

Public Health Summer School Wellington



University
of Otago
ŌTĀKOU WHAKAIHU WAKA

Pōneke
Wellington
Campus

GIS techniques in health research and policy — DAY 2

Te whakamahi GIS (Pūnaha Mōhiohio
Matawhenua) i roto i te rangahau
hauora

Day 2, 17 February 2026



Programme

Time	Session
9.00 am	Introductions, purpose of GIS
9.30 am	S1: Generating covariates – Tutorial 3
10.30 am	<i>Morning break</i>
11.00 am	S2: Point and area pattern analysis
11.30 pm	Tutorial 4 – Kernel density
12.30 pm	<i>Lunch break</i>
1.30 pm	S3: Interpolation and spatial autocorrelation
2.15 pm	Tutorial 5 – Spatial autocorrelation
3.00 pm	<i>Afternoon break</i>
3.30 pm	S4: Considerations for regression modeling
4.15 pm	AI and geospatial developments
4.45 pm	Discussion/questions, your own work
5.00 pm	Finish

Welcome announcements

- Useful information
 - Orientation – bathrooms, lunch & refreshments
- Keeping to time
- Evaluation forms
- Certificates of attendance
- Special room requirements - computer lab
- Introductions

Emergency Evacuation

Mein Street

Link Block Levels 4, 5, 6, 7, 8 and 9 (D, E, F, G, H & J) through Vibe (The Health Deli) Café.

WSB levels 5,7,8 and 9 (E, G & J) Deli) Café.

Lecture theatre (Nordmeyer)

Ambulance

Ask Otago and Level 3 (C) link carpark

Exit through to WSB, use the stairs down to level 3 and exit

Use the stairs down to level 3 and exit through Vibe (The Health

Exit down the stairs and use the fire exit opening out to the exterior of the building and report to carpark near bay.

Exit through the main entrance by Askotago and report to near Ambulance bay.

Prefabs

Exit through doors and gather in the main car park at front of building.

Hanson Street

Exit through doors and gather on street at front of building.

Considering postgraduate study?

Talk to us about the options available at our 3 campuses

Wellington | Dunedin | Christchurch

- Certificate in Public Health
- Diploma in Public Health
- Master of Public Health
- PhD



To find out more, ask at the registration desk, pick up a brochure or visit otago.ac.nz/publichealth

Access UO_GUEST

UO_GUEST is the network available for use by visitors to the University of Otago who do not have eduroam access, and provides internet access.

How do I access the UO_GUEST network

To access the UO_GUEST network:

1. Select the **UO_GUEST** network from the list of available networks on your device. A webpage should open automatically with the login screen. If not, open a web browser manually and the page should open.
2. Click on the *Don't Have an Account?* link on the login screen and follow the prompts to create a new account. If you already have an existing UO_GUEST account, enter your login details here. You will be required to provide a working email address and contact phone number.
3. Your UO_GUEST credentials will be emailed to the account provided. You will need to be able to access your email to collect your password. Once you have received your username and password, click the *Sign On* button.
4. You are required to read and accept the University network regulations. You will be unable to access the internet until this is done.
5. After you click *Accept* you will be redirected to the ITS Visitor Network Access page, and you should now be able to access the internet.

Welcome! Haere mai!

- Goals for today:
 - To understand which spatial tools and techniques are available for addressing epidemiological questions, health care service decisions, etc.
 - To apply some spatial tools using data
 - To begin to think of ways of applying these tools and spatial thinking to the work you already do
 - To provide resources for further learning

Primary purposes of GIS

- Record and store information DAY 1
- Present information and communicate findings DAY 1
- Generate geographic covariates
- Analyse locational distributions and spatial patterns

Generation of geographic covariates

- Key concept: where we live affects our lives and our health
- We often extract information from GIS to be used in other software programs for statistical analyses (eg, regression in Stata or R)
- This gives us information about places or spatial patterns
- Helps us characterise exposures
- Helps us account for non-independence in our observations

Analyse location and patterns

- 1854 - London
- Cholera outbreak
- Unknown source
- Mapped locations of patients & water pumps
- Intervention



Generation of geographic covariates

- As epidemiologists or public health practitioners,
 - We often think about exposures at the individual-level (eg, genetics, behaviour)
 - GIS helps us assess exposure at a larger scale (eg, air pollution) and across a population, which can be assigned to individuals

Generation of geographic covariates

- As epidemiologists or public health practitioners
 - We often think about exposures at the individual-level (eg, genetics, behaviour)
 - GIS helps us assess exposure for individuals at larger scales (eg, air pollution)
 - Based on where they live, work, attend school
 - Based on where they spend their time (GPS tracking data)

Generation of geographic covariates

- Generate variables which describe:
 - Neighbourhood context
 - Work environment
 - Exposure to environments (built, physical, social)
- Variables are then used to:
 - Assess exposure (assign to individuals or to areas)
 - Assess health (usually to areas)
- Variables used in regression analyses, for example:
 - Land use regression to predict air pollution
 - Access to greenspace and depression
 - Neighbourhood walkability and CVD

Example: Neighborhood conditions and the microbiome

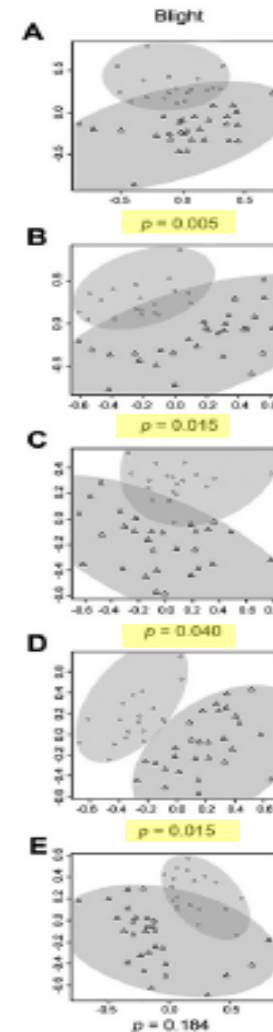
- Little is known about the potential influence of neighborhood environments on the bacterial & archaea communities that live in and on the human body -- the microbiome
- Unique opportunity – partnership with Medical Examiner's Office in Wayne County (Detroit)
- In a city with declining population, vacant lots, abandoned buildings, but also grassroots 'greening' programs

Example: Neighborhood conditions and the microbiome



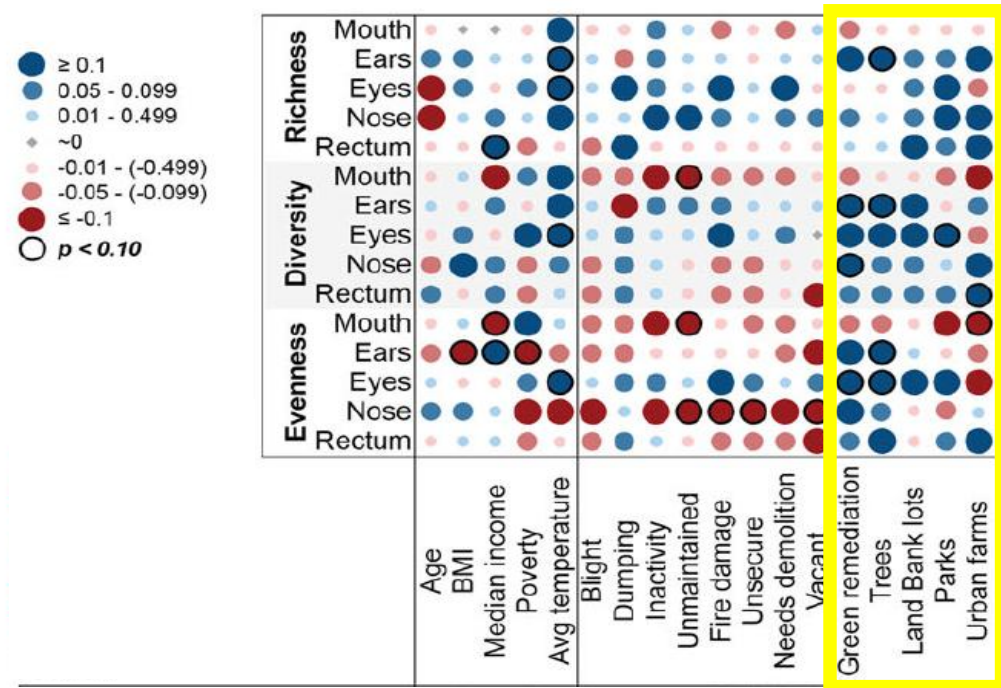
Example: Neighborhood conditions and the microbiome

- Significant clustering of composition (genera) of the microbiome for all anatomical regions except the rectum



Example: Neighborhood conditions and the microbiome

- Microbial biodiversity was significantly and positively correlated with green remediation



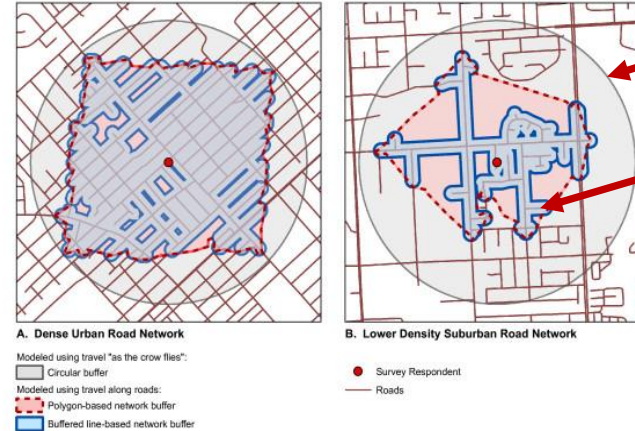
Section 1: Generating spatial covariates

Making geographic calculations

- We want to know the number of cancer cases within 200 metres and 500 metres of green space in the Hamilton. We also want to know the distance from cases' home locations to the nearest oncology clinic.
- Say we got the data for cancer cases from the New Zealand Health Survey, green space data from aerial imagery and we geocoded the addresses of clinic locations.
- You could make these calculations using spatial tools.

Buffer

- Boundaries placed around areas (polygons), lines or points
- May use a centroid or point as centre
- Set radius
- Then euclidean or network buffering technique used to assign buffer at given distance from centre point



Source: Oliver et al. *International Journal of Health Geographics* **2007** 6:41.

Making geographic calculations

- We want to know the number of cancer cases living within 200 metres and 500 metres of green spaces in Auckland.
 - Make buffers around green spaces.
 - Join the case locations (points) to the greenspaces. Use a *sum function* within this tool.
 - We end up with the sum of cases within 200 metres and 500 metres of each greenspace.

Making geographic calculations

- Other useful calculations:
 - Calculate the area of a polygon
 - Proportion of a neighbourhood occupied by greenspace
 - Calculate distance from point to a point or polygon
 - Distance from home to nearest oncology clinic
 - Calculate number of SA1s within each SA2
- You can select units (m, mile, km)
- Many tools available – cannot cover all of them

Making geographic calculations

- Before measuring any distances or calculating scale, ArcPRO must be configured to report measurements in units of the user's choice
- ArcPRO must be told what units the data are stored in
- Without knowing what the units are, it is impossible to calculate scale or to perform any measurements. **For this reason, you should *always* know the projection and coordinate system of any datasets you are using.**

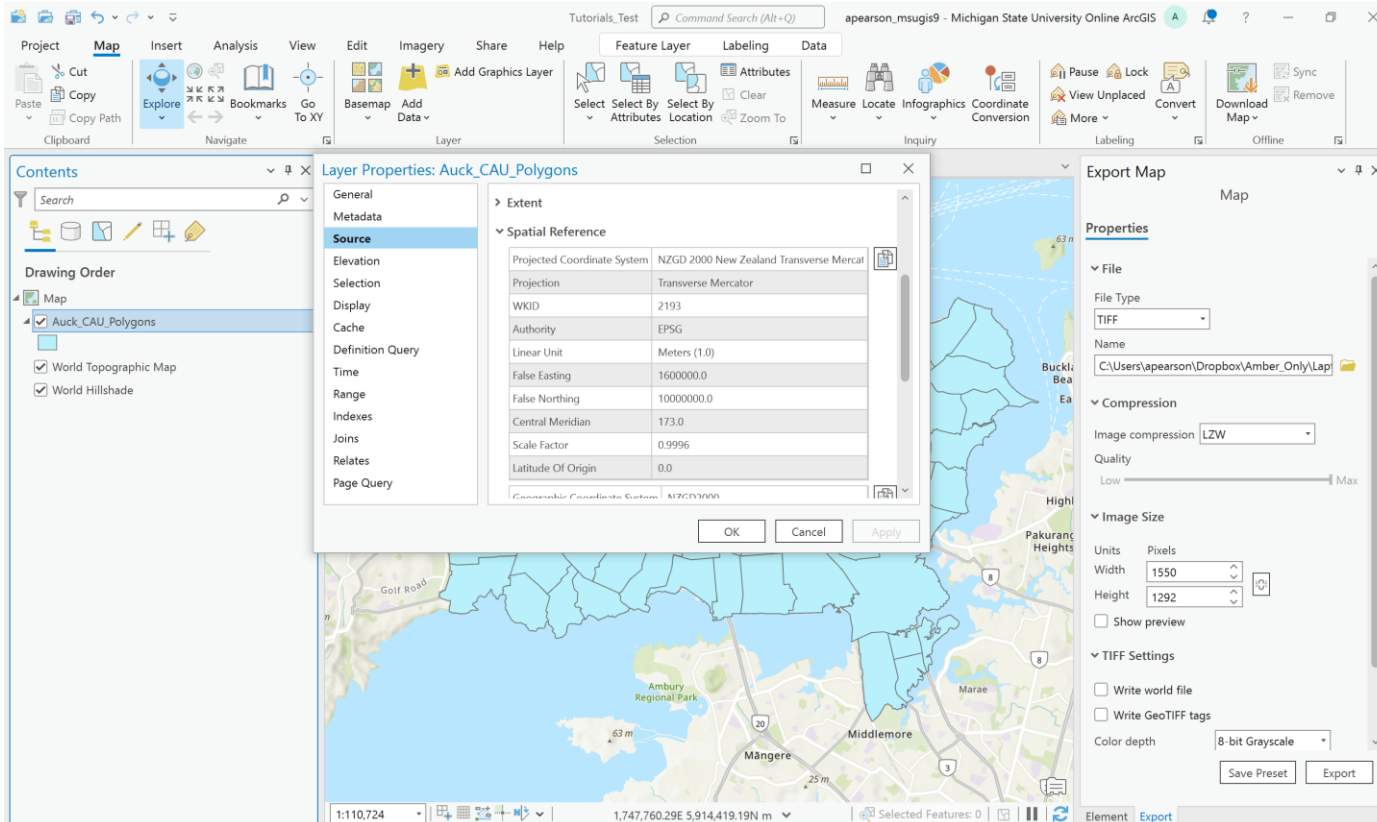
Coordinate systems

- **Geographic Coordinate System** - uses an angular unit of measure (e.g. latitude and longitude), a prime meridian and a datum based on a spheroid
- **Projected Coordinate Systems** converts the latitude and longitude that fall on the 3-D surface of the earth onto a 2-D surface for use in GIS (important for distances)
- In NZ, the two most common projections are :
 - **New Zealand Map Grid**
 - **New Zealand Transverse Mercator**

Making geographic calculations

- Checking/setting coordinate system
- You can always right click a dataset title in the table of contents
- **Properties**
 - **Source → Spatial reference**
 - Predefined > Geographic Coordinate Systems > World WGS 1984
- *You do not need to do this for any tutorial in this course*

Projection used in tutorials



Making geographic calculations

- You can make calculations on an **un-projected** dataset – but less realistic without curvature of the earth
- Still, all projections distort map properties (which is why there are so many; to reduce bias in some places, bias in other areas are increased).

Making geographic calculations

- We are not covering all:
 - Projection systems
 - Transformations
 - Bias
- In these tutorials, all the data are in the same projected coordinate system – NZ Transverse Mercator

Special considerations – health or sensitive data

- When working with health data, we may be required by data providers (and ethics) to consider:
 - Preserving anonymity of survey participants
 - Rarity of event (infantile astrocytoma)
 - Stigmatism of places or population groups
- We can effectively and ethically do this:
 - Jitter individual point locations
 - Aggregate to larger population/geographic unit

Jittering

- Lots of different options (refer to course book)
 - Paul A. Zandbergen, “Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data,” *Advances in Medicine*, vol. 2014, Article ID 567049, 14 pages, 2014. doi:10.1155/2014/567049
- Complex algorithms to shift point locations but maintain similar pattern
- More important if then using points for analysis or sharing
- Simple method available – **BONUS tutorial 3**
 - Should **not** use new locations for spatial analysis

Tutorial 3

- You will be learning a few tools now:
 - Buffer
 - Sum points within polygon
 - Distance calculation
- BONUS MATERIAL: jittering sensitive data

Tutorial 3 – buffers, sum points and distance

A. What might be the scientific significance of using different buffer sizes?

For discussion

Tutorial 3 – buffers, sum points and distance

B. For any given park (greenspace), what is the highest number of obesity cases within 200 metres, 500 metres?

C. How close is the nearest food outlet, of all the obesity cases? How far is the greatest distance?

D. Why would a health researcher be interested in knowing this information? Would this be an exposure or an outcome and why (provide a scenario)?

Tutorial 3 – buffers, sum points and distance

B. For any given park (greenspace), what is the highest number of obesity cases within 200 metres, 500 metres?

200m=5, 500m=15

C. How close is the nearest food outlet, of all the obesity cases? How far is the greatest distance?

0.0003km, 264.151km

D. Why would a health researcher be interested in knowing this information? Would this be an exposure or an outcome and why (provide a scenario)?

Morning break

Public Health
Summer School
Wellington



Lake Michigan

Section 2: Point and area pattern analysis

Point and area pattern analysis

- Geographers use points and areas (or polygons) to examine patterns
- Can be done descriptively (e.g., mean center of points)
- Or inferentially

Statistical pattern recognition

- Involves calculating a statistic (e.g. spatial cluster statistic, autocorrelation statistic, etc.) that quantifies a relevant aspect of spatial pattern data.
- This might be
 - a health outcome (e.g. case/control locations, incidence or mortality rate)
 - exposures (e.g. exposure to agricultural pesticides)
 - risk factors (e.g. smoking prevalence)
- The numerical value of this statistic is then compared to the distribution of that statistic's value under a null spatial model.
- This provides a probabilistic assessment of how unlikely an observed spatial pattern is under the null hypothesis (p-value).

Clustering

Visualisation, descriptive recognition

- Map display - eye-balling technique, Kernel density functions

Statistical pattern recognition – most common method

- Statistics quantifying an observed spatial pattern – compare the expected and observed spatial distribution under a null spatial model
- Provides probabilistic assessment that an observed spatial pattern is not just random chance variation from the expected pattern

Clustering

- Event-based (point locations of cases)
- Population-based (rate of cases/population at risk, usually within polygons)
- Field-based (continuous distribution over space eg, air pollutant highs, usually raster)

Applications

- Disease surveillance
- Exposure assessment
- Crime hotspots/coldspots
- Traffic crashes
- Zoonotic infections – hotspots of contact

Defining nearest neighbours

- Many approaches
- Nearest based on adjacency
- Nearest based on a set distance (within 5 miles)
- Nearest k neighbours (useful when working across rural and urban)
- Use spatial weights to indicate strength of 'neighbourliness'
 - Inverse-distance
 - Inverse distance squared

Nearest neighbour analysis

- Compare each distance to nearest neighbour to the mean for all points.
- Compare the mean observed distance to some expected mean distance (eg the distance expected for a random distribution)

TABLE 14.1
Coordinates and Nearest Neighbor Information for Example*

Point	X	Y	NN	NND
A	1.3	0.9	E	3.94
B	3.2	4.4	C	2.00
C	3.3	6.4	B	2.00
D	5.6	3.8	E	1.36
E	4.8	2.7	D	1.36
F	8.1	7.4	G	4.21
G	9.4	3.4	D	3.82
SUM				18.69

where NN = nearest neighbor
NND = nearest neighbor distance

$$\overline{NND} = \frac{\sum NND}{n} = \frac{18.69}{7} = 2.67$$

* See figure 14.1 for graph of point locations.

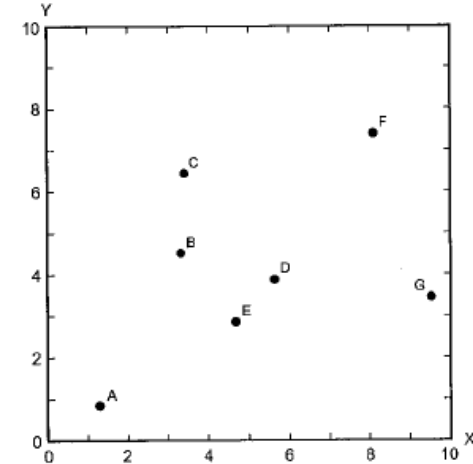
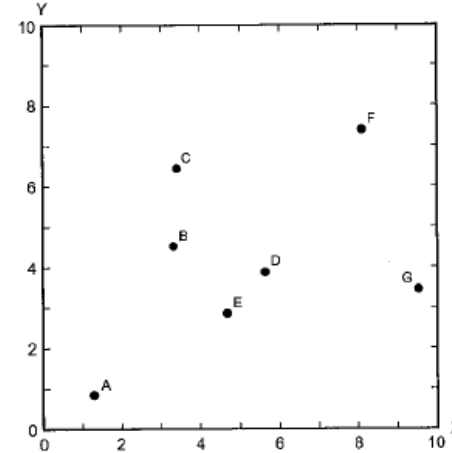


FIGURE 14.1
Location of Points for Nearest Neighbor Problem

Nearest neighbour analysis

- This value of 2.67 must be compared to a theoretical point distribution.
 - Perfectly (or purely) dispersed
 - Perfectly random
 - Perfectly clustered



- If this example was perfectly random (n= 7 points)...

$$\overline{NND}_r = \frac{1}{2\sqrt{Density}} = \frac{1}{2\sqrt{0.07}} = 1.89$$

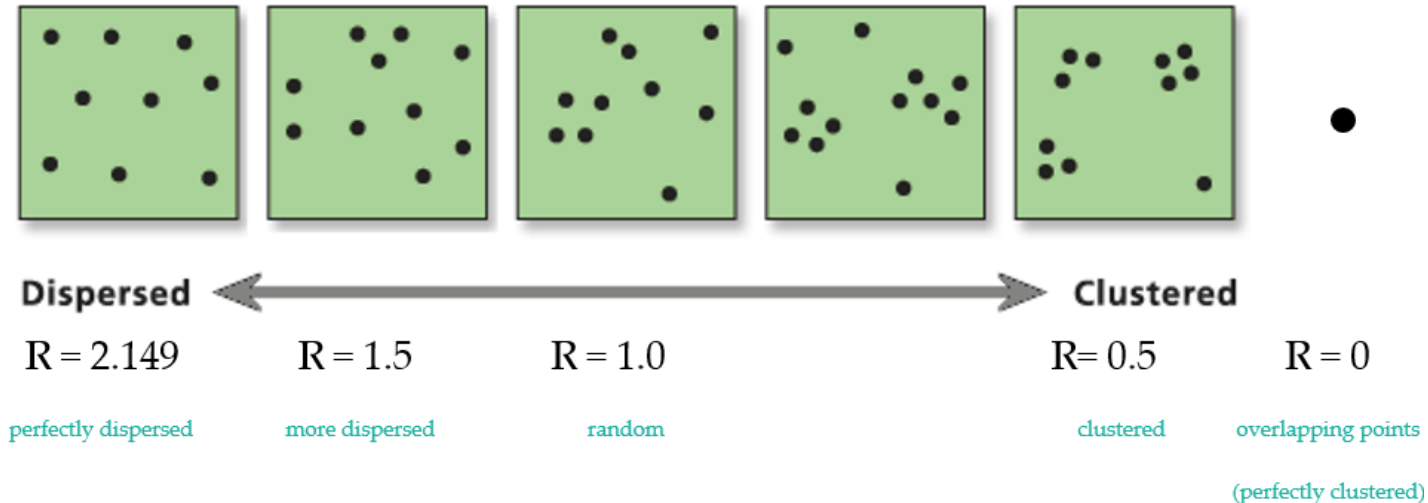
Note: 10 x 10 units = 100
7/100 = 0.07

- But this value will vary based on the number of points, etc
- So, we calculate a standardized nearest neighbour index (R)

Nearest neighbour analysis

$$R = \frac{\overline{NND}}{NND_r} = \frac{2.67}{1.89} = 1.41$$

We compare this value to a continuum.



Nearest neighbour analysis

We can also use the R value to test the null hypothesis that the observed nearest neighbour pattern is equal to a random distribution.

To do this, we calculate:

$$Z = \frac{NND - NND_r}{\sigma NND} \quad \text{and} \quad \sigma NND = \frac{0.26136}{\sqrt{n(Density)}}$$

$$Z = \frac{2.67 - 1.89}{\sigma NND} \quad \text{and} \quad \sigma NND = \frac{0.26136}{\sqrt{7(0.07)}}$$

$$Z = \frac{0.78}{0.373} = 2.09$$

Then, we can use a z-table to find the corresponding p-value for this z-statistic.

$$Z = \frac{0.78}{0.373} = 2.09$$

$$p = 0.018$$

However, we need to consider whether we have a one-tailed or two-tailed hypothesis.

Nearest neighbour analysis – one-tailed vs two-tailed

- Two-tailed hypothesis
 - We think our distribution is not random. So we simply test the null hypothesis that the observed pattern = a random pattern
- One-tailed hypothesis
 - We think our distribution is clustered (or dispersed).
 - Think about the tails on the continuum.

Nearest Neighbor Analysis	
Primary Objective:	Determine whether a random (Poisson) process has generated a point pattern
Requirements and Assumptions:	
1. Random sample of points from a population	
2. Sample points are independently selected	
Hypotheses:	
$H_0 : \overline{NND} = \overline{NND}_R$	(point pattern is random)
$H_A : \overline{NND} \neq \overline{NND}_R$	(point pattern is not random)
$H_A : \overline{NND} > \overline{NND}_R$	(point pattern is more dispersed than random)
$H_A : \overline{NND} < \overline{NND}_R$	(point pattern is more clustered than random)
Test Statistic:	
$Z_R = \frac{\overline{NND} - \overline{NND}_R}{\sigma_{NND}}$	

Since we did not have a clear theory about the hypothesized distribution a priori, we use a two-tailed test.

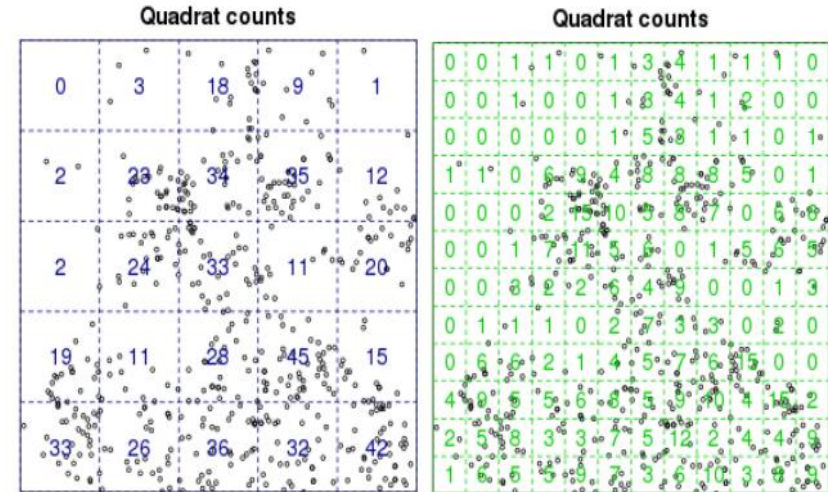
$$p = 0.018 * 2 = 0.036$$

So we conclude that the observed distribution is significantly different from a random distribution.

Nearest neighbour analysis

- Can also use quadrants superimposed on points to assess patterns
- Which compares variability of points within each cell
- This approach can lead to issues (e.g., if cells are different sizes)

Quadrat Mapping



Quadrant maps

- Classic approach
 - analogy to employment of sampling frames in field surveys
Partition study region into equal-area subregions called quadrants
- Approach 1: completely tile the study area with non-overlapping quadrants
- Approach 2: randomly position n quadrants in the region
 - Count number of points in each subregion; divide by quadrant area to obtain frequency counts
 - Summarizes the distribution - sort of a 2-D histogram

Quadrant maps

- Advantages:
 - Point pattern essentially transformed to areal data
 - Many analysis techniques exist for such data
- Disadvantages:
 - Much information is lost
 - No obviously ideal quadrant size

Density

- A measure of intensity – exposure to features of the built environment
- Relates to accessibility
 - Sum (count) of features within specified area (buffer) (e.g. fast food outlet per km²)
 - Count per population/geographic area
 - Kernel density
- Continuous measures assume: linear relationship between feature and health outcome

Kernel density

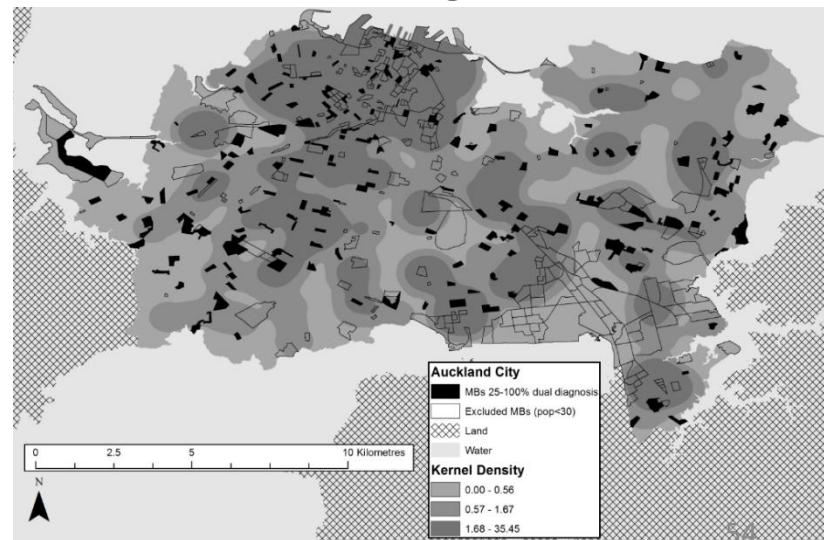
- An extension of the quadrant approach
- Idea: Pass a moving window (quadrant) over the region. Assign weights to points falling within the window according to how close to the center of the window they are
- Theory: Intensity estimation of spatial point pattern is similar to estimating a bivariate probability density
- Must have population data, not sample data

Kernel density

$$p(x) \cong \frac{k}{NV} \text{ where } \begin{cases} V & \text{volume surrounding } x \\ N & \text{total \#examples} \\ k & \text{\#examples inside } V \end{cases}$$

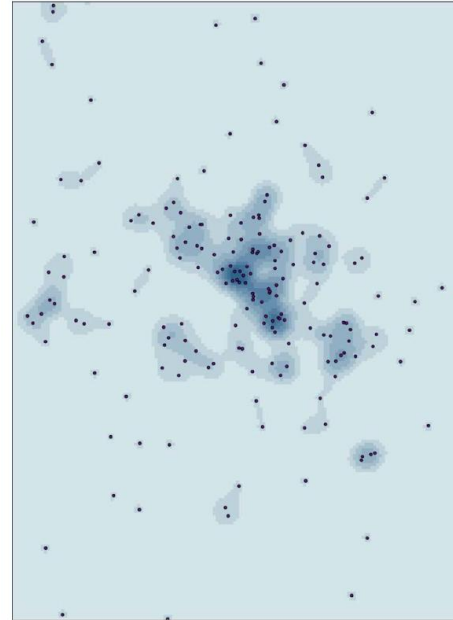
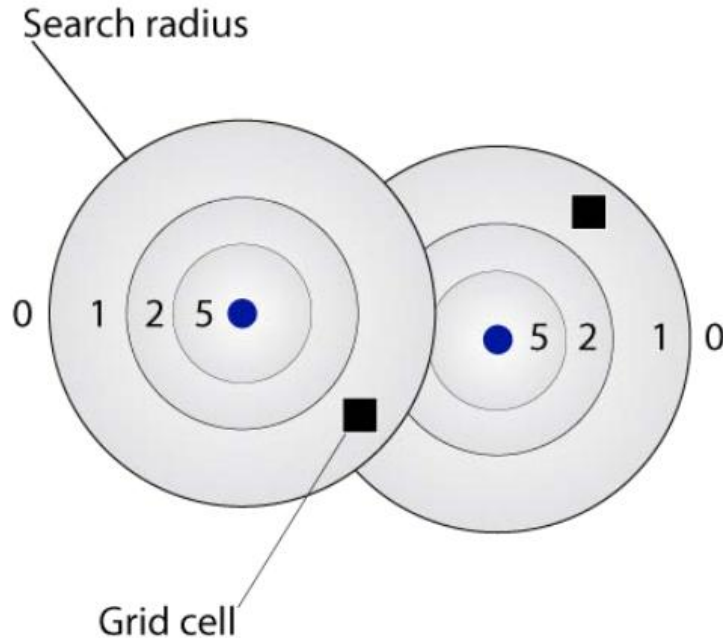
- Similar to histogram
 - Bin width and starting point = limitations
- Kernel density overcomes this ‘bumpiness’
 - But is still subject to width issues - Relies on defining the interpolation interval (1km)

Thus, a zero indicates that there were no outlets within one kilometre.



Kernel density

- Calculates a magnitude per unit area from point [or line] to fit a smooth surface

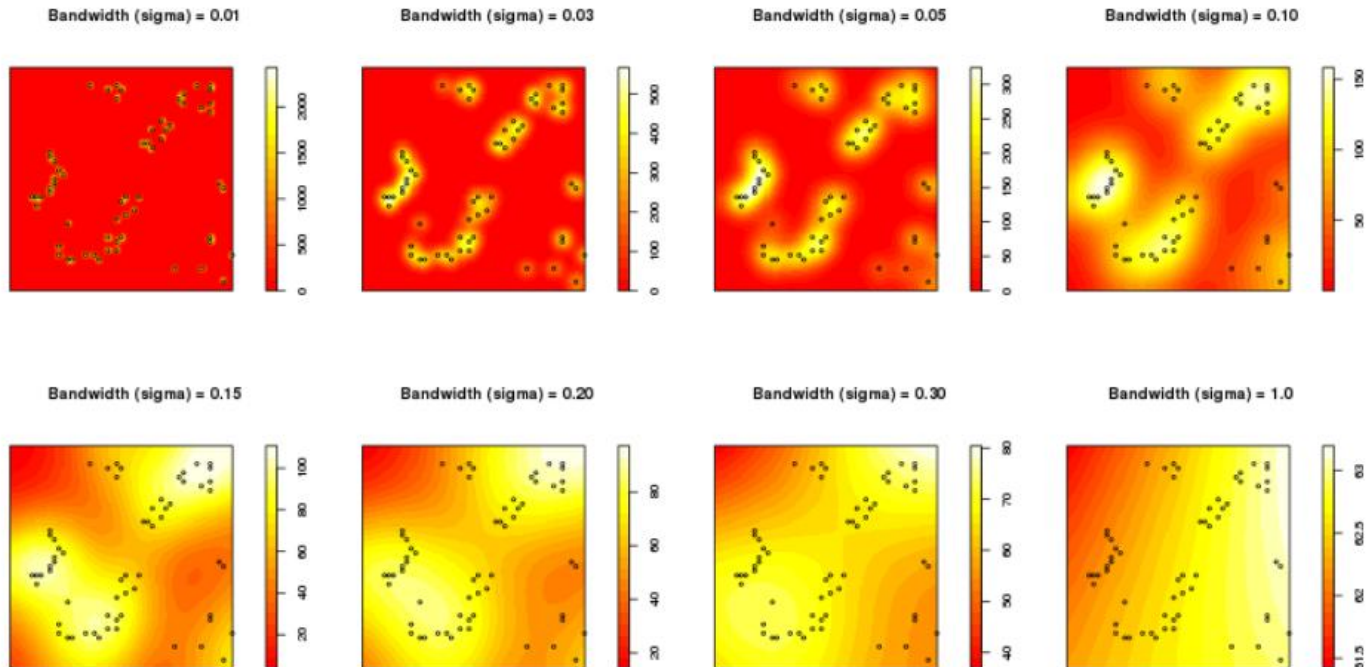


Kernel density

- The bandwidth determines the radius, and is critical to the performance of the kernel
- Setting large bandwidth lessens influence of nearby points on kernel
- Setting small bandwidth gives poor notion of overall intensity

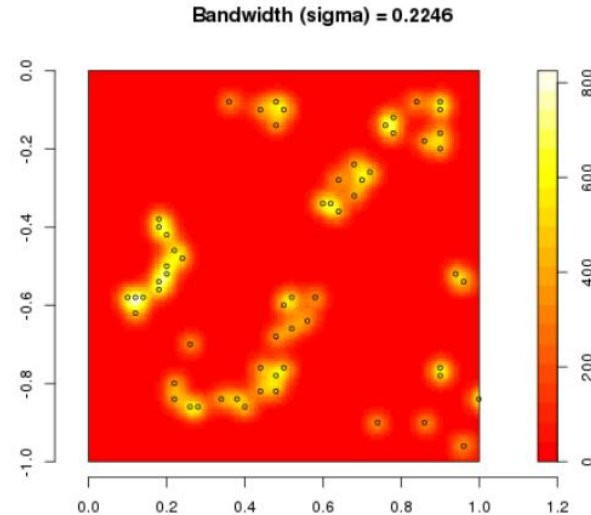
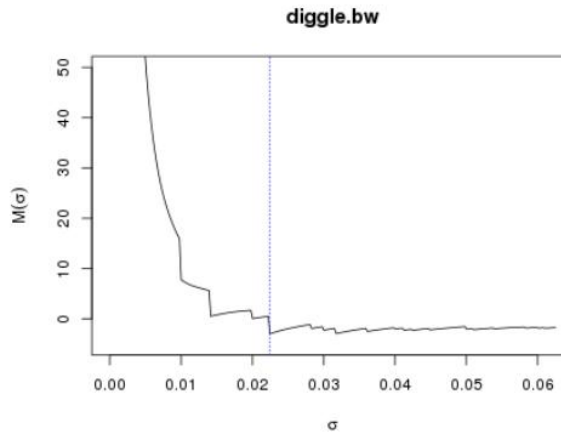
Kernel Maps

Changing the bandwidth (here, sigma for a Gaussian kernel) has a major effect on the kernel.



Kernel density

- Optimal bandwidth minimizes mean squared error (using criteria by Peter Diggle)
- Output is a smoothed, continuous [raster] surface of the phenomenon of interest



Pattern analysis

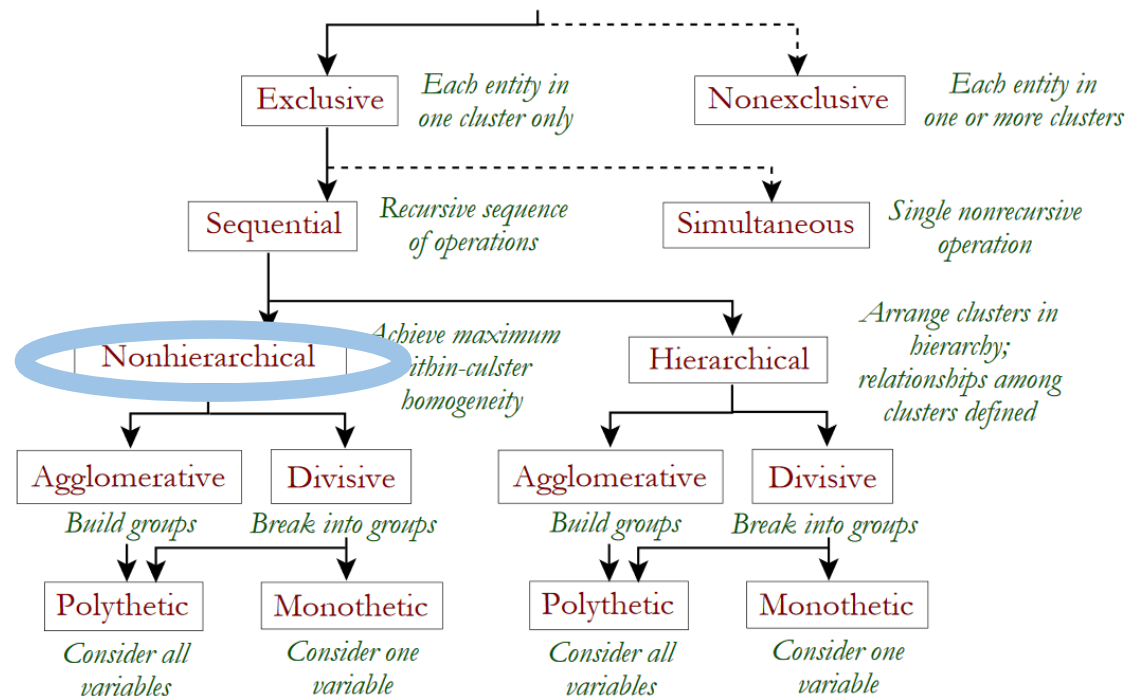
- Suppose we want to understand patterns in the rates of anxiety across CAUs in Northland
- We could use area pattern analysis to determine if there appears to be area-level exposures leading to high levels of anxiety
- Or whether there is a perfectly even distribution of anxiety across the TLA

Clustering

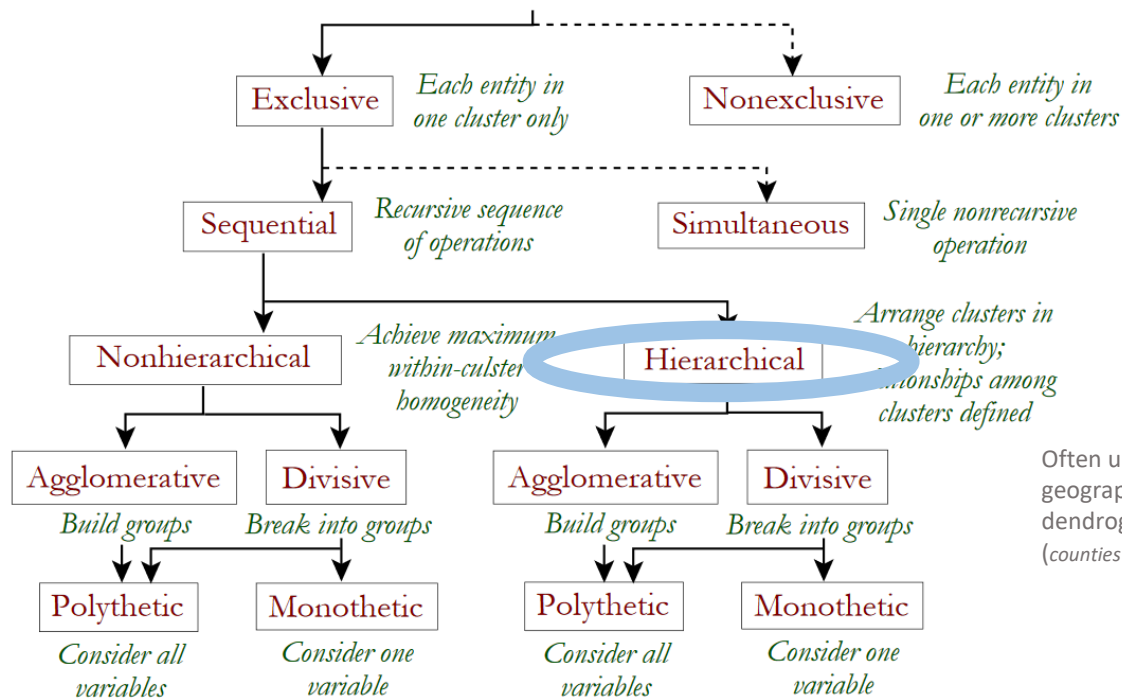
- Often a way of reducing data
- Grouping – by characteristics and/or location
- Maximize within group similarity, and among group similarity minimized
- Reduce information on the whole set of N observations to information about k groups of nearly similar entities (where hopefully $k < N$)

Cluster Techniques

Maximizes within-cluster homogeneity, thus inference based on cluster membership



Cluster Techniques



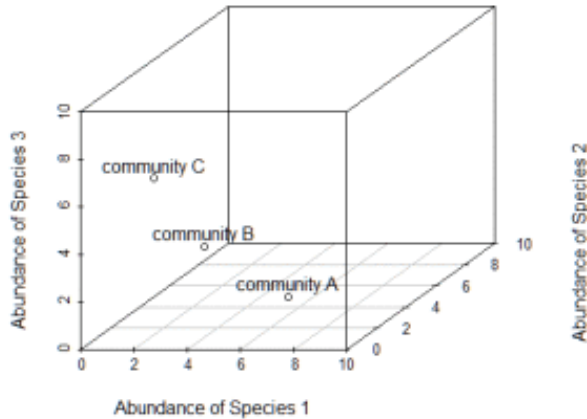
Often used in
geographic analyses –
dendrogram
(counties within states)

Clustering

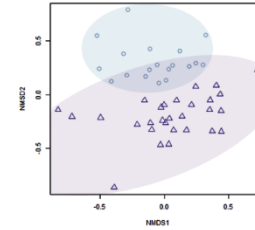
- Identify outliers – by leaving them in solitary or small clusters; possible remove from further analyses
- Eliminate noise – by clustering similar entities without exact similarity
- So many methods....
 - Maximum likelihood clustering
 - Minimum variance partitioning
 - Composite clustering
 - K-means clustering
- Often must define number of allowable clusters (cut-off)

Clustering of non-spatial data

- Ecology

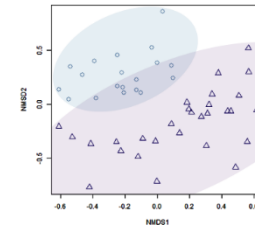


A. Buccal

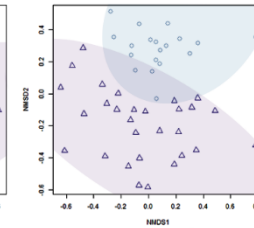


DECAY

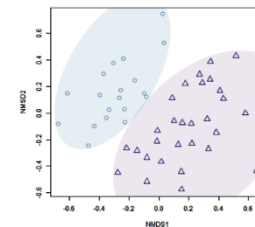
B. Ears



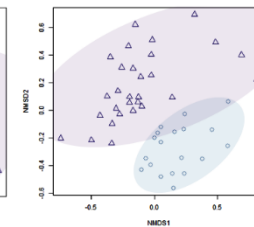
C. Eyes



D. Nares

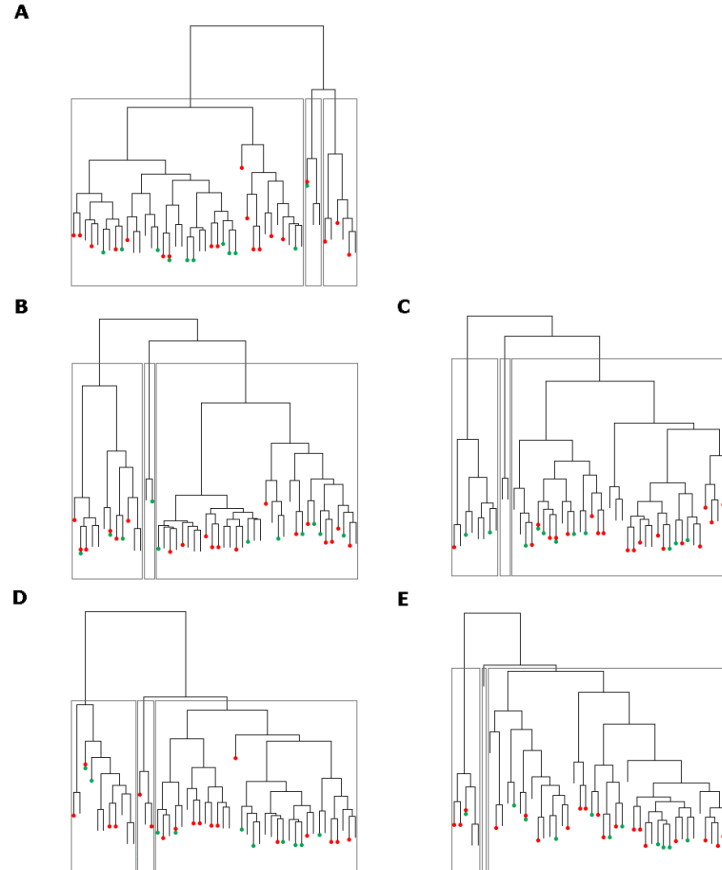


E. Rectum



Hierarchical clustering in non-spatial data

- Dendrograms



Cluster analysis of spatial data

- Goal of cluster detection is to uncover interesting patterns in multivariate data space and to classify observations into groups based on their internal similarity or external distinction
- One area of similarity/dissimilarity is spatial location
- Lots of techniques have been developed; many can be grouped (clustered?!) into various sorts:
 - Hierarchical Methods: begin with one large group and iteratively divide this down into subgroups, or alternatively begin with each observation in its own group and iteratively merge them up. Either approach results in a multiple-scale classification tree
 - Partitioning Methods: assign objects into a particular number (k) of groups
- These approaches typically consider the overall variance of the objects as being the sum of the within-group variance and the between-group variance
- A good result will minimize within-group while maximizing between-group variation

If interested... SaTScan (free)

- Cluster detection method
 - Can assess space or space-time clusters
 - Uses circles of varying sizes (cylinders if including time)
 - Compares risk within versus outside circle
 - Most likely cluster -> circle with maximum likelihood (more than expected number of cases)
 - Hot and cold spots
 - Background population-at-risk
- No pre-specified cluster size
- Can take confounding into account

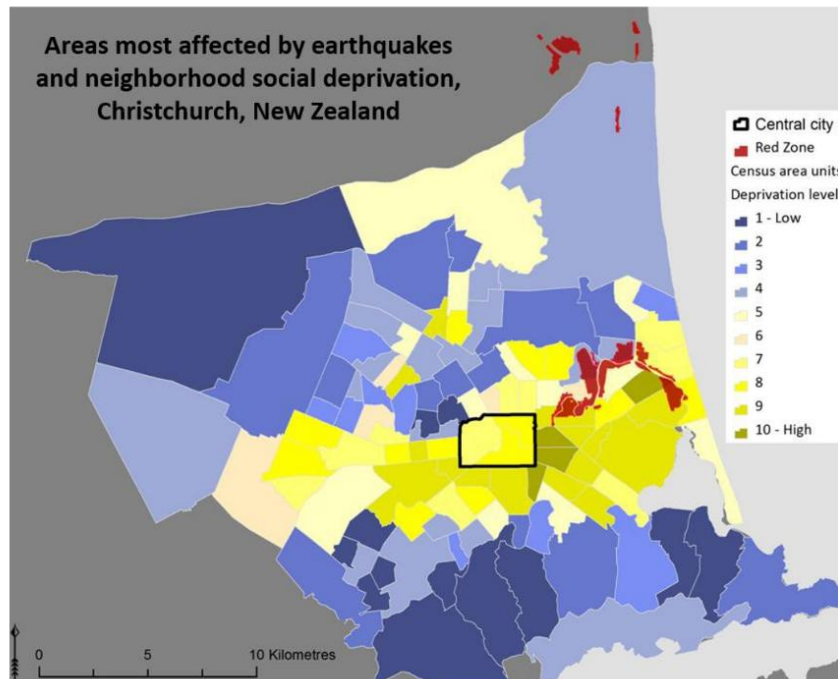
Example: Clustering of pneumococcal pneumonia

- Christchurch, New Zealand
- Devastating earthquakes 2011-2012
- Hundreds of lives lost
- Almost entire central city destroyed



Example: Clustering of pneumococcal pneumonia

- Areas most affected by damage were also the poorest areas



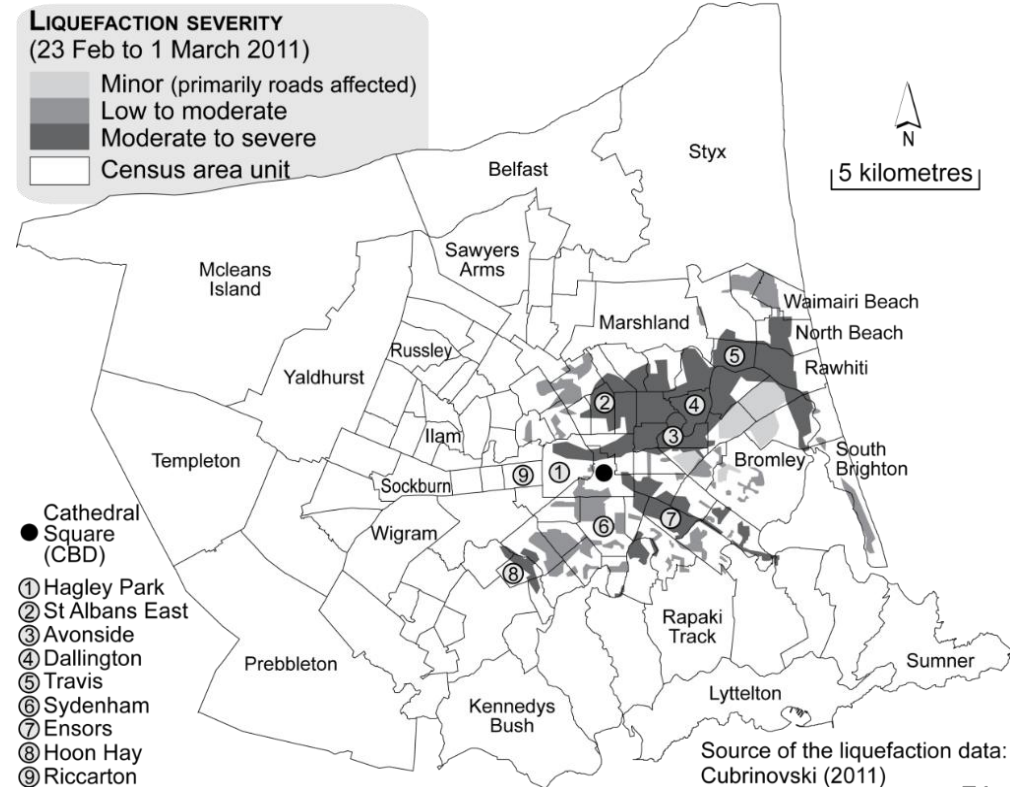
Example: Clustering of pneumococcal pneumonia

- Liquefaction – wet then air borne
 - For many months



Example: Clustering of pneumococcal pneumonia

- Liquefaction ejecta
- A source of airborne dust



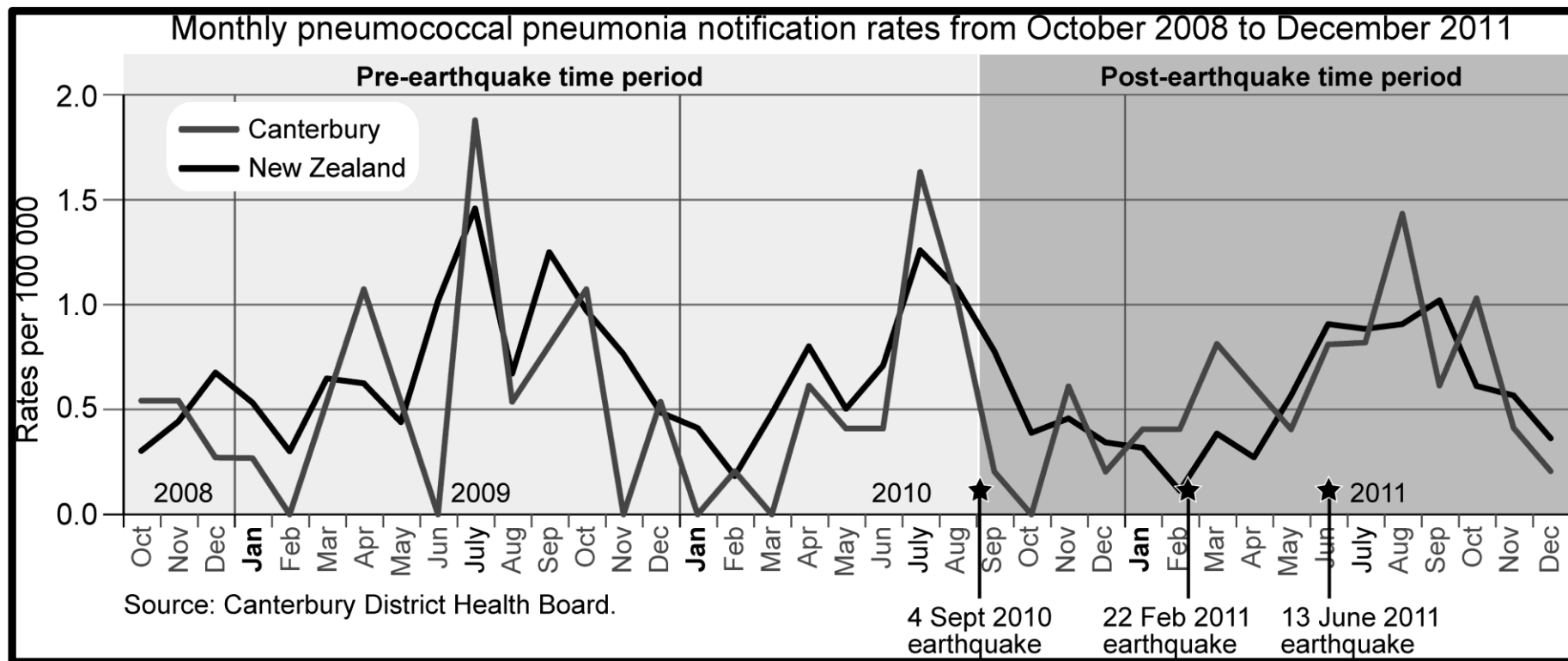
Example: Clustering of pneumococcal pneumonia

- The etiology of pneumococcal pneumonia (PP) is well-known.
- Some events may increase its incidence.
- Natural disasters may worsen air quality, a risk factor for PP.

Example: Clustering of pneumococcal pneumonia

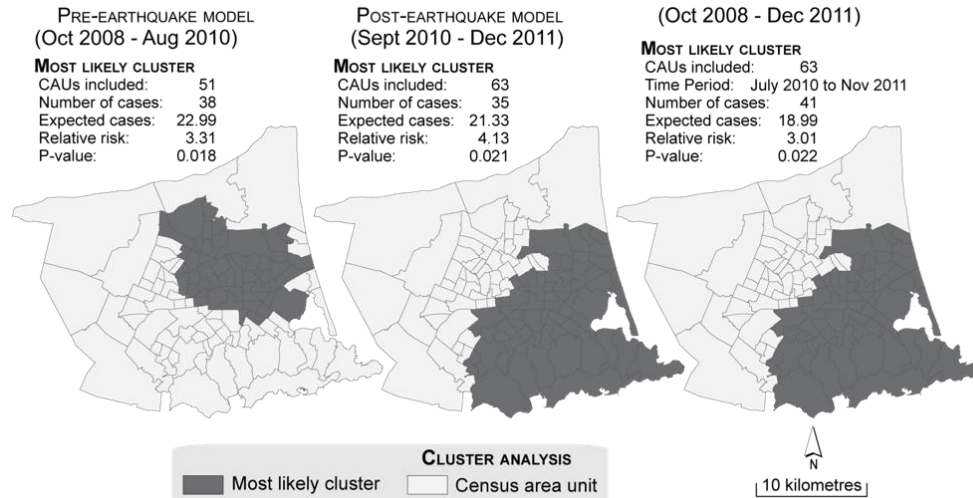
- Data
 - Cases
 - Population at risk
 - Age composition of neighbourhoods (older people more susceptible)
 - Deprivation status of neighbourhoods (poor areas tend to have higher rates)
 - Liquefaction data (spatial)
 - 53 in the pre-earthquake months (22.5 months)
 - 44 in the post-earthquake months (16 months).

Example: Clustering of pneumococcal pneumonia



Example: Clustering of pneumococcal pneumonia

- Results
 - Space-time cluster same as post EQ models
 - Increased relative risk in the fully-adjusted post-earthquake compared to the pre-earthquake cluster
 - Long-term impacts of ejecta inhalation? Seismic activity continues



Challenges of cluster analysis

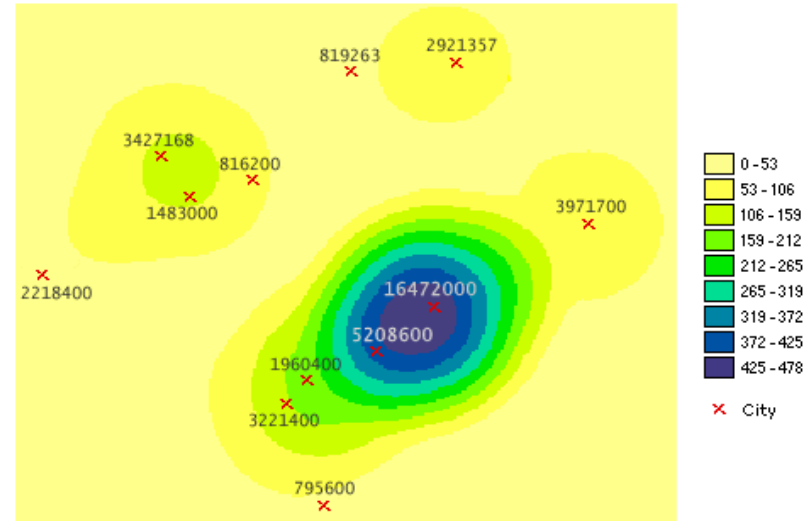
- Link between a causal factor (perhaps a point source) and a disease cluster has been extremely difficult to substantiate in court
 - disease etiology often not fully understood
 - lack of proof of causation - only correlation
 - difficulty modeling underlying population at risk
 - variation of results when spatial resolution is changed (coming later)
 - clustering at a particular place may be due to chance
 - event locations may not indicate site of exposure

Clustering

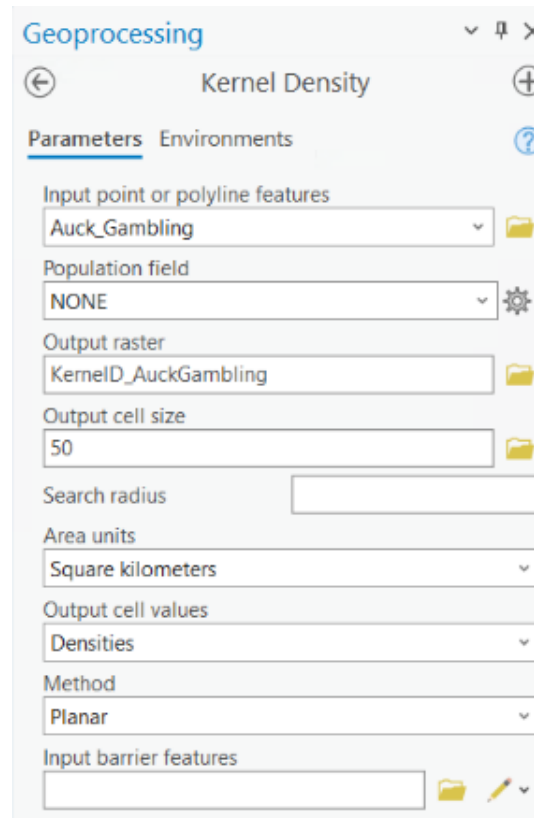
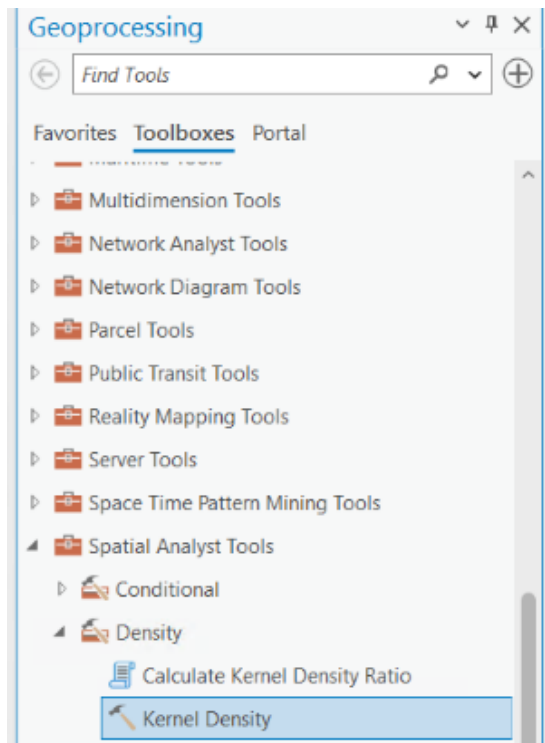
- Typically, one uses geographical statistical programs (ArcPRO, SaTScan, Geoda) to calculate these statistics
- Although some ‘conventional’ statistics programs offer add-ons or packages – increasingly common
 - R
 - Stata

Kernel density in ArcPRO

- Under Spatial Analyst → Kernel Density
- The kernel function is based on the quadratic kernel function described in Silverman (1986, p. 76, equation 4.5)



Kernel density in ArcPRO



Tutorial 4 – kernel density

A. How close is the nearest gambling outlet to another one? How far is the greatest distance between outlets?

- Nearest _____
- Farthest _____

Tutorial 4 – kernel density

A. How close is the nearest gambling outlet to another one? How far is the greatest distance between outlets?

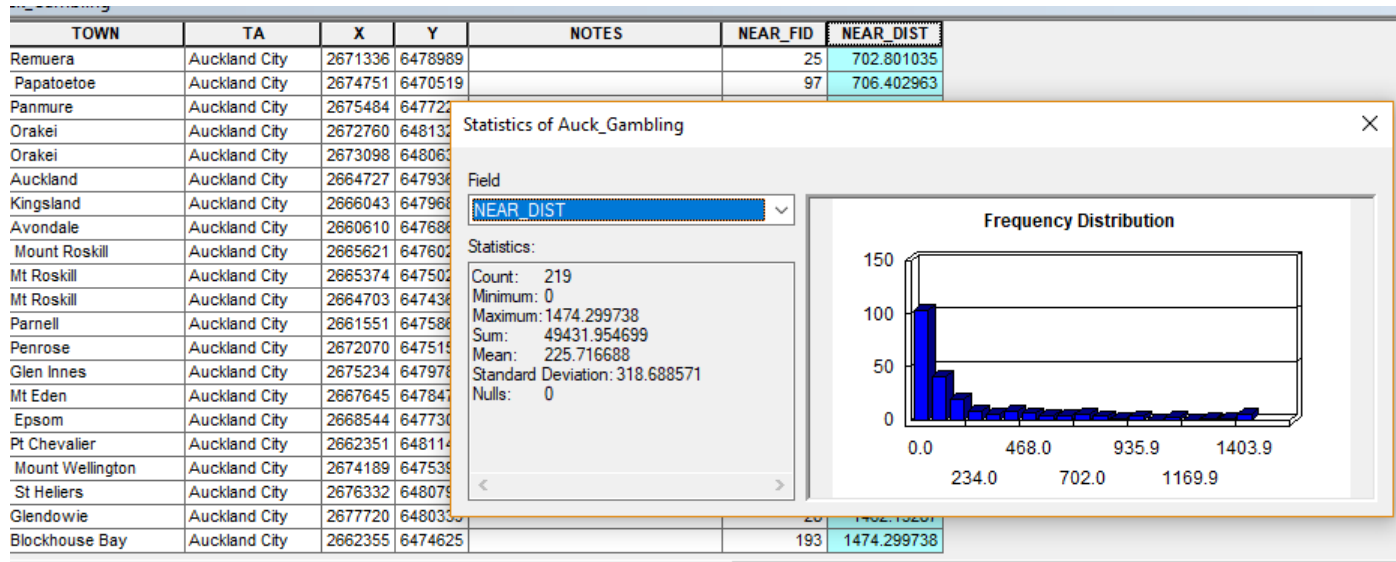
- Nearest = 0m (in the same building, at the same address)
- Farthest = 1474m

Tutorial 4 – kernel density

B. Is the distribution normal? Is there a positive or negative skew?

Tutorial 4 – kernel density

B. Is the distribution normal? Is there a positive or negative skew?



Tutorial 4 – kernel density

C. Were some outliers eliminated? What is the difference between the two raster datasets? Which might be most appropriate? Why?

Tutorial 4 – kernel density

C. Were some outliers eliminated? What is the difference between the two raster datasets? Which might be most appropriate? Why?

Yes, the outlets at the positive tail were not included.

This created tighter areas of high density and more areas of low density.

The most appropriate dataset would depend on the purpose of the research.

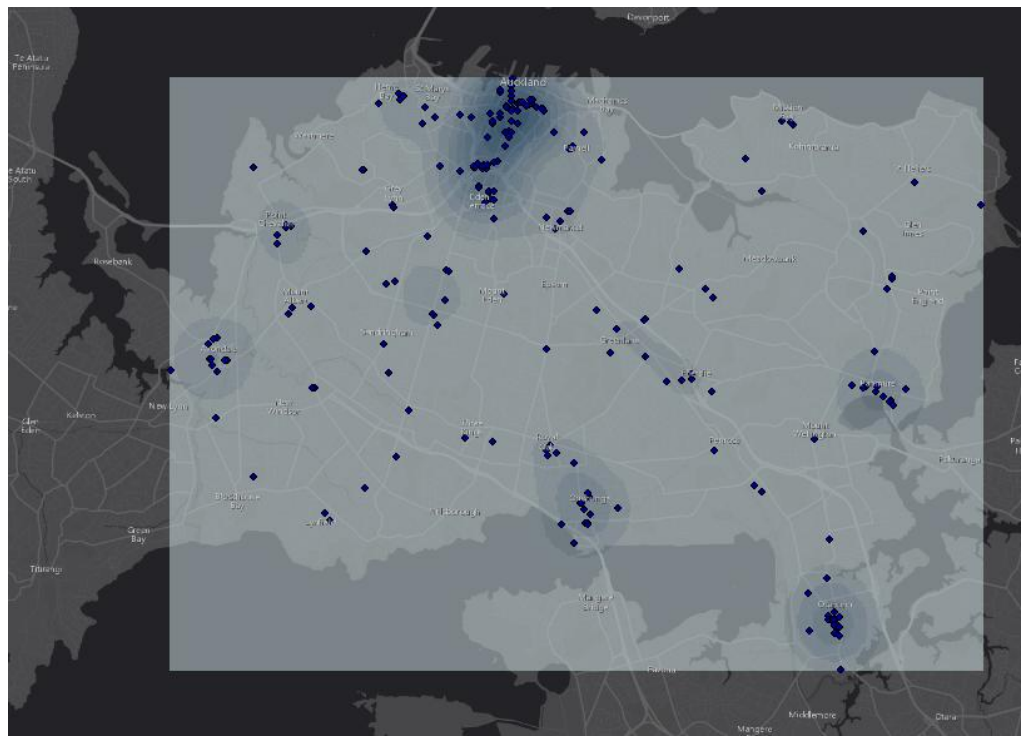
Tutorial 4 – kernel density

D. Across both raster datasets, are there notable areas of high density? Low density?

Tutorial 4 – kernel density

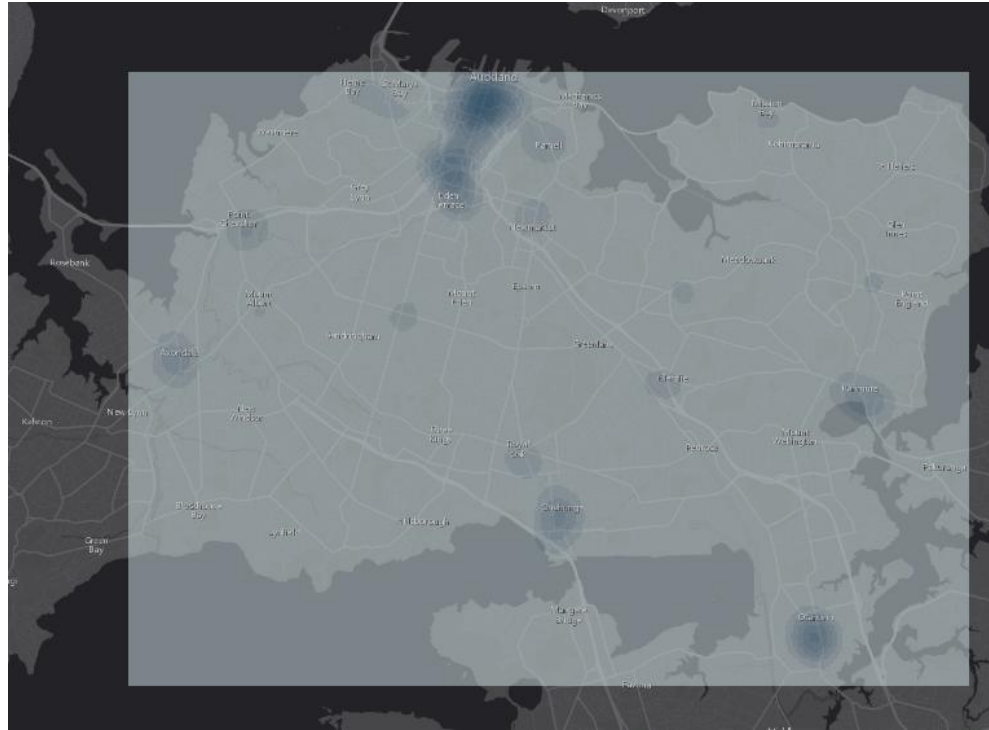
D. Across both raster datasets, are there notable areas of high density? Low density?

Notice the “edge effects”. It can be useful to use a spatial extent larger than your data points to reduce edge effects.



Tutorial 4 – kernel density

D. Across both raster datasets, are there notable areas of high density? Low density?



Tutorial 4 – kernel density

E. What could these raster datasets be used for in health research?

Tutorial 4 – kernel density

F. How is the kernel density result different when we use data on the number of machines?

Tutorial 4 – kernel density

F. How is the kernel density result different when we use data on the number of machines?



Lunch

“Out West”, USA



Section 3: Interpolation and kriging

Kernel density

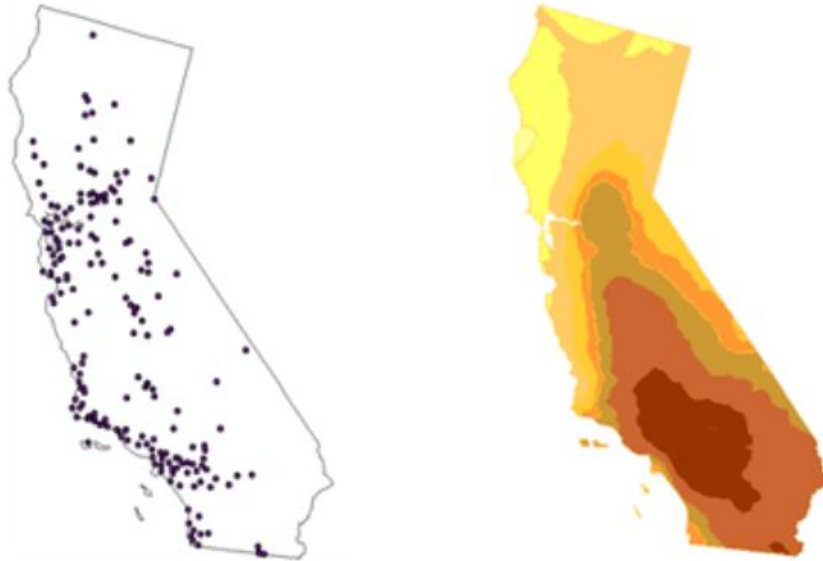
- Note that when we used this technique, we had population data on gambling outlets (we knew all of them across the study area)
- We used this technique to describe the pattern of outlets (or machines), on a continuous surface

Sampled data

- We may seek to use a data of sampled locations to model, and predict attribute values at unknown locations
- This is a way of interpolating values across a surface
- To do so, we need to consider both first and second order spatial trends

Why use interpolation?

- To estimate values where there are no data
- Visualisation of complex point data
- Analysis of trends over space and time



First order v second order spatial trends

- First order

- Relates to patterns across the study area
- E.g. south-north increase in average temperatures across a region
- Often varies with spatial and temporal scale
- Can be modeled using trend surface analysis

- **in practice, we often assume this is not the case, unless clear scientific reason it may exist*

First order v second order spatial trends

- Second order

- Pattern attributed to interaction effects among elements in the pattern
- Alternatively, it is what is left of the pattern once first order variation or trend effects have been removed
- Against some background trend, the second order effects are more localized variations in pattern intensity
- E.g., tree density changing with elevation is a first-order effect, while trees clustering due to seed dispersal is a second-order effect

Interpolation methods

1) Inverse distance weighting and Spline interpolation tools are deterministic interpolation methods

- Directly based on the surrounding measured values or on specified mathematical formulas that determine the smoothness of the resulting surface.

2) Kriging is based on statistical models that include autocorrelation (statistical relationships among the measured points)

- So, can both produce a prediction surface and the uncertainty or accuracy of the predictions
- Heaps...We will cover practical applications but I can give you material with equations and theory
- Other further reading: <http://pro.arcgis.com/en/pro-app/tool-reference/3d-analyst/comparing-interpolation-methods.htm>

Kriging

- Based on the idea that closer points are more similar
- Kriging uses a statistical model to weight sample points
- Models values based on the spatial autocorrelation between points
- The model produces an error surface on the estimated values

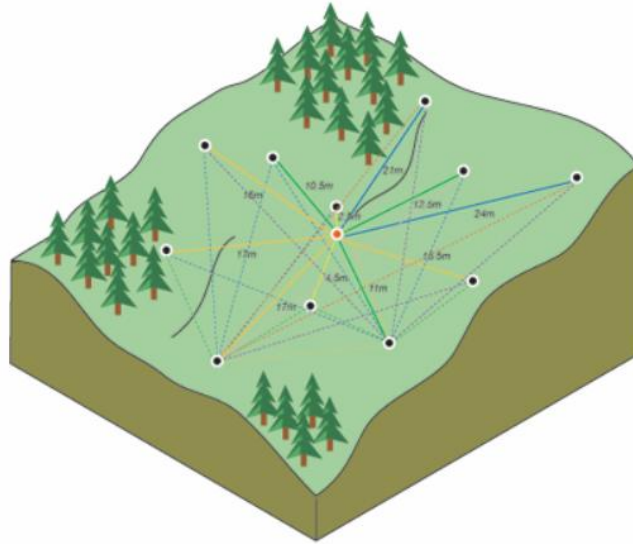
- To predict values at un-sampled locations, given the weighted average of sampled locations ‘in the neighbourhood’
 - Defining neighbourhood
- Similar to regression analysis (squared error prediction)
- Form of Bayesian inference

Kriging

- Kriging involves a two-step process:
 - 1) Creation of the variograms and covariance functions to estimate the statistical dependence (spatial autocorrelation) values that depend on the model of autocorrelation (fitting a model)
 - 2) Prediction of the unknown values

Step 1: Semivariogram

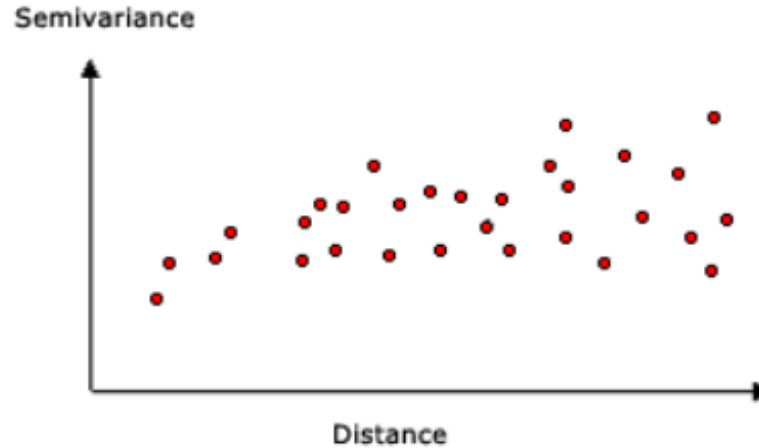
- Calculate squared difference between the values for each point.



- Example, pairing of red point with all other measured locations. Repeat for each point. Define the 'neighbourhood' for such comparisons.*

Step 1: Semivariogram

- Estimation of spatial dependence
 - Things closer together are more similar



Empirical semivariogram graph example

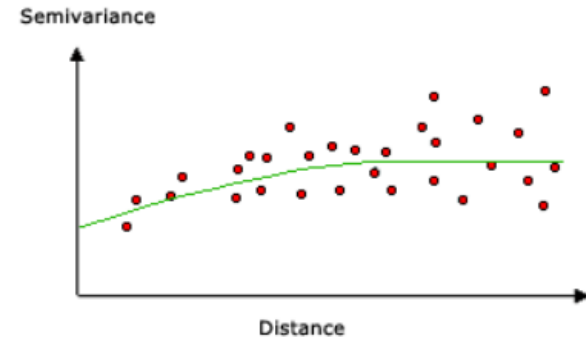
Step 2: Fitting a model to the semivariogram

- The empirical semivariogram provides information on the spatial autocorrelation of datasets
- However, it does not provide information for all possible directions and distances
- It is necessary to fit a model—a continuous function or curve—to the empirical semivariogram
- Similar to fitting a regression line in linear regression modeling

Step 2: Fitting a model to the semivariogram

- The ArcPRO Kriging tool provides the following choices for modeling the empirical semivariogram:

- Circular
- Spherical
- Exponential
- Gaussian
- Linear



- The selected model influences the prediction of the unknown values
 - Tip: the steeper the curve near the origin, the more influence the closest neighbours will have on the prediction. As a result, the output surface will be less smooth. Each model is designed to fit different types of phenomena more accurately.

Making predictions

- Kriging forms weights from surrounding measured values to predict values at unsampled locations
- Sampled values closest to the unsampled locations have the most influence
- Weights for the surrounding sampled points come from the semivariogram
- This creates a continuous surface of predicted values

Primary types of kriging

- Ordinary
 - most widely used
 - assumes the constant mean is unknown
- Universal
 - assumes that there is an overriding spatial trend in the data (e.g., prevailing wind) that can be modeled by a deterministic function
 - should only be used when you know there is a trend in your data and you can provide a scientific justification

Assumptions of kriging

- Assumes that the spatial variation in the phenomenon is statistically homogeneous throughout the surface
- The same pattern of variation can be observed at all locations in the study area
- This hypothesis of spatial homogeneity is fundamental to the regionalized variable theory
- *Think about assumptions about homoscedasticity in regression modeling*

Further reading

- This is a brief introduction, if you wish to use this method, I recommend further reading
- Burrough, P. A. Principles of Geographical Information Systems for Land Resources Assessment. New York: Oxford University Press. 1986.
- Heine, G. W. "A Controlled Study of Some Two-Dimensional Interpolation Methods." COGS Computer Contributions 3 (no. 2): 60–72. 1986.
- McBratney, A. B., and R. Webster. "Choosing Functions for Semi-variograms of Soil Properties and Fitting Them to Sampling Estimates." Journal of Soil Science 37: 617–639. 1986.
- Oliver, M. A. "Kriging: A Method of Interpolation for Geographical Information Systems." International Journal of Geographic Information Systems 4: 313–332. 1990.
- Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. Numerical Recipes in C: The Art of Scientific Computing. New York: Cambridge University Press. 1988.
- Royle, A. G., F. L. Clausen, and P. Frederiksen. "Practical Universal Kriging and Automatic Contouring." Geoprocessing 1: 377–394. 1981.

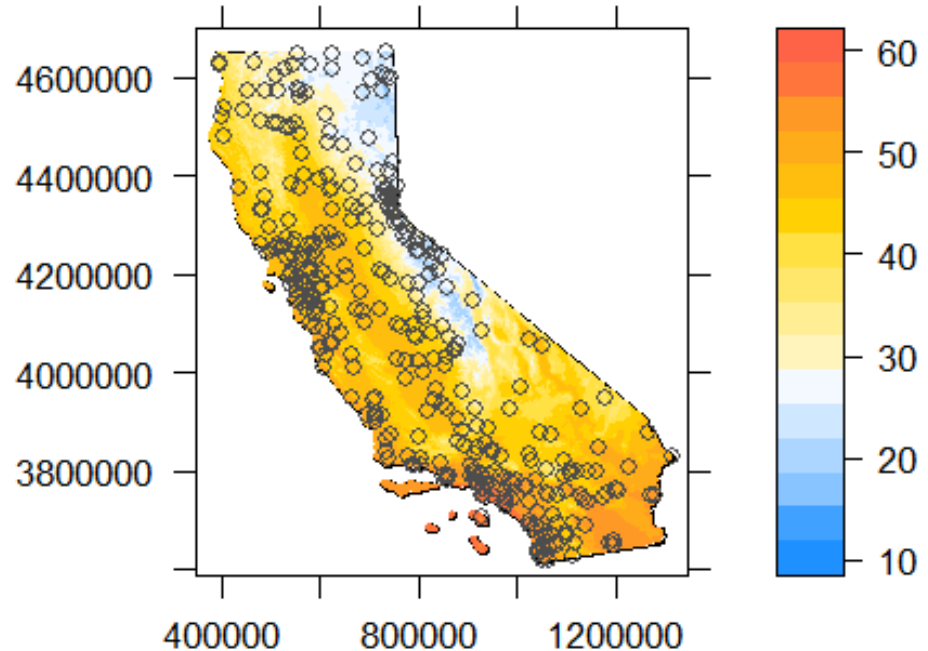
Example - Kriging

- We have climate data from California. We are interested in whether we can estimate average monthly temperatures from weather stations scattered throughout the state
- We will do universal kriging to create a continuous surface with monthly mean temperature across the entire state of California

Example - Kriging

- We created a surface of estimated average temperature
- We now can have an idea of how monthly temperatures differ across the state

CA MNTM: Universal Kriging Surface



Example - Kriging

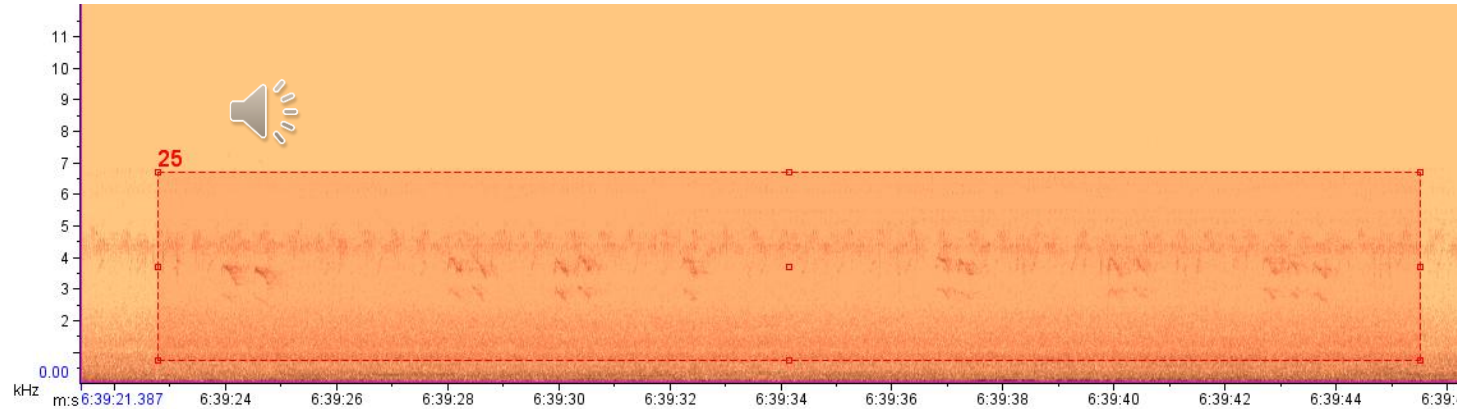
- Lab evidence suggests that natural sounds (eg, birdsong) appear to have positive effects on health
 - Stress (recovery from stress) – cortisol, heart rate variability, skin conductance
 - Sleep
 - Pain intensity
 - Recovery from surgery
- We wanted to know if everyday soundscapes also affect health

Example - Kriging

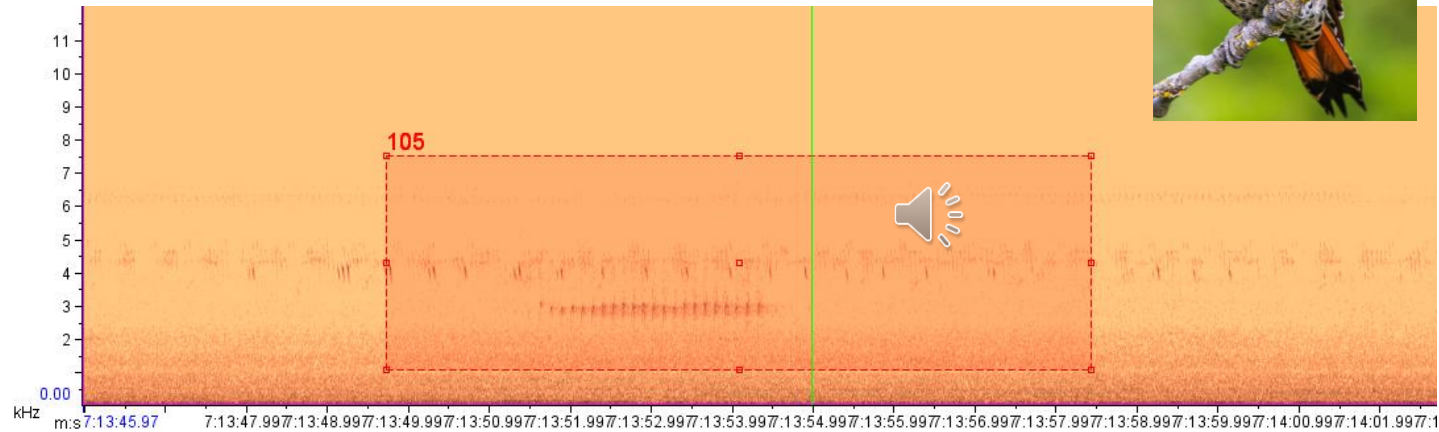
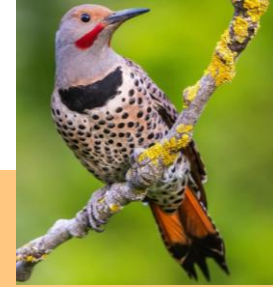
- So, we collected audio recordings at locations in Detroit
- Deployed units for one week
- Listened to hours of recordings over 4 days (dawn and dusk chorus)
- Identified bird and insect calls



Blue Jay, (*Cyanocitta cristata*)

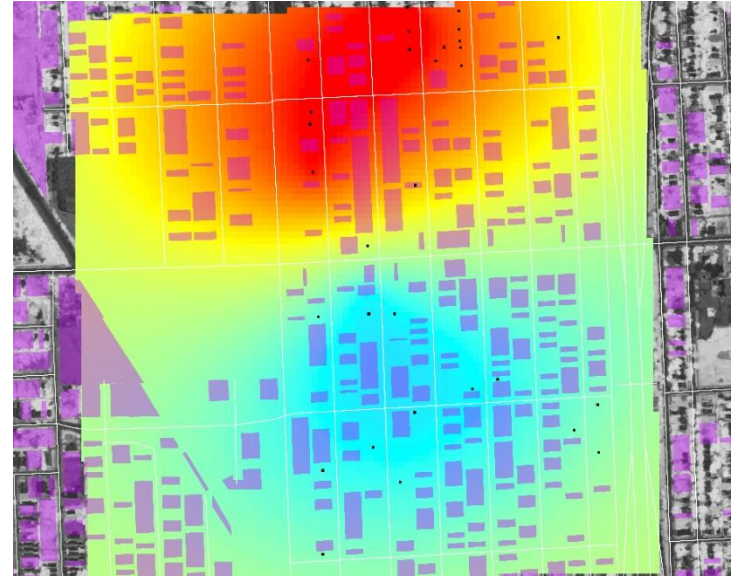
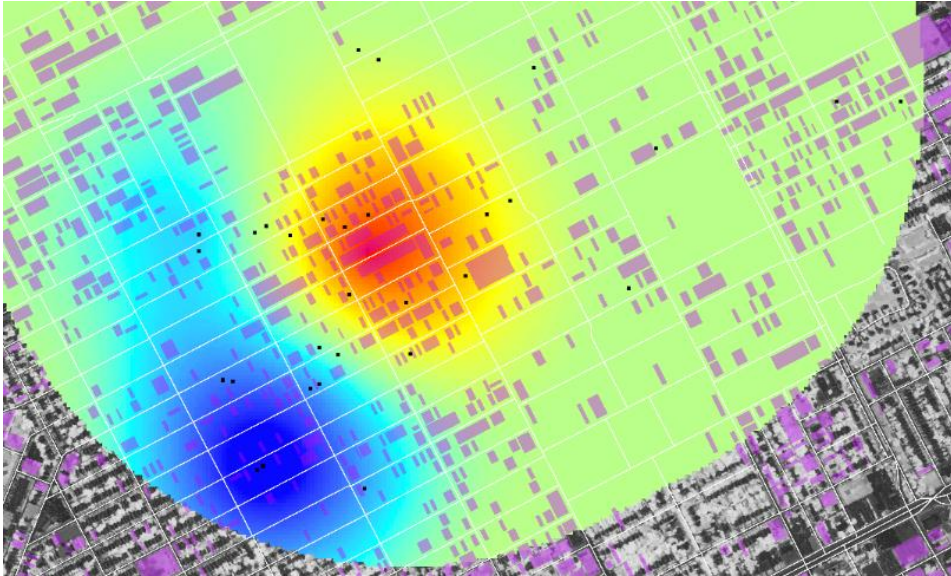


Norther Flicker, (*Colaptes auratus*)



Example - Kriging

- Kriged surface of bird species count
- Assigned values to study participant's home location



Example - Kriging

- Largest effect – blighted lots
- Similar effect size for:
 - marital status
 - neighborhood deprivation
 - bird species
- For every bird species, decrease in PSS of over 1 point
 - Ref: normal in this pop ~7
 - High stress = 27-40
 - Our mean = 17

	Perceived Stress Score	
	Std β	p
Bird species	-0.28	0.038
age	0.02	0.915
female	-0.08	0.554
length residency	0.13	0.868
employed	0.02	0.293
married or partnered	0.31	0.067
own home	-0.23	0.143
blighted lots within 100m	0.42	0.017
blighted structures within 100m	0.16	0.255
crime	-0.13	0.311
neighborhood deprivation	-0.32	0.046

Summary of interpolation objectives

1. Identify significant first-order trends across study area
2. Characterize local (second-order) variability
3. Establish factors that "explain" these trends
4. Predict values of z at unmeasured locations
5. Attach uncertainty estimates to such predictions

Spatial autocorrelation

- What is it?
 - Spatial autocorrelation measures the degree **to which one object is similar to other nearby objects OR things near one another have contrasting values.** Global Moran's I (Index) measures spatial autocorrelation.
- Global Moran's I can be classified as positive, negative and no spatial auto-correlation.
 - Positive = similar values cluster together
 - Negative = dispersion
 - Just like Pearson's or Spearman's
- Inferential spatial statistic

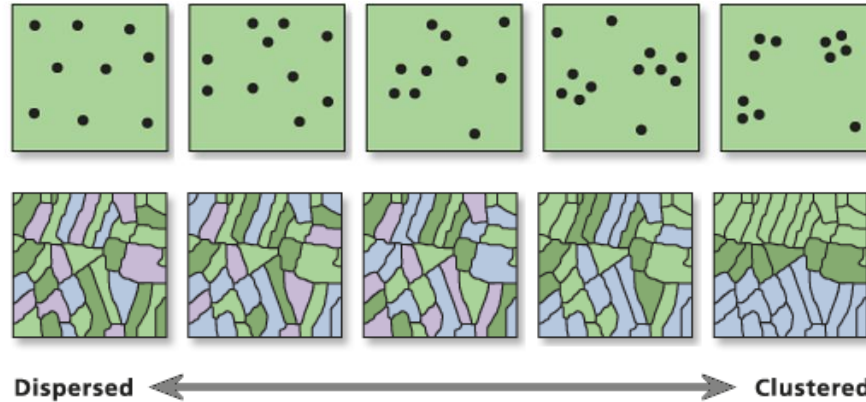
Moran's Index	
Primary Objective:	Identify significant spatial patterns within a study area
Requirements and Assumptions:	
1.	Minimum of (30) geographic features
2.	Attribute values measured on an ordinal or interval/ratio scale
Hypotheses:	
H_0 :	Attribute values are randomly distributed across features in the study area
H_A :	Attribute values are not randomly distributed across features in the study area
Test Statistic:	
$I = \frac{n \sum (x_i - \bar{x})(x_j - \bar{x})}{f \sum (x - \bar{x})^2}$	
Interpretation:	
Assuming a significant p-value:	
$I < 0$	(observed pattern is dispersed)
$I = 0$	(observed pattern is random)
$I > 0$	(observed pattern is clustered)

Moran's I Index (Global)

- Evaluates both the pattern of the geographic locations **and** the attribute values to determine clustering across a study area
 - Can be applied to points and polygons
 - Attributes can be interval/ratio or ordinal
- Identifies each pair of neighbouring locations across a study area and compares attributes
- More to come...
 - Can define 'neighbours' as those sharing a border, or adjacent
 - Rook vs queen contiguity

Spatial autocorrelation

- We have positive, negative and no autocorrelation



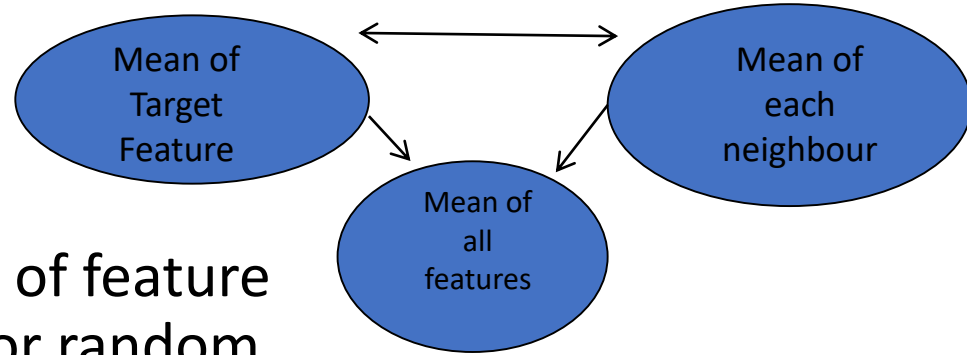
- Null hypothesis: the attribute being analyzed is randomly distributed across the study area
- P-values indicate accept or reject null

Spatial autocorrelation

- Why does it matter?
 - Statistics rely on observations being independent from one another. If autocorrelation exists, then **this violates the assumption**. A nuisance
 - Analyzing clusters and dispersion. Of interest

Spatial Autocorrelation

- **Global** statistic
- Measures whether the pattern of feature *values* is clustered, dispersed, or random.
- Compares the difference between the mean of the target feature and the mean for all features to the difference between the mean for each neighbour and the mean for all features.
- For more information on the equation, see [ESRI online help](#)



Spatial autocorrelation

- Positive spatial autocorrelation
 - Global Moran's I values near +1
 - Indication of clustering
 - Ex: Elevation values; Moran's I = 0.60, $z = 4.95$, $p < 0.01$ that pattern is random



Clustered Image Spatial Autocorrelation

Spatial autocorrelation

- Negative spatial autocorrelation
 - Global Moran's I values near -1
 - Indication of dispersion
 - Ex: Elevation values; Moran's I = -1.00, $z = -7.59$, $p < 0.01$ that pattern is dispersed

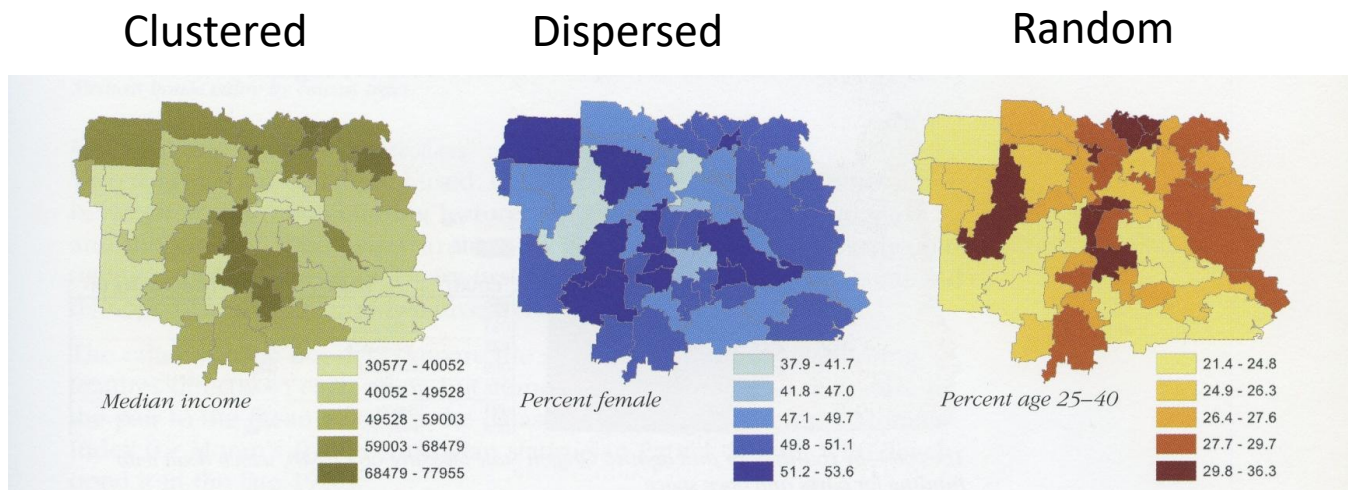


Checkerboard Pattern: Spatial Autocorrelation

$I=0$ =random distribution

$I<0$ =Values dispersed

$0<I$ =Values clustered



Spatial autocorrelation

- Global Moran's I yields one value for entire study area (does not tell you the location of clusters)
- May also be interested in patterns within a subset of the study area
 - Are there pockets of spatial autocorrelation?
 - Often called hotspot analysis
 - We can find clusters at the local level (and determine their location)...
 - Local Moran's I
 - More complex space-time scan statistics (for Bernoulli or Poisson distributions, eg cases)
 - Lots of other techniques (recommend SaTScan)

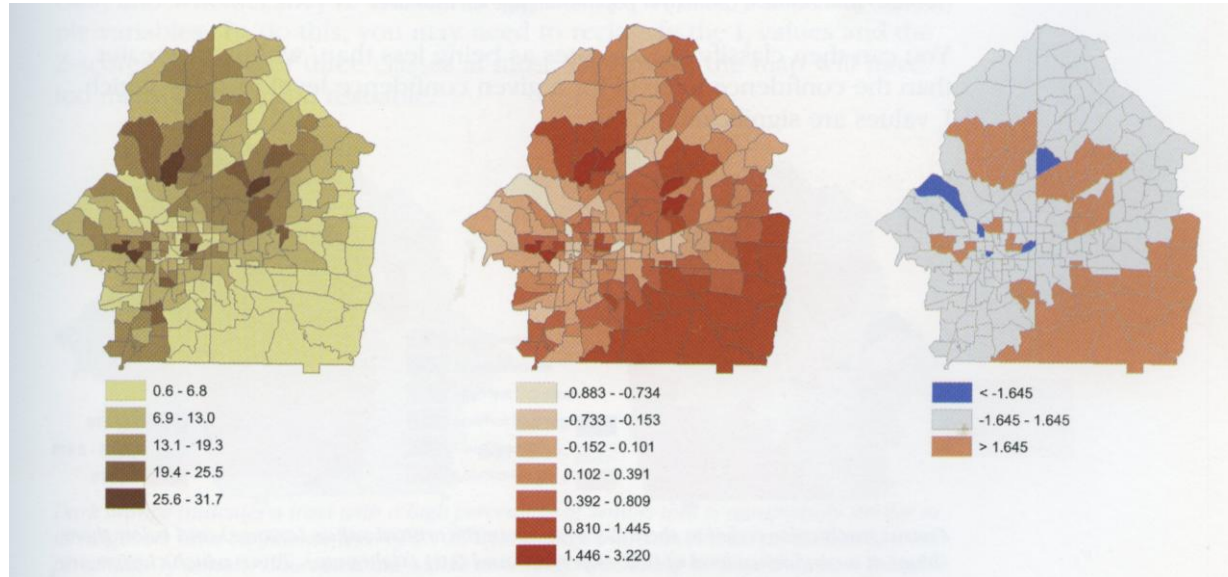
Anselin Local Moran's I

- **Local** statistic
- Measures the strength of patterns for each specific feature.
- Compares the value of each feature in a pair to the mean value for all features in the study area.
- See [ESRI online help](#) for the equation.

Anselin Local Moran's I

- Positive I value:
 - Feature is surrounded by features with similar values, either high or low.
 - Feature is part of a cluster.
 - Statistically significant clusters can consist of high values (HH) or low values (LL)
- Negative I value:
 - Feature is surrounded by features with dissimilar values.
 - Feature is an outlier.
 - Statistically significant outliers can be a feature with a high value surrounded by features with low values (HL) or a feature with a low value surrounded by features with high values (LH).

Anselin Local Moran's I



Census tracts
for percentage
65 and above

I values

Z-scores

Spatial Neighbourhoods and Weights

- Neighbourhood = area in which the GIS will compare the target values to neighbouring values
- Neighbourhoods are most often defined based on adjacency or distance, but can be defined based on travel time, travel cost, etc.
- You can also define a cutoff distance, the amount of adjacency (borders vs. corners), or the amount of influence at different distances
- A table of spatial weights is used to incorporate these definitions into statistical analysis.

Distance Models

- Inverse distance – all features influence all other features, but the closer something is, the more influence it has
- Distance band – features outside a specified distance do not influence the features within the area
- Zone of indifference – combines inverse distance and distance band

Adjacency Models

- K Nearest Neighbours – a specified number of neighbouring features are included in calculations
- Polygon Contiguity – polygons that share an edge or node influence each other
- Spatial weights – specified by user (ex. Travel times or distances)

Types of Contiguity

- Rook = Share edges
- Bishop = share corners
- Queen = share edges or corners
- Secondary order contiguity = neighbour of neighbour

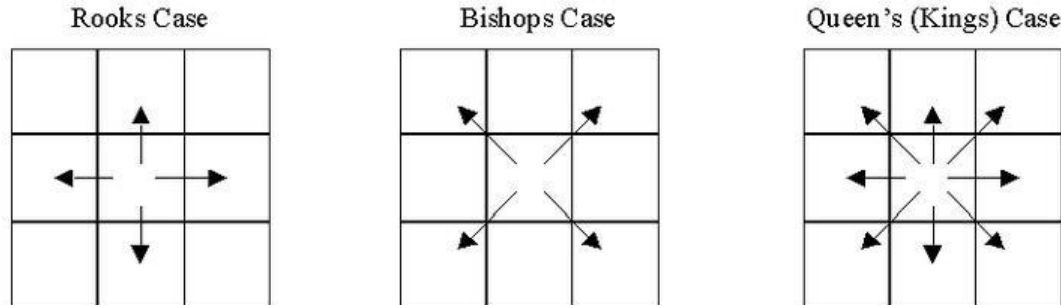


Image from: <http://www.lpc.uottawa.ca/publications/moransi/moran.htm>

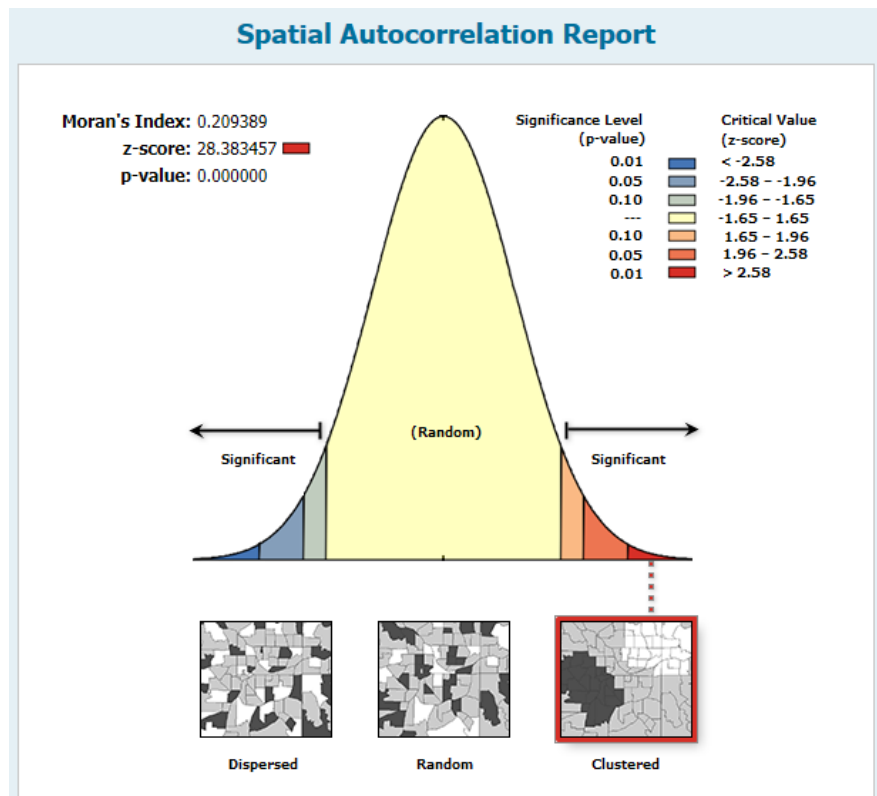
Spatial autocorrelation

- We believe that the location of alcohol outlets might influence crime in the neighbourhood
- It is possible that people who are consuming alcohol, late into the night may be prone to engage in criminal behaviours, particularly near outlets
- We want to test whether there is an association between the amount of alcohol outlets and the crime rate
- What's our dependent variable?
- What is one of the assumptions of regression analysis?

Tutorial 5 – spatial autocorrelation

A. What are the results? How do you interpret these results; what are your conclusions?

Tutorial 5 – spatial autocorrelation



Given the z-score of 28.3834569179, there is a less than 1% likelihood that this clustered pattern could be the result of random chance.

Tutorial 5 – spatial autocorrelation

Global Moran's I Summary

Moran's Index:	0.209389
Expected Index:	-0.000716
Variance:	0.000055
z-score:	28.383457
p-value:	0.000000

Dataset Information

Input Feature Class:	test2
Input Field:	AVGALLCRIM
Conceptualization:	INVERSE_DISTANCE
Distance Method:	EUCLIDEAN
Row Standardization:	False
Distance Threshold:	725.2600 Meters
Weights Matrix File:	None
Selection Set:	False

Tutorial 5 – spatial autocorrelation

B. Did the analysis using k-nearest neighbours change your conclusions?

Tutorial 5 – spatial autocorrelation

B. Did the analysis using k-nearest neighbours change your conclusions?

No.

Global Moran's I Summary

Moran's Index: 0.202561

Expected Index: -0.000716

Variance: 0.000165

z-score: 15.813303

p-value: 0.000000

Tutorial 5 – spatial autocorrelation

B. Did the analysis using k-nearest neighbours change your conclusions?

Here, we are treating spatial autocorrelation as a nuisance.

Tutorial 5 – spatial autocorrelation

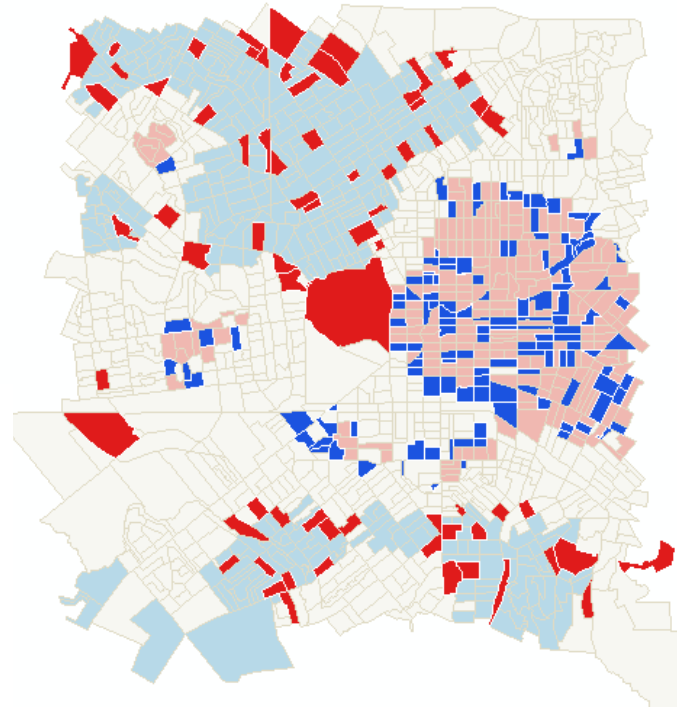
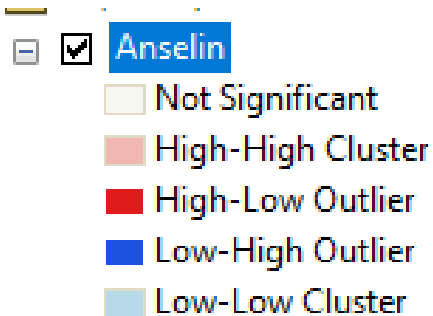
C. Are there significant clusters?

D. Outliers?

Tutorial 5 – spatial autocorrelation

C. Are there significant clusters?

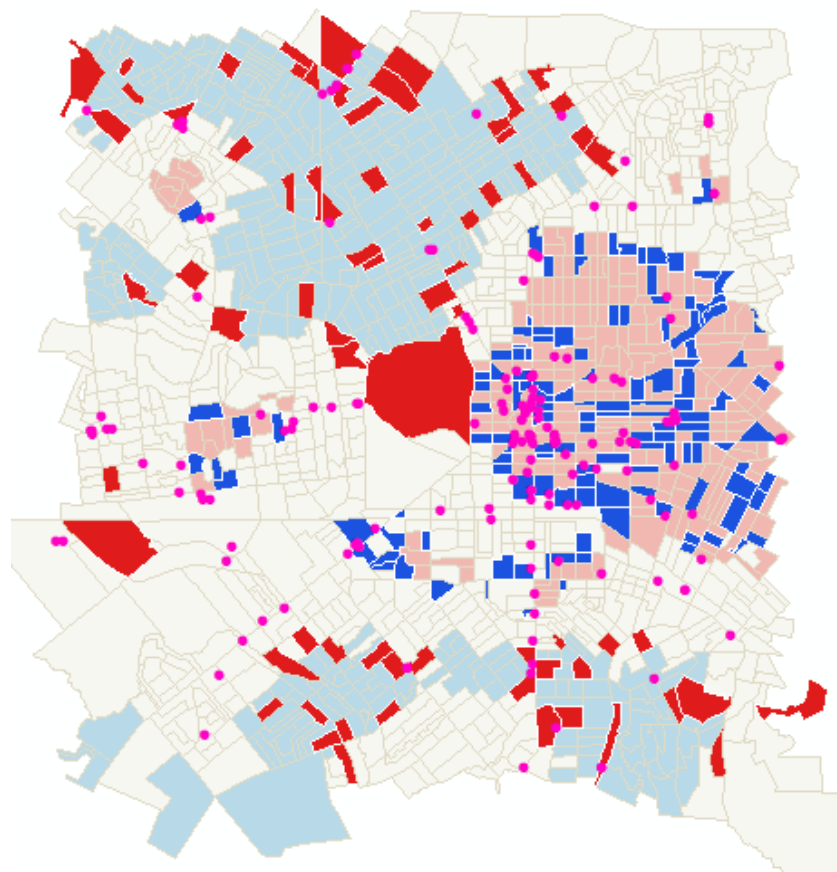
D. Outliers?



Tutorial 5 – spatial autocorrelation

E. Do you see any patterns worth exploring?

Tutorial 5 – spatial autocorrelation



Tutorial 5 – spatial autocorrelation

F. Can you think of three possible next steps or limitations in terms of answering your original research question?

Tutorial 5 – spatial autocorrelation

F. Can you think of three possible next steps or limitations in terms of answering your original research question?

Limitation: Assumptions of independence in outcome variable not met.

Next steps: Could look at overlap in clustering of alcohol outlets. Could look at whether the count of alcohol outlets with 200m of a H-H cluster has significantly more alcohol outlets than buffers around L-H clusters.

Afternoon break



Central Turkey

Section 4: Considerations for regression modeling

Topics to consider in your own regression modeling

- Spatial autocorrelation, nested units, complex sampling and regression
- Ecological fallacy
- Modifiable Area Unit (MAUP) problem

OLS regression assumptions, scientific steps

1. Dependent variable measured continuous/quantitative
2. Observations of are independent of one another [[sampling, theory](#)]
3. The relationship is linear [[scatterplot, Pearson's correlation, theory](#)]
4. Data show homoscedasticity [[scatterplot and white's test](#)]
5. Residuals: [[residual plot, map](#)]
 - a) have a mean of zero (there is no systematic misspecification or bias in the population regression equation)
 - b) show homoskedasticity and are uncorrelated;
 - c) have a fairly normal distribution

What does 'independent observations' mean?

- The value of one observation does not affect the value of other observations
- Non-independent observations can make your statistical test give too many false positives
- Usually, this is violated when samples are taken too close in space or time
- Example:
 - You want to know whether calico cats have a different mean weight than black cats
 - You get five calico cats, five black cats, weigh them, and then fit a regression to test whether cat color predicts weight
 - If the five calico cats are all from one litter, and the five black cats are all from a second litter, then the measurements are not independent
 - Some cat parents have small offspring, while some have large
 - This is the result of clustered sampling [[hierarchical regression](#)]

What does 'independent observations' mean?

- The most common violation of this assumption in regression is in time series data, where Y has been measured at different times
- Example:
 - Biologists have counted the number of moose on Isle Royale, a large island in Lake Superior, every year
 - Moose live a long time, so the number of moose in one year is not independent of the number of moose in the previous year, it is highly dependent on it; if the number of moose in one year is high, the number in the next year will probably be pretty high, and if the number of moose is low one year, the number will probably be low the next year as well
 - This non-independence, or "autocorrelation," can yield significant results, even when the null hypothesis of no relationship between time and Y is true
 - If both X and Y are time series—for example, you analyze the number of wolves and the number of moose on Isle Royale—you can also get a "significant" relationship between them much too often.

What does 'independent observations' mean?

- What about 'spatial autocorrelation'?
- Can occur when you measure a variable at locations that are close enough together that nearby locations will tend to have similar values
- Example:
 - You want to know whether the abundance of dandelions is associated with the level of phosphate in the soil
 - You count dandelions and measure soil phosphate in 1 m² quadrants in a field
 - However, if one quadrant has a lot of dandelions, its neighbouring quadrants will also have a lot of dandelions, for reasons that may have nothing to do with phosphate (seed spread)
 - Similarly, soil composition changes gradually across most areas, so a quadrant with low phosphate will probably be close to other quadrants that are low in phosphate
 - It would be easy to find a significant correlation between dandelion abundance and phosphate concentration, even if there is no real relationship

Modeling options

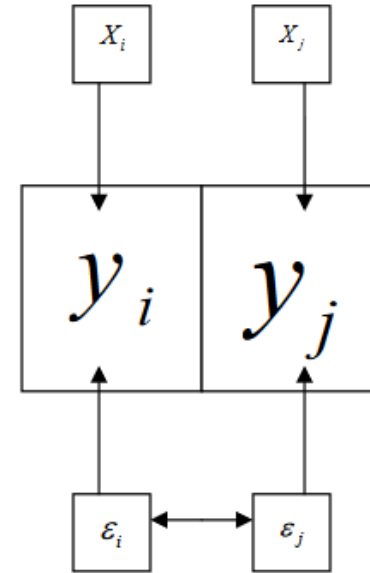
- If sampled from the same units (school, hospital, neighbourhood), include a random error term for that unit = hierarchical modeling
- Must also evaluate spatial dependence in the residuals
- If the residuals are spatially autocorrelated, this indicates that the model is misspecified
- Either:
 - improve model by adding (and perhaps removing) variables
 - use a regression model that accounts for spatial autocorrelation

Complex sampling

- Clusters – often geographic units
- Weights
- Svy code in Stata
- Sometimes limited post-estimation commands, etc.
- Eg, New Zealand Health Survey

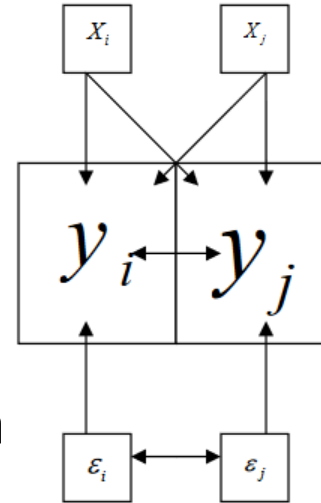
Spatial error

- Errors across two geographic units are correlated
- Assumption of OLS violated
 - Uncorrelated errors
- Indicates omitted covariates
- Affects inference



Spatial lag

- Dependent variable in one place is affected by independent variables in neighbouring places
- Assumptions of OLS violated
 - Uncorrelated errors
 - Observations are independent
- Estimates biased
- Events in one place predict likelihood of similar events in neighbouring places



Lag v error

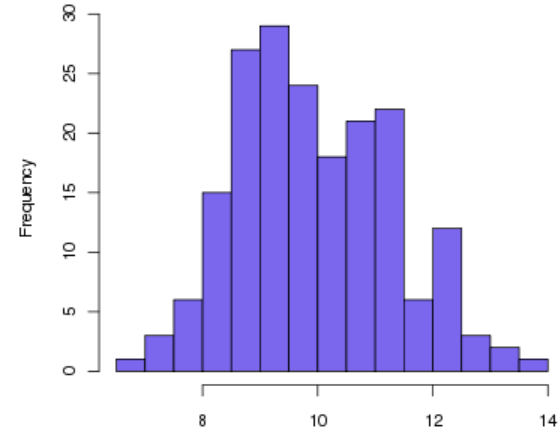
- One spatial dependence identified
- Use theory to inform choice
- Is spatial dependence of interest (in terms of the relationships between places) or simply a nuisance in the data?

Many types of geographic regression modeling

- Just like non-spatial statistics, heaps of options depending on the data and the intention
- Not going into all options
 - Generalized Least Squares
 - Simultaneous Autoregressive Models (SAR)
 - Conditional Autoregressive Models (CAR)
 - Empirical Bayes Estimation
 - Geographically weighted regression (GWR)

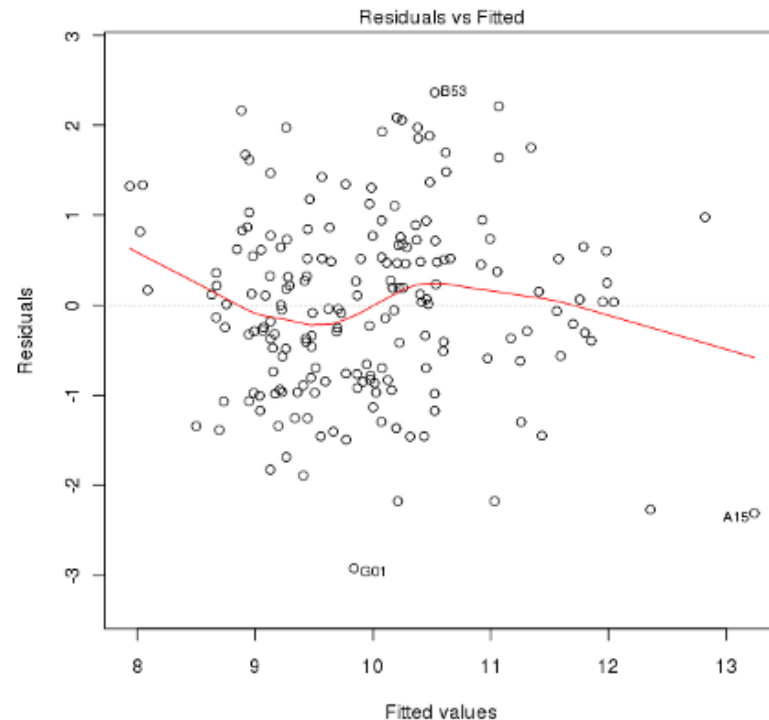
Example: Spatial regression 1

- We have data on all heart attack rates across 190 districts in England in 2017
- We want to fit a regression to predict MI:
 - lung cancer rates
 - age composition
 - neighbourhood deprivation



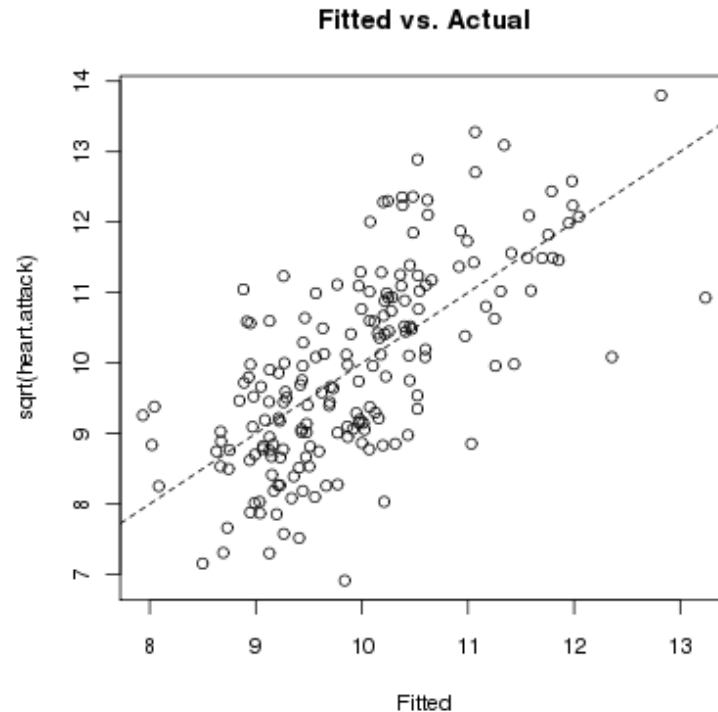
Example: Spatial regression 1

- Fitted v residuals



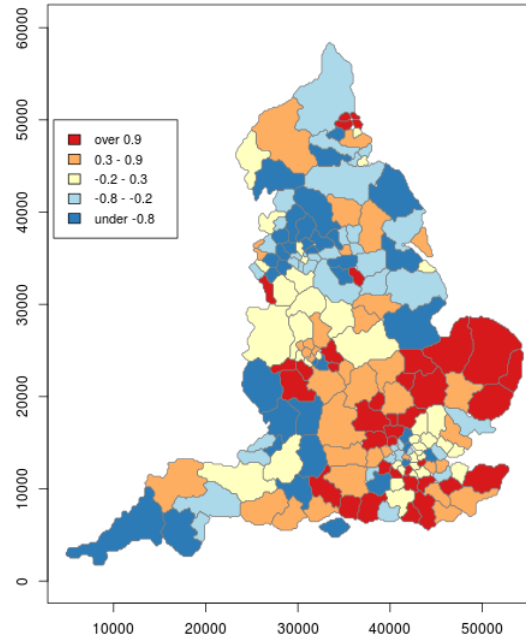
Example: Spatial regression 1

- Fitted v actual



Example: Spatial regression 1

- Map of residuals



- Are the residuals autocorrelated?
- Moran's $I = 0.26$, $p\text{-value} < 0.001$

Example: Spatial regression 1

- What are the implications?
- What should we do next?

Example: Spatial regression 2

- We were interested in whether there are neighbourhood characteristics which might relate to
 - Higher than expected premature mortality rates
 - Lower than expected premature mortality rates
 - Among Māori
- Step 1 – build a robust model to predict premature mortality rates
 - Indep vars: NZDep, Background population (offset)

Example: Spatial regression 2

- First, our data were compiled for each CAU
- We were concerned about possible non-independence of death counts within the same District Health Board
 - differences in healthcare services between DHBs that could impact survival, particularly with injuries and acute causes of death in those in younger age groups

Example: Spatial regression 2

- So, we utilised a multi-level model (each CAU was 'nested' in a DHB)
 - Intra-class correlation coefficient (ICC) was used to assess non-independence
 - cluster robust standard errors to account for it

Example: Spatial regression 2

- We were also concerned that neighbouring CAUs might also be more similar to one another in terms of death counts and the drivers of premature death
- To rule out the potential for our regression analyses to lead to spatially correlated errors (at smaller scales than DHB level and specifically for CAUs neighbouring one another), we calculated the Moran's I statistic

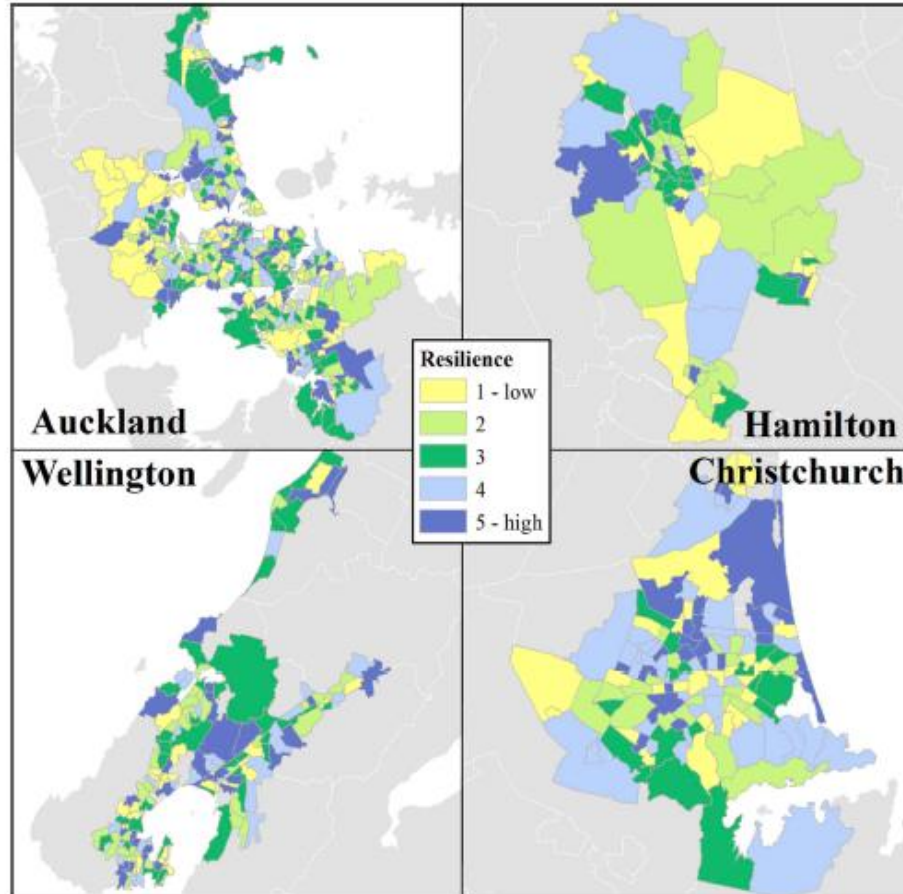
Example: Spatial regression 2

- Moran's I Statistic 0.158, $p=0.001$
- Conclusion?
- What would you do next, based on this statistic and the theoretical relationships?

Example: Spatial regression 2

- We then generated a spatial weights matrix to calculate the weighted average of all neighbouring values for premature death for all CAUs
- This spatial lag variable was then treated as an independent variable in the regression model
- After fitting the model, we generated deviance residuals for each CAU and ranked them from high to low to examine trends across these neighbourhoods

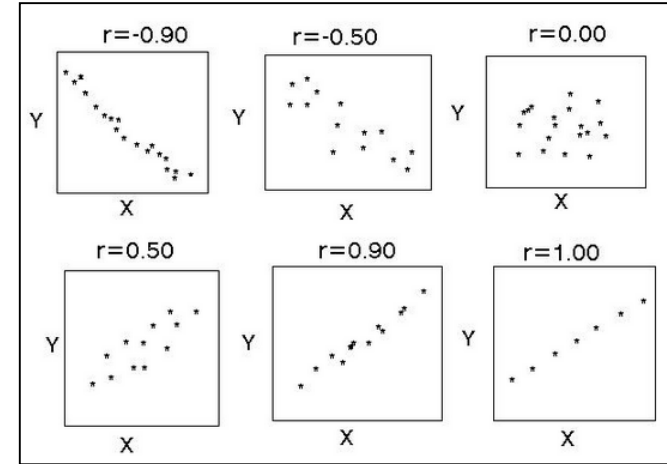
Example: Spatial regression 2



- Definition: is a logical fallacy in the interpretation of statistical data where inferences about the nature of individuals are deduced from inference for the group or area within which the individuals belong
- Exposure: measured at area level
- Outcome: measured at area level

Ecological fallacy

- For each US state, we have:
 - the literacy rate
 - the proportion of the population born outside the US
- Correlation = 0.53
- Higher % immigrants → Higher literacy rates
- We also have a sample of individuals:
- Correlation = -0.11
- Fallacy:
- The positive correlation at the state level was because immigrants tended to settle in states where the native population was more literate.



1950 ["Ecological Correlations and the Behavior of Individuals"](#). *American Sociological Review* 15 (3): 351–357.

Ecological fallacy

- **Cannot** make inference:
 - About individuals
- **Can** make inference:
 - About areas or neighbourhoods
 - Ex: High density of fast food outlets in neighbourhoods was associated with higher neighbourhood rates of obesity

Does passport ownership prevent diabetes?

- Exposure: passport ownership by state

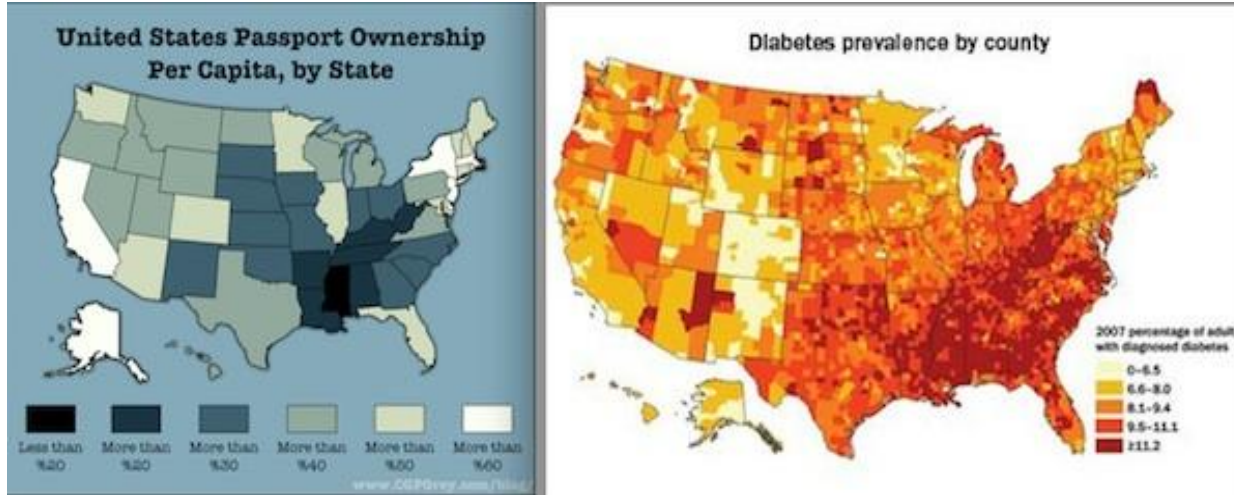


Does passport ownership prevent diabetes?

- Outcome: diabetes prevalence by county



Conclusion?



- Does this mean that if I own a passport, I can prevent developing diabetes?

Conclusion?

- Higher levels of passport ownership in states with higher income
 - Having the money to travel also usually means having access to healthy foods and sufficient health insurance
- Correlation v causation, other factors, no biological pathway
- No inference at individual-level

Ecological studies

- Limitations
 - Observations made at the group level may not represent the exposure-disease relationship at the individual level
 - Risk of ecological fallacy
 - Can't control for some potential confounding factors
 - Correlation does not equal causation

GIS $\neq$ ecological study

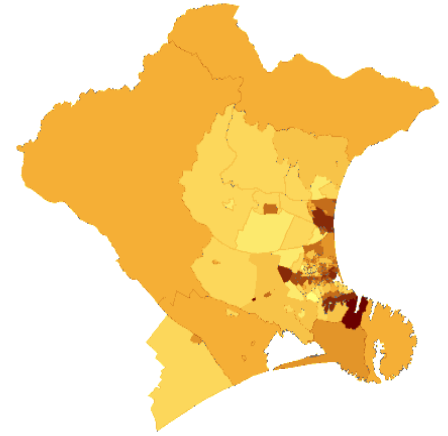
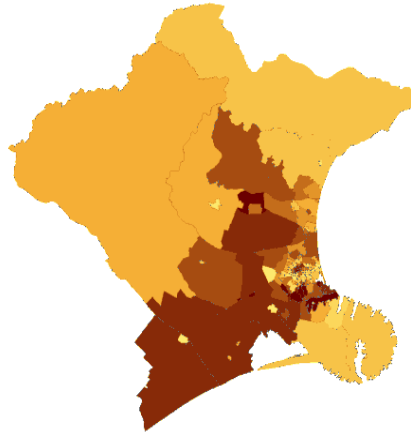
- We do not have to conduct an ecological study if using GIS!
- If we have:
 - Individual-level outcome data (eg, BMI)
 - Geographic covariate characterising exposure (eg, neighbourhood walkability)
- We could test whether one's neighbourhood walkability is associated with individual BMI
- May need to account for shared neighbourhoods through hierarchical modeling

MAUP

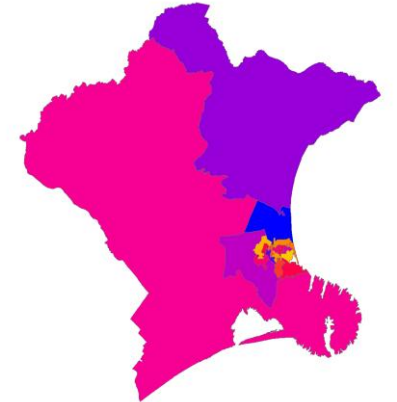
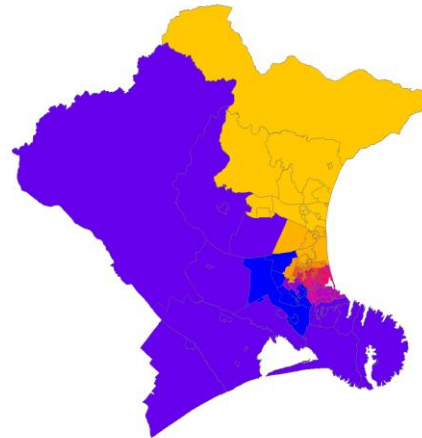
- Modifiable area unit problem
 - Unit (scale) of aggregation impacts results
- Scale effect
 - Census meshblock, Census area unit (CAU)
- Aggregation effect
 - How neighbouring units are aggregated together

MAUP

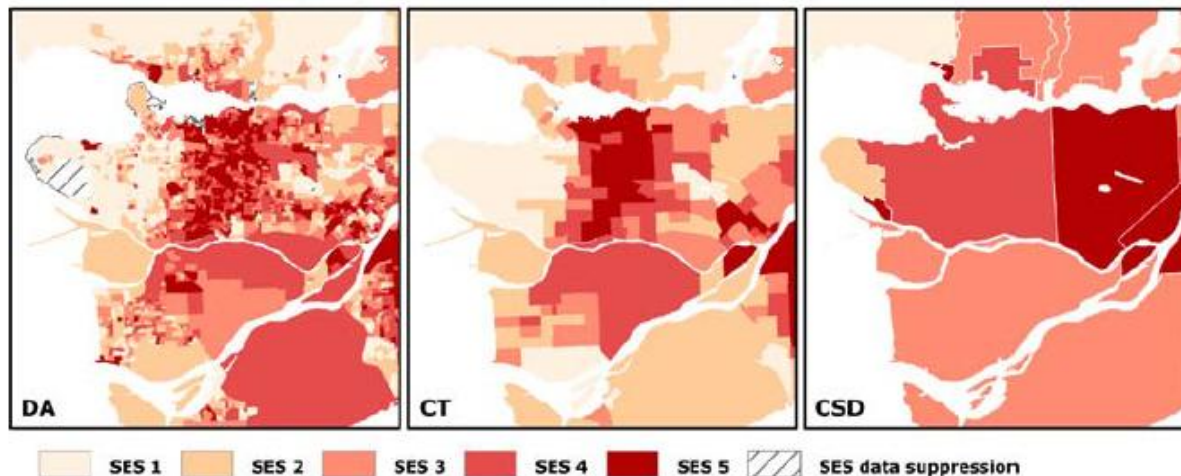
Aggregation/Grouping



Scale



MAUP – Scale



- What happens as our units become larger areas?

Shuurman N, Bell N, Dunn JR, Oliver L. Deprivation Indices, Population Health and Geography: An Evaluation of the Spatial Effectiveness of Indices at Multiple Scales, J Urban Health. 2007 84(4): 591–603.
(refer to course book)

MAUP – Aggregation/Grouping

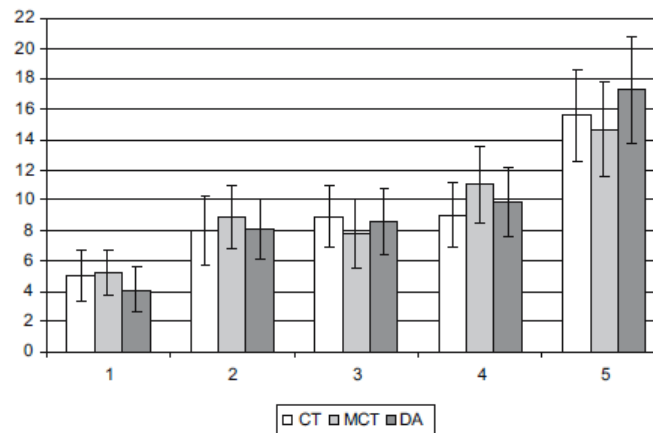


FIGURE 4. Percent “Fair or Poor” Self-rated health by quintiles of VANDIX index for Dissemination Area (DA), Census Tracts (CT), and rezoned Census Tracts (MCT). 1 = least deprived SES quintile; 5 = most deprived SES quintile.

*Note that CT and MCT are roughly same scale unit

- Do our conclusions change?

- Selection of scale of unit should be based on **theory**
- Or **existing evidence**

What is a neighbourhood?

At what scale would processes/exposures of interest be operating?

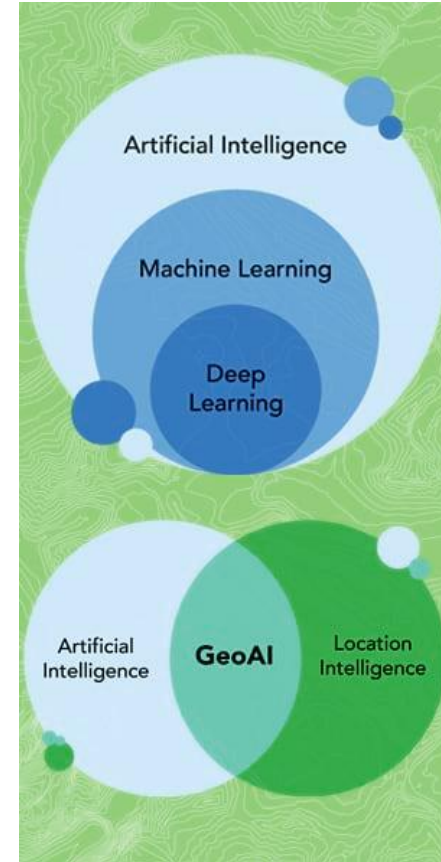
Sometimes those are not available – so one option is a sensitivity analysis

Free opportunities to continue learning

- <https://www.gislounge.com/learn-gis-for-free/>
- Visit my website for Stata code
(spacehealthandcommunity.org)

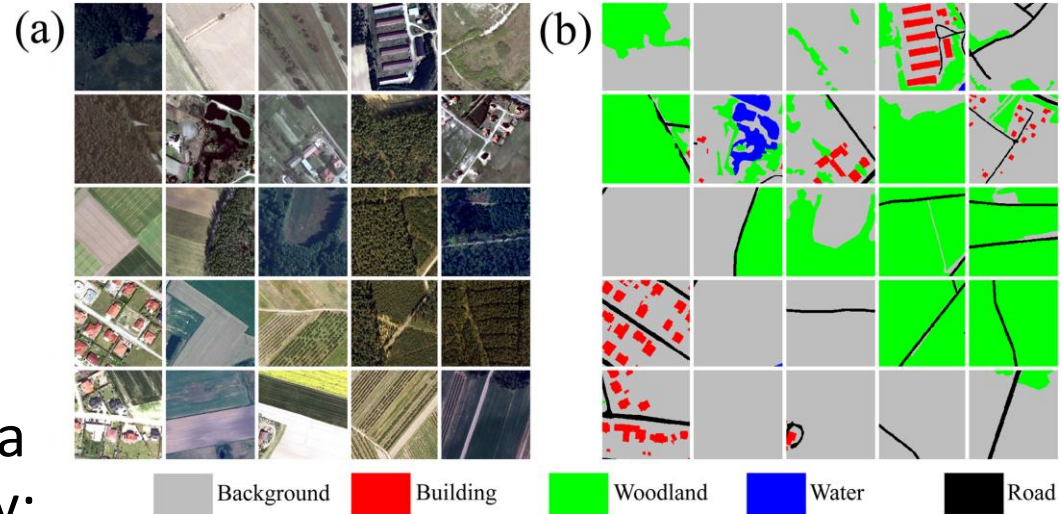
AI and geospatial developments

- Geospatial Artificial Intelligence (GeoAI)
 - Machine learning, deep learning, and computer vision within GIS
- Imagery analysis and extraction of data (eg, objection detection)
- Performing geospatial analysis



Example: semantic segmentation

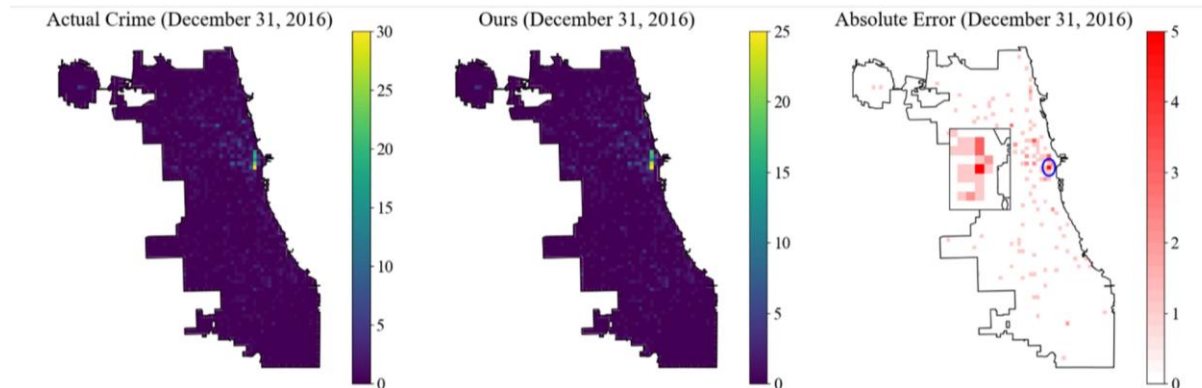
- geodl R package
- Pixel-level classification
- Satellite or streetlevel imagery can be used
- Supervised training deep learning
- Output is same imagery data where the pixel is a category: Background, building, woodland, water, or road



Source: Maxwell AE, Farhadpour S, Das S, Yang Y (2024) geodl: An R package for geospatial deep learning semantic segmentation using torch and terra. PLOS ONE 19(12):e0315127. <https://doi.org/10.1371/journal.pone.0315127>

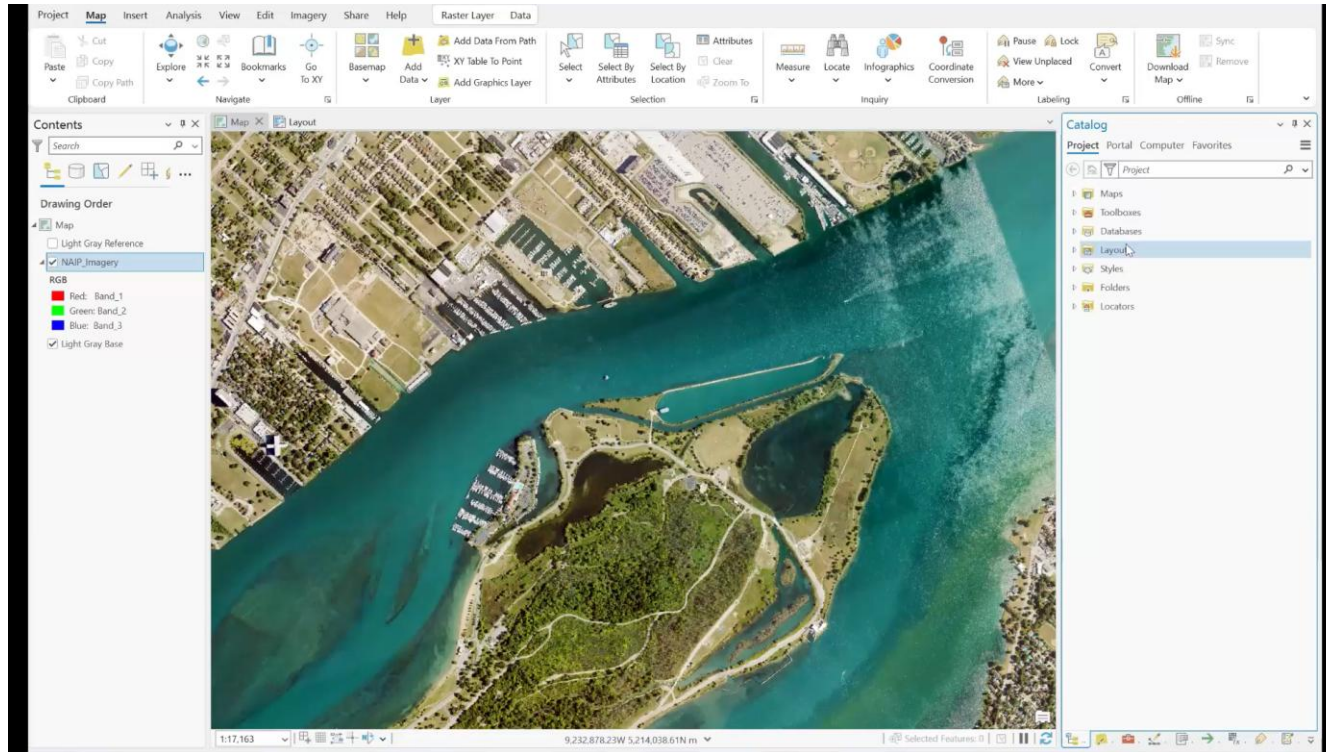
AI and geospatial developments

- Predictive modeling of crime hot spots
- Provide datasets of historic crime location, day of crime, weather, parks, etc.
- ST-ResNet with LSTM processing



Source: Shahmoradi, N., Alesheikh, A.A., Jafari, A. *et al.* Hybrid ST-ResNet and LSTM approach for precise crime hotspot prediction. *Sci Rep* 15, 40754 (2025). <https://doi.org/10.1038/s41598-025-24559-7>

AI and geospatial developments



AI and geospatial developments

- Applications:
 - Pedestrian infrastructure maintenance
 - Disaster response and prevention
- Caveats and concerns
 - Can perpetuate or exacerbate biases
 - Deepfake geography
 - Can just be wrong... need to monitor and validate

Discussion

Your own
applications?

Please feel free to
email us!

Thank you

More time to work on tutorials