrelation: similarity between data that’s close together in space

Ways to describe spatial autocorrelation:

* Neighbourhoods – rooks move, queens move, distance based, k nearest (slide 9-12). Can assign weights to neighbours (optional)
* Correlation matrix – see later (generalised least squares)

Global spatial autocorrelation measures (using neighbours)

* Global Moran’s I – strong corr at I = 1
* Global Geary’s C – strong corr at C = 0

Local autocorrelation

* Local indicators of spatial autocorrelation (LISA)
* Can tell you if some areas are more spatially autocorrelated than others
* Sensitive to region size you choose, as increase area, will approach original global measure of spat autocorr

## Effects of spatial autocorrelation in data

* Data points not independent
* Degrees of freedom reduced – effective sample size is smaller
  + Means standard errors and significance testing affected: larger errors and harder to find signif as lower effective degrees of freedom
* Not equally weighted
  + Parameter estimation affected

## Dealing with spatial autocorrelation

* Modify degrees of freedom in significance testing (only possible for simple tests)
* Account for autocorrelation in models:
  + Simultaneous autoregressive models
  + Generalised least squares
  + Eigenvector fitting
  + Geographically weighted regression

### Spatial autoregression

* + **Graphical user interface, application

    Description automatically generated**Estimate slope of model based on the value of each point but also weighted by the values of its neighbours (slide 20)

### Generalised least squares

* + Similar to linear models but points are allowed to be correlated - have a function to describe it
  + Uses correlograms/variograms (see diagrams below)
  + Models correlation as a function of distance ie. Find a curve that fits well to correlogram/variogram
  + Diagram

    Description automatically generatedChart, line chart

    Description automatically generatedgenerate a correlation matrix - Correlation between all pairs of values based on model, which can feed into generalised least squares model
  + **Different shapes possible to describe spatial autocorrelation:**
    - Gaussian
    - Exponential
    - Spherical
    - Linear
  + **Chart

    Description automatically generated**Example of Gaussian below:
  + **Parameters**:

Nugget = how much correlation is there at 0 distance (minimum correlation)

Range = at what distance the model approaches maximum variance

Sill = asymptotic value of variance

See slide 26 for stats output of model (similar to lm)

**Assumptions of spatial autoregression and generalised least squares:**

* Stationarity – assumes same process is happening across all the different locations
* Isotropy – assumes same process is happening across all directions
* Can solve problem using below methods:

### Eigenvector filtering

* + Take eigendecomposition of a spatial weights model
    - Eigendecomposition = Way of decomposing a matrix of data to find key axes of variation
  + Use eigenvectors (key axes of variation) as variables in the model to describe the spatial structure
    - Each describes a particular pattern in the data
  + Use a selection process to identify important eigenvectors to include – each one uses a df so only want essential ones
    - Put a bunch into a basic linear model then take out the ones with non-signif p-vals

**Text

Description automatically generated with low confidence**

* + **Difficulties**:
* hard to describe what spatial structure is - just putting eigenvectors in to soak up spatial structure. Should look at residual autocorr to see if theyre doing a good job
* Generates lots of Evs, hard to decide which ones to include as no set rules. There are functions in R that will compare lots of them and their effect but take a long time to run

### Geographically weighted regression

* + Controversial technique - good way to explore data but need to be careful about taking results at face value
  + Fit a model for every cell:
    - Define local region size and a weighting function and fit a weighted regression for each cell
  + Look how coefficients vary in space – shows how effect of each explanatory variable on the response variable varies throughout space
  + Also shows explanatory power of the model throughout space

## Problems with spatial modelling

* Large number of overlapping packages with different data structures required
* Sometimes poor documentation in these packages – either highly mathematical or incomplete
* Often very slow to calculate
  + Speed of calculation decreases as a square of size of dataset – as dealing with large matrices
* Memory hungry (large matrices)
* Too many options – hard to know when you’ve found the ‘right’ model