BI&SemWeb'25

UCL

Day 1: Introduction to Spatial Data Analysis

Dr Anwar Musah (a.musah@ucl.ac.uk)
Lecturer in Social and Geographic Data Science
UCL Geography

Structure of this session

Today's Lecture

- An introduction to what Spatial Data Analysis is, and scientific (or geographically) theory the area is grounded on
- What are the different types of spatial data, and their features
- What are the common spatial analytical techniques
- Using RStudio as a GIS

Schedule

- Wednesday 17th (15:45-17:00): Lecture – Introduction to Spatial Data Analysis
- Thursday 18th (11:00-12:30): Practical on Spatial Data Analysis in RStudio (Part 1)
- Friday 19th (14:00-15:30): Practical on Spatial Data Analysis in RStudio (Part 2)

The goal of the practical is to introduce you to the basics of RStudio and to get you to generate some maps in a programmatic

What is Spatial Data Analysis?

Definition:

The field of Spatial Statistics (or Analysis) is built on the assertation that nearby geographical observations (or objects) are somewhat associated in someway in space.

- The field is interdisciplinary as it brings to together theories from **statistics, geography, computer science** and integrates **evidence-based research** methodologies (i.e., study designs)

- Usage for describing spatial distribution of **areal/point/gridded** outcomes, as well as **interactions** between objects in space, but also how **an object has an impact on other nearby objects in geographic space**.

First Law of Geography [1]

Waldo Rudolph Tobler (1930 – 2018)

Tobler's First Law of Geography on "everything is related to everything else, but near things are more related to distant things"

This first law is the foundation of the fundamental concepts of spatial **dependence.** It is integrated to most of the families of models in spatial statistics

Very important concepts to keep in mind when you think about **spatial dependence:**

- **Spatial Autocorrelation**
- **Distance Decay**
- **Spatial Spillover**

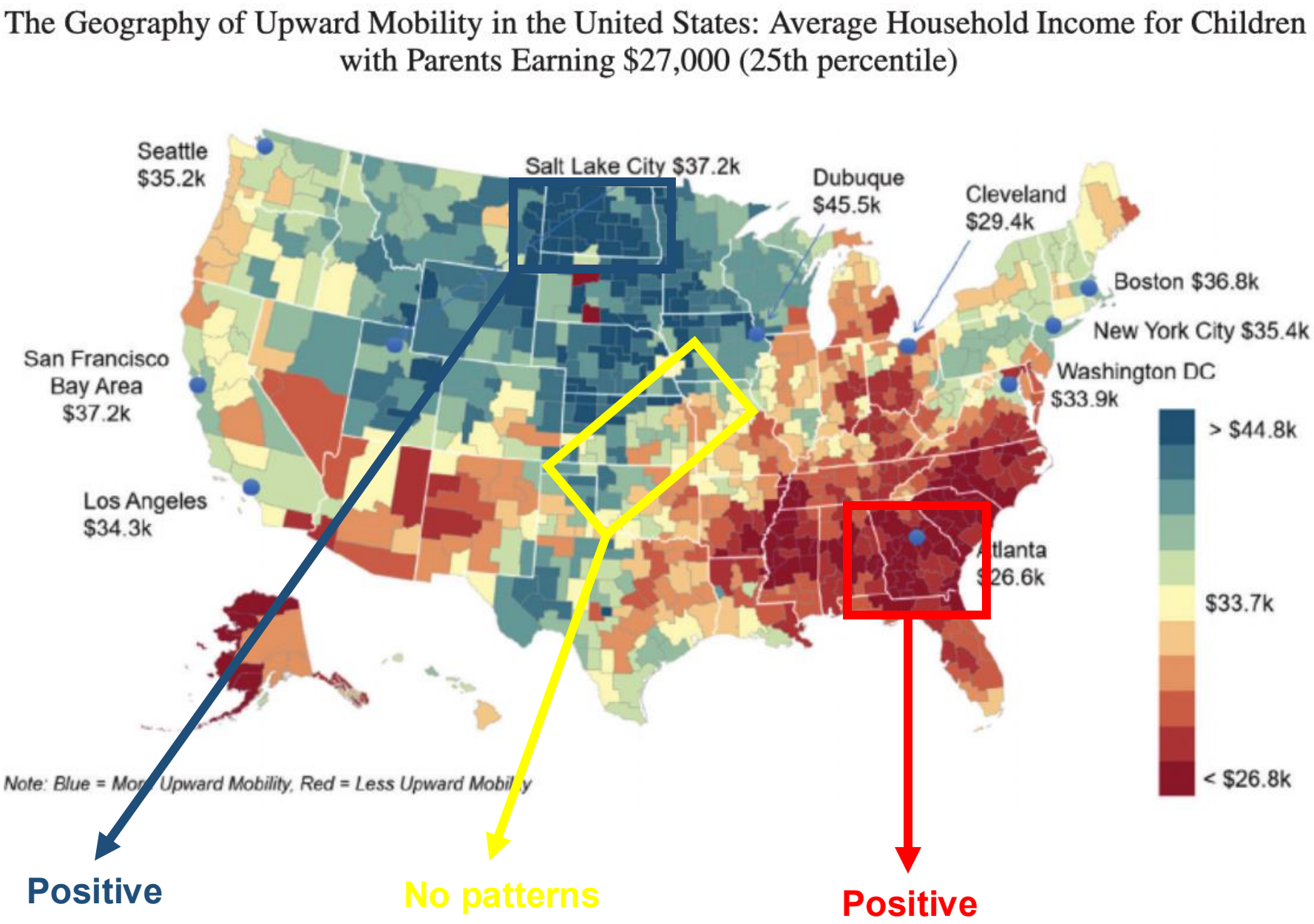
What is **Spatial autocorrelation** in the context of Tobler's theory?

"This simply refers to whether (or not) similar values cluster together over geographic space"

Etymology ("**Auto**" means self; and "**correlation**" means the degree of relative correspondence)

Similar values that cluster together are said to have **positive spatial autocorrelation (& spatially dependent)**

Random patterns or values that cluster together are said to have **no spatial autocorrelation (and thus no dependence)**



What is **Distance decay** in the context of Tobler's theory?

"This refers to the distance between two events (or objects) and their degree of interaction in space as distance vary."

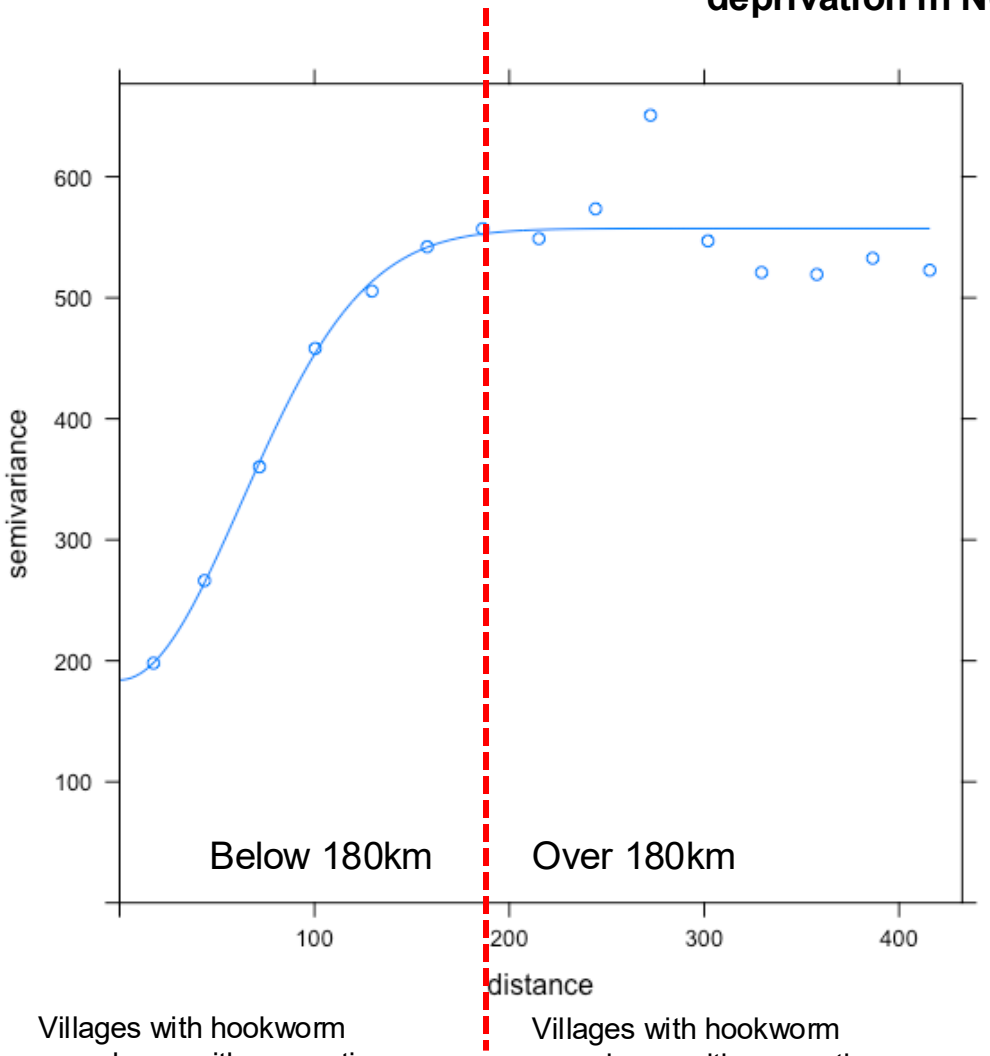
In short, what distance decay (in the context of Tobler's theory) says**:**

- It dictates how objects interact. The larger the distance between two events in a space – the less is their interaction & *vice versa*

Examples of where Distance decay manifests in studies cities and urban space:

- As distance from the focal point of a city center increases – high population density, taller buildings, and accessibility to multiple modes of transport decreases
- Housing marketing – house prices decreases, and in turn, residential mobility increases from expensive to affordable areas

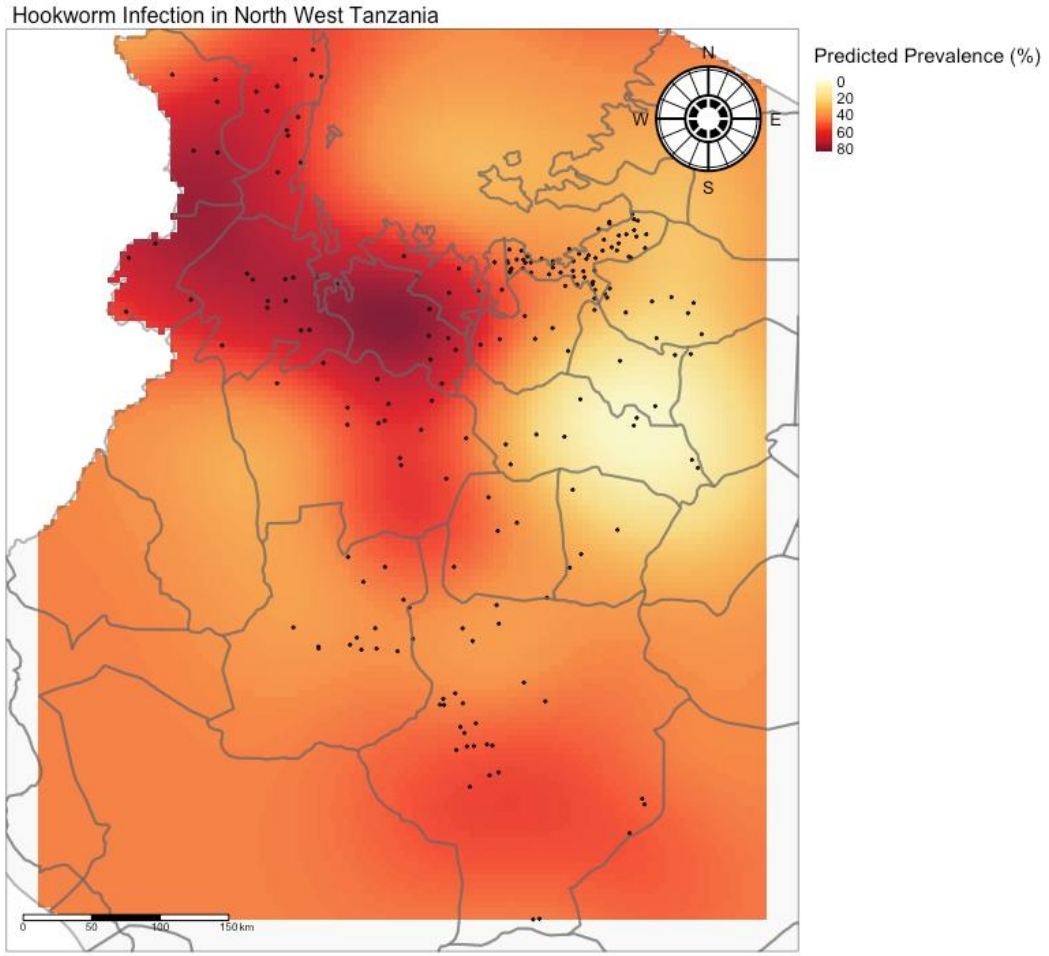
Using Kriging to spatially predict areas with intense hookworm infection associated with socioeconomic deprivation in Northwestern Tanzania



Villages with hookworm prevalence with separation distance below 180km have similar prevalence

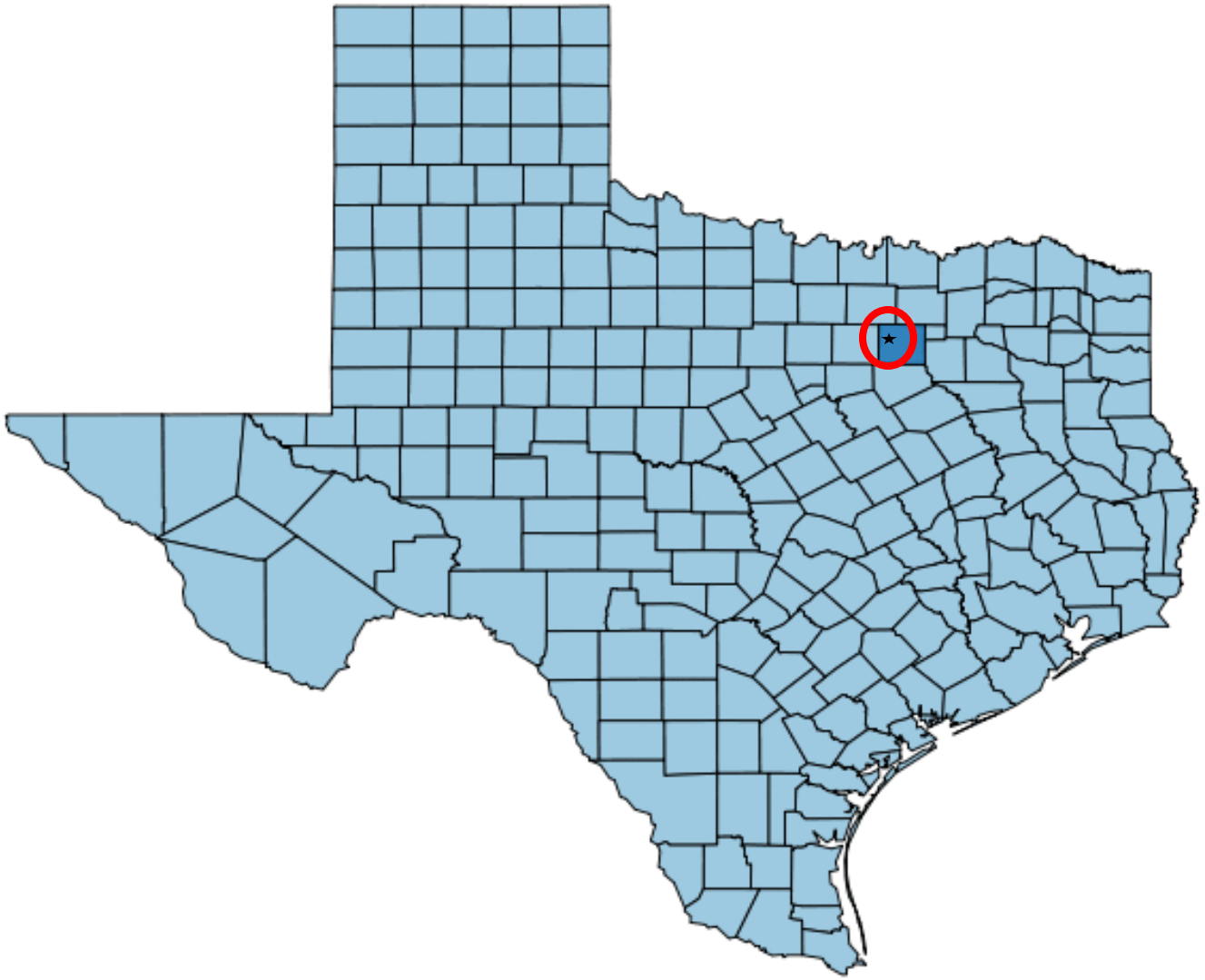
Villages with hookworm prevalence with separation distance 180km and over don't have related prevalence

We used the concept of distance decay, or in this 'separation distance' to account for spatial dependence between our survey points



We have survey points (of villages) reporting prevalence of hookworm in Northwestern Tanzania. We assumed there some spatial dependence in the prevalence of hookworm in these locations to predict prevalence where there are no information (or points). We use this information on form the left panel to build our model for making geostatistical predictions.

Spatial Spillover: "…where you are in geographic space matters" an event in one location can someway have an impact on other events in neighboring areas

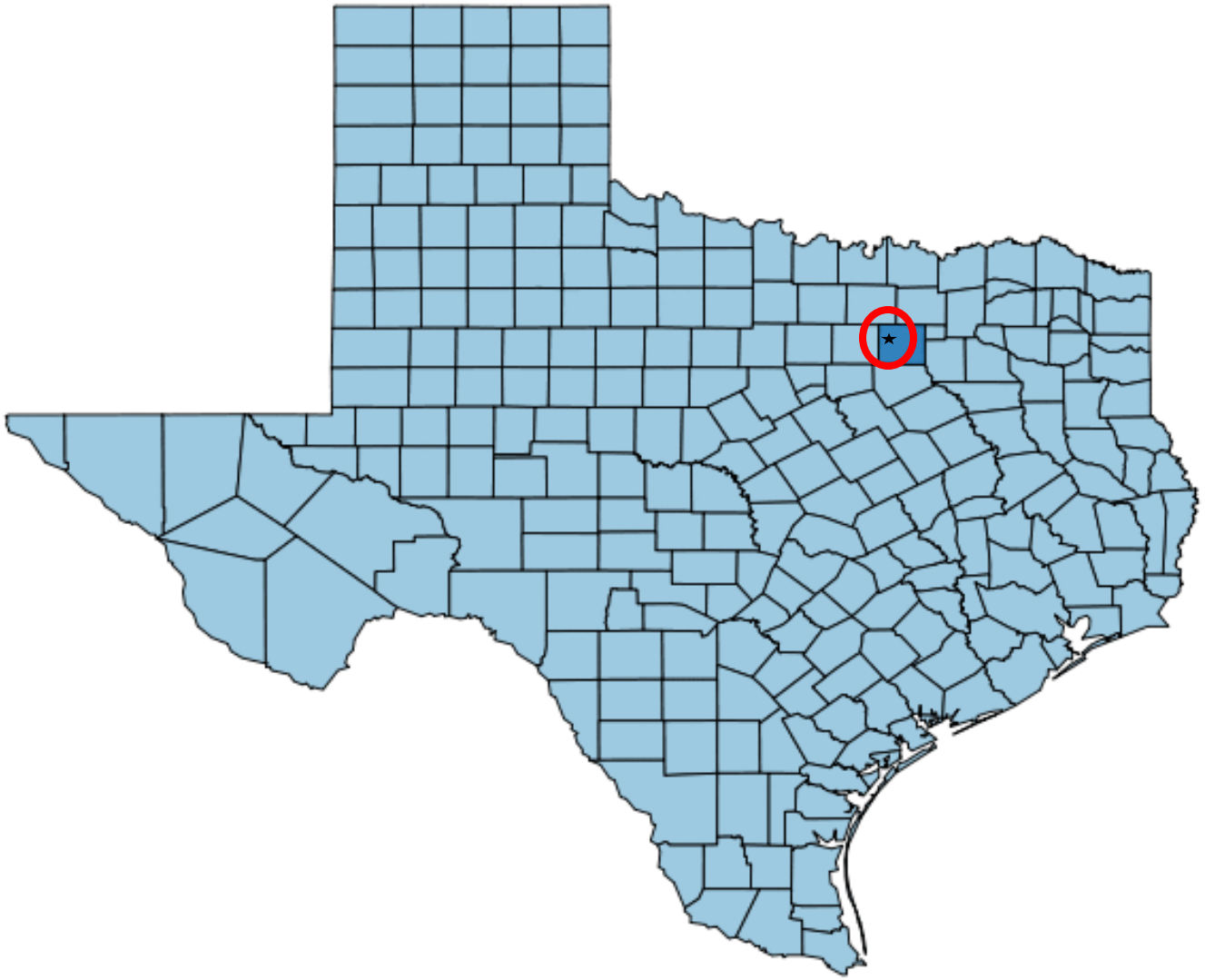


In Spatial Statistics, especially in Spatial Regression, Geostatistics – we often try to account for **spatial spillover** effects in our models

There are four types of spillovers:

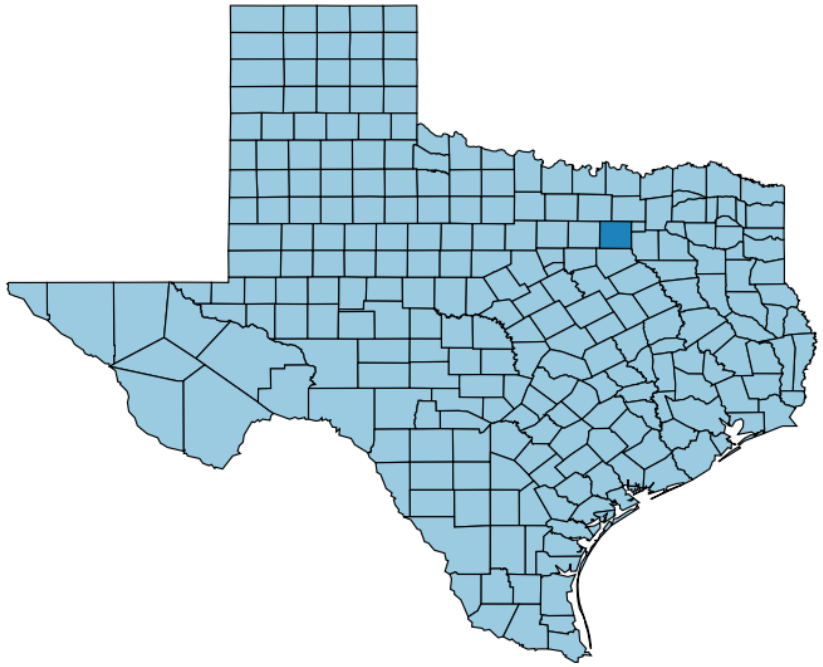
- No spillover
- Local spillover
- Global spillover
- Rippling spillovers

Spatial Spillover: "…where you are in geographic space matters" an event in one location can someway have an impact on other events in neighboring areas



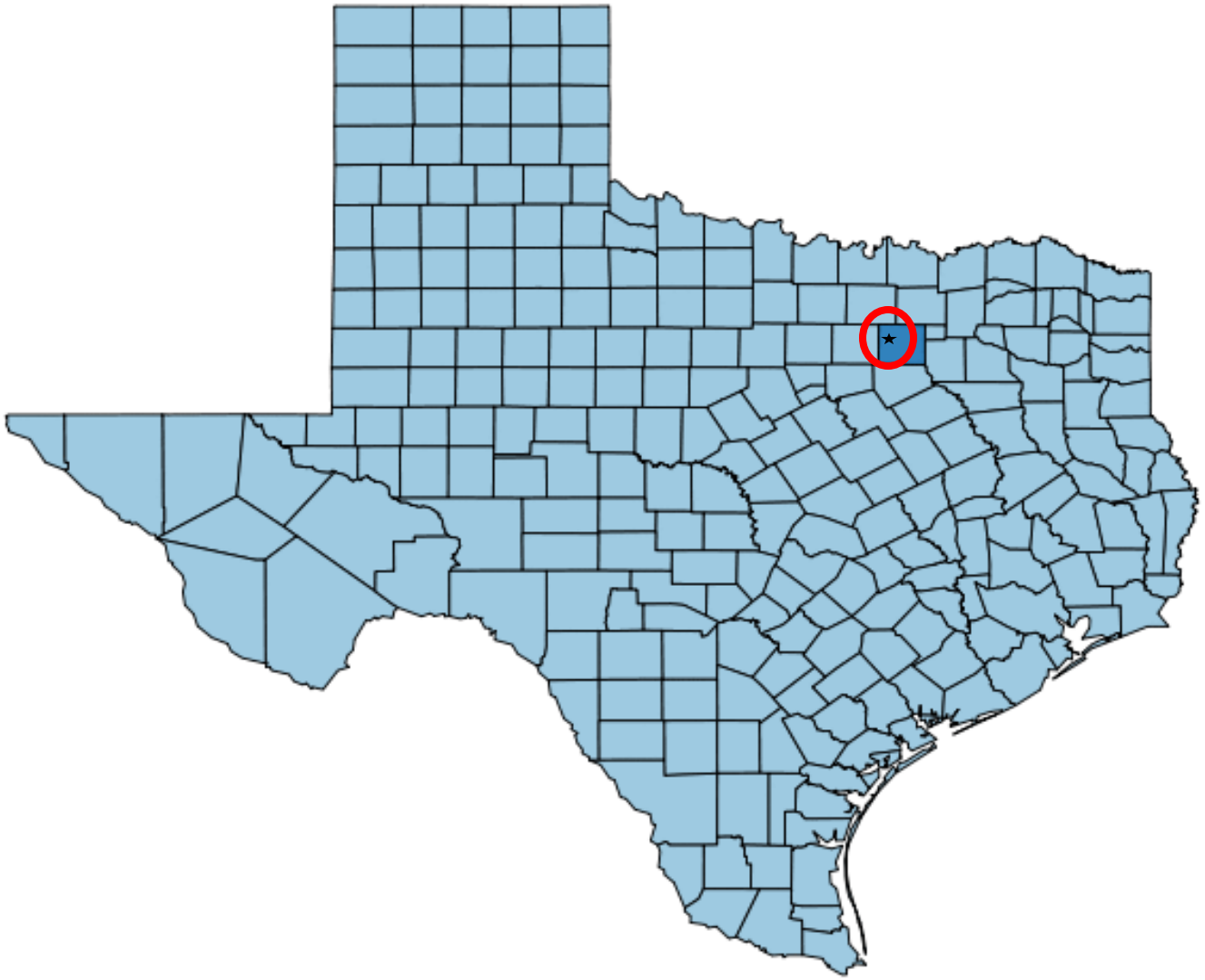
In Spatial Statistics, especially in Spatial Regression, Geostatistics – we often try to account for **spatial spillover** effects in our models

No spillover (independence)



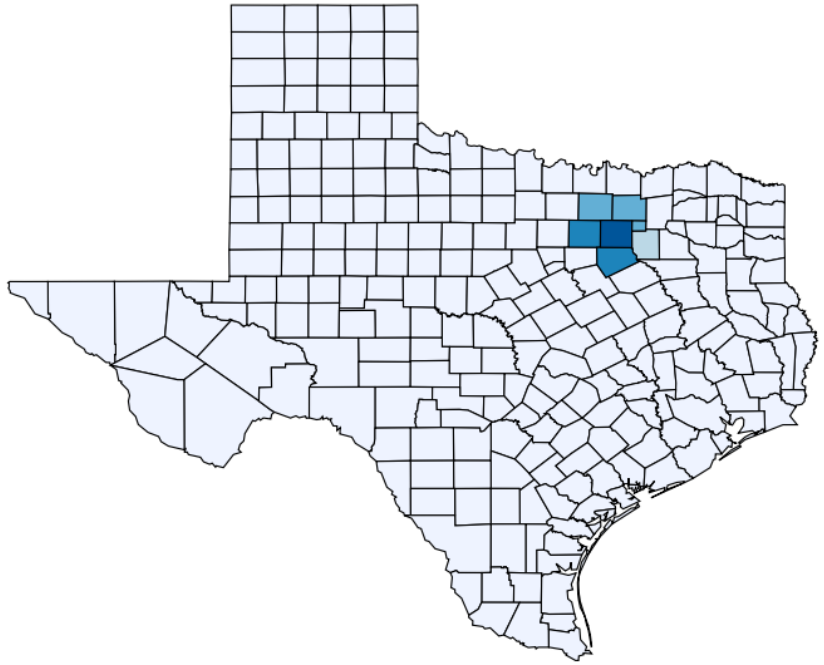
A change in unemployment in that dark blue area will only have an impact on crime rates within its own area, but does not influence crime rates in neighboring areas

Spatial Spillover: "…where you are in geographic space matters" an event in one location can someway have an impact on other events in neighboring areas



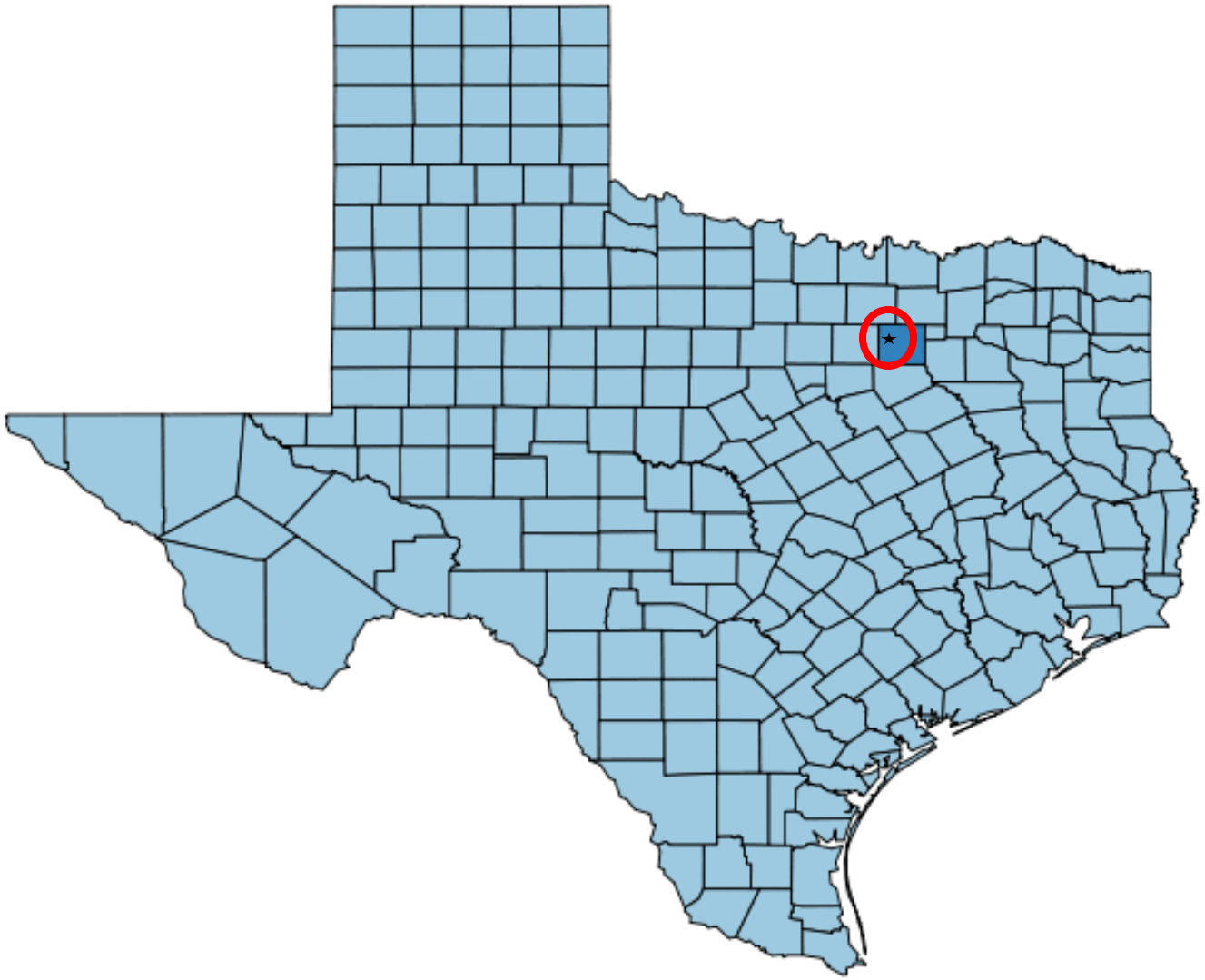
In Spatial Statistics, especially in Spatial Regression, Geostatistics – we often try to account for **spatial spillover** effects in our models

Local spillover (dependence)



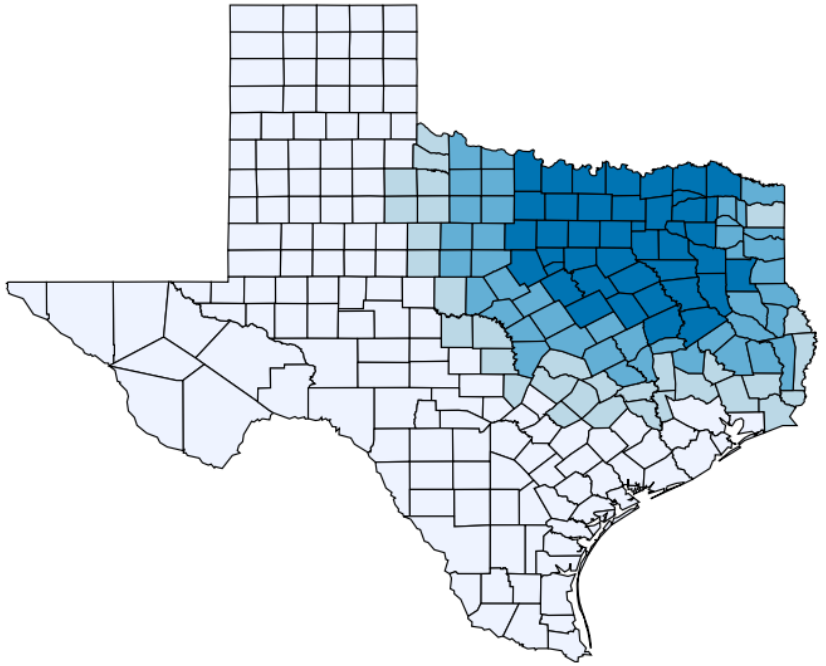
A change in unemployment in that dark blue area (in the center) will not only have an impact on crime rates within its own area, but it will also have a direct influence on crime rates in neighboring areas only.

Spatial Spillover: "…where you are in geographic space matters" an event in one location can someway have an impact on other events in neighboring areas



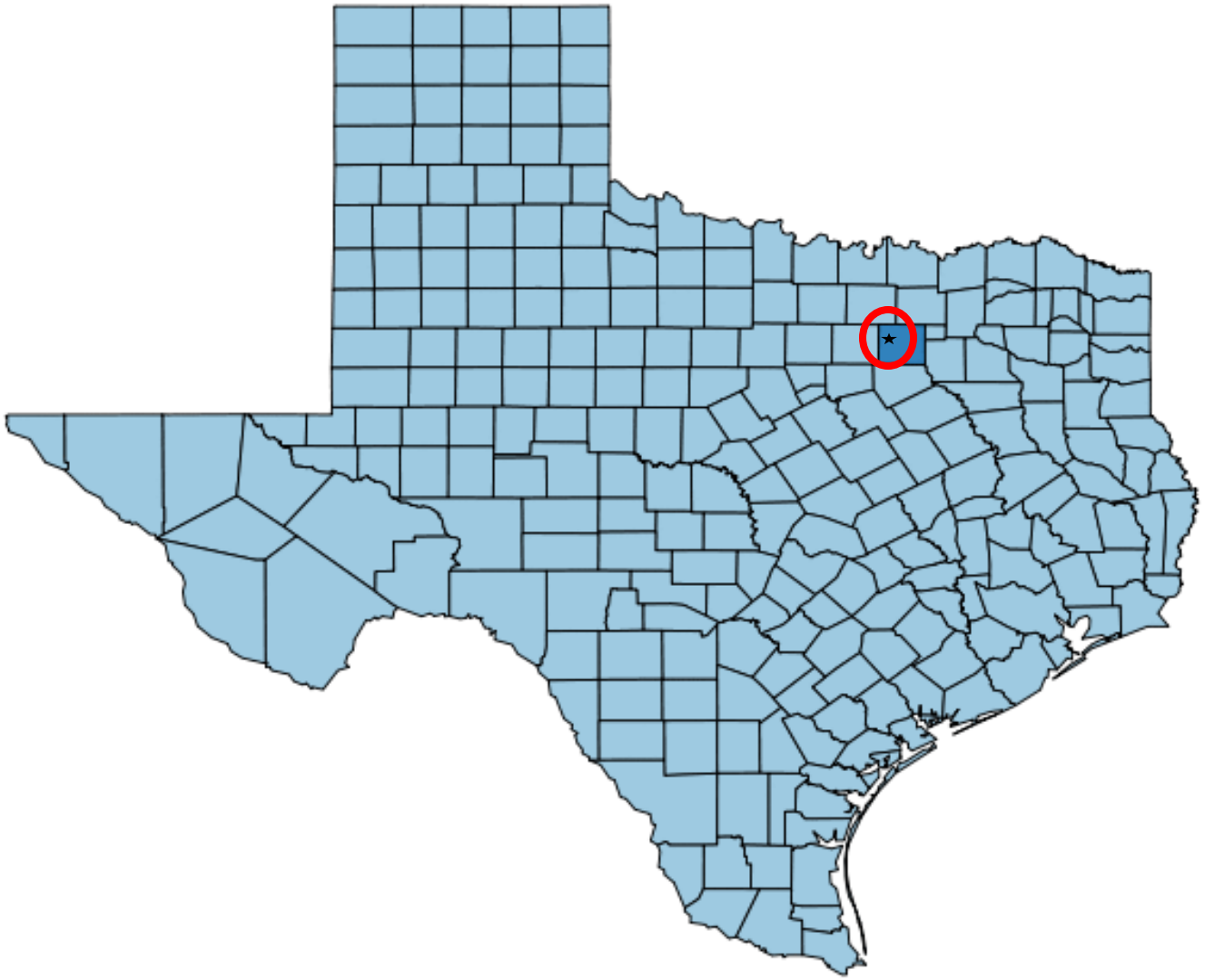
In Spatial Statistics, especially in Spatial Regression, Geostatistics – we often try to account for **spatial spillover** effects in our models

Global spillover (dependence)



A change in unemployment in that dark blue area (in the center) will not only have an impact on crime rates within its own area, but it will also have a wider influence on crime rates beyond its direct neighbors.

Spatial Spillover: "…where you are in geographic space matters" an event in one location can someway have an impact on other events in neighboring areas



In Spatial Statistics, especially in Spatial Regression, Geostatistics – we often try to account for **spatial spillover** effects in our models

Rippling spillover (dependence)

Where there is a focal point for an event and its influence may have a rippling (or trickle down) effect across space, triggering other events, which then diminishes with time and distance.

e.g.,

- Natural disasters - an earthquake and building destruction.
- Disease spread and outbreaks
- World financial markets and crashes

What is Spatial Data & its Features?

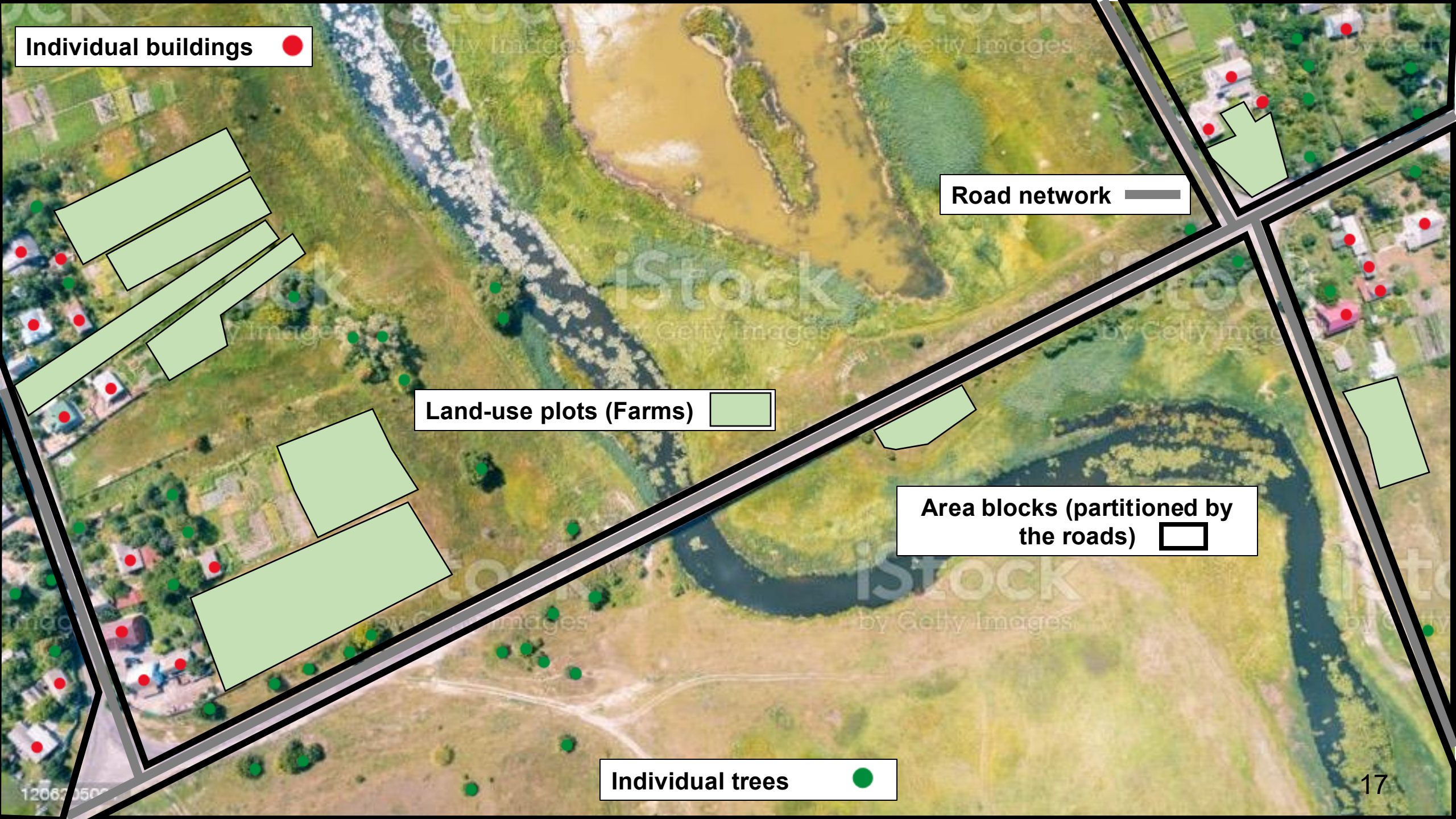
Suppose we want to map the following from this landscape:

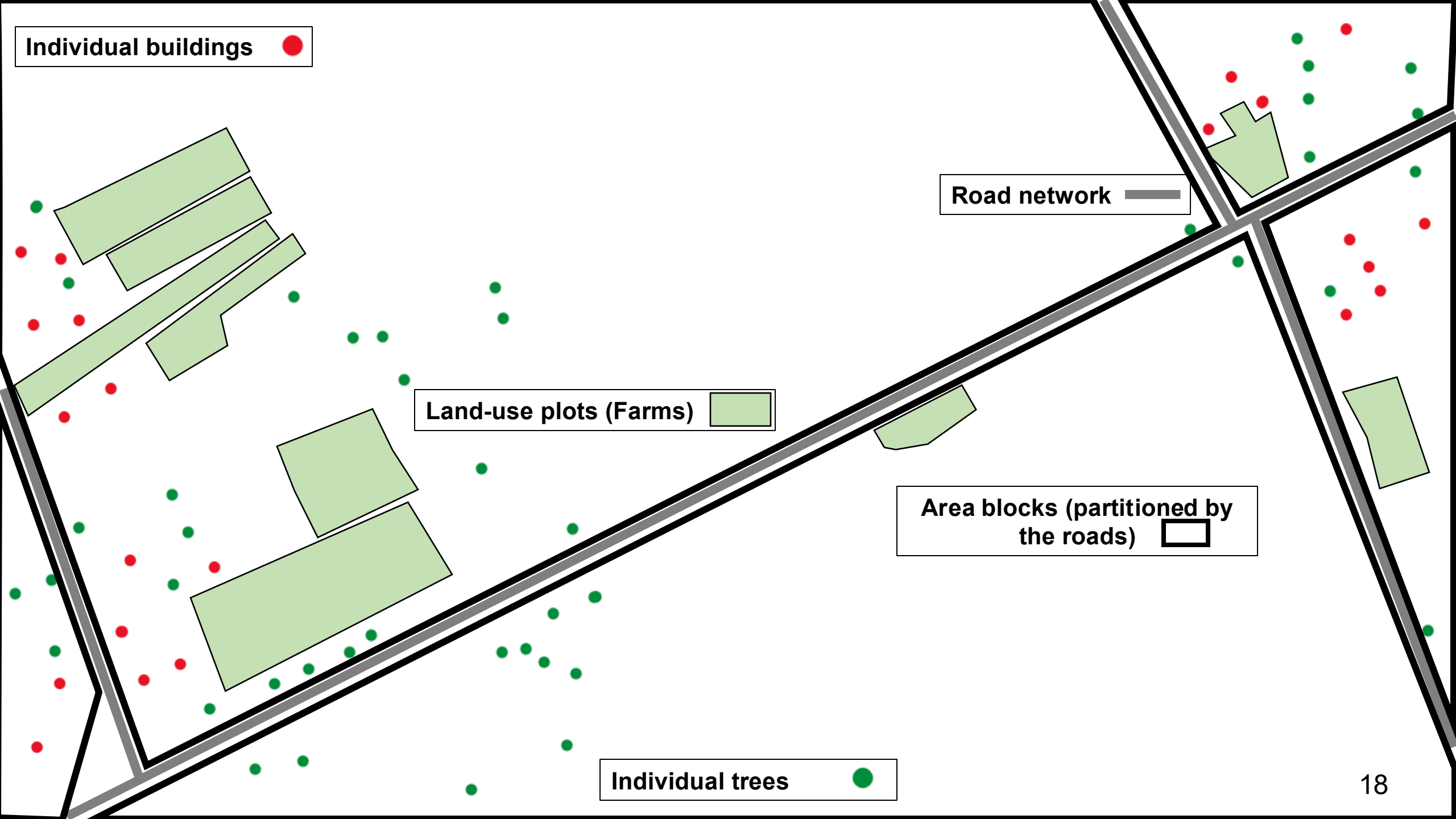
1. Physical objects:

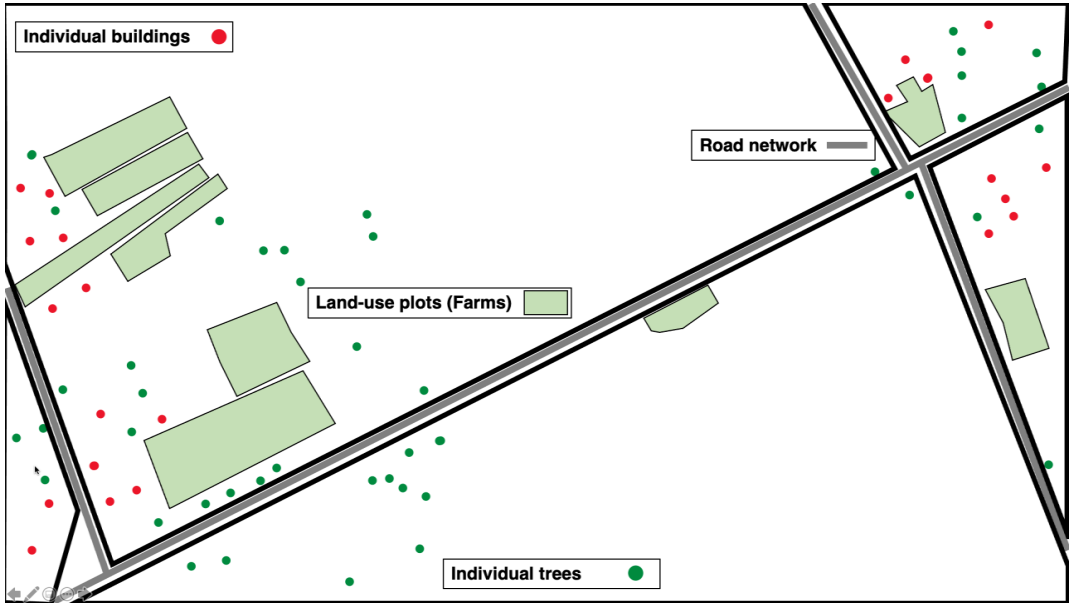
- Location of buildings
- Farm plots
- Locations of trees
- Road network
- Block areas (divided by the road)

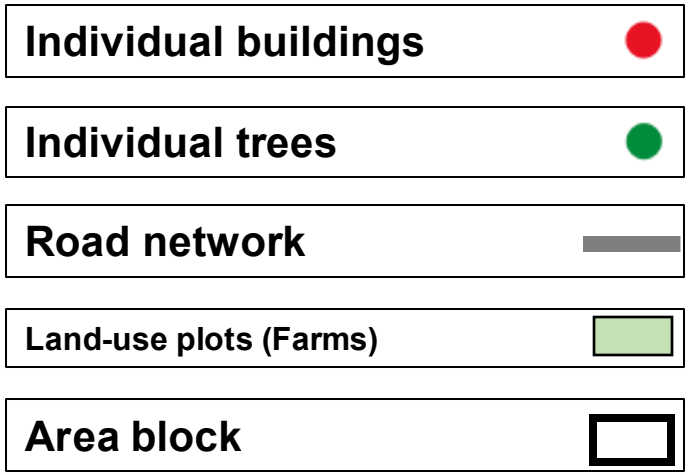
2. Levels of soil moisture across the landscape


Individual buildings
Road network
Land-use plots (Farm)
Area blocks (partitioned by the roads)
Individual trees


Individual buildings
Road network
Land-use plots (Farms)
Area blocks (partitioned by the roads)
Individual trees


Individual buildings
Road network
Land-use plots (Farms)
Area blocks (partitioned by the roads)
Individual trees





The above objects listed are called "**Features**". A feature can be described according to its characteristics which is termed an "**Attribute**" in GIS. The attribute of a feature can be a **numeric** or **text** observation.

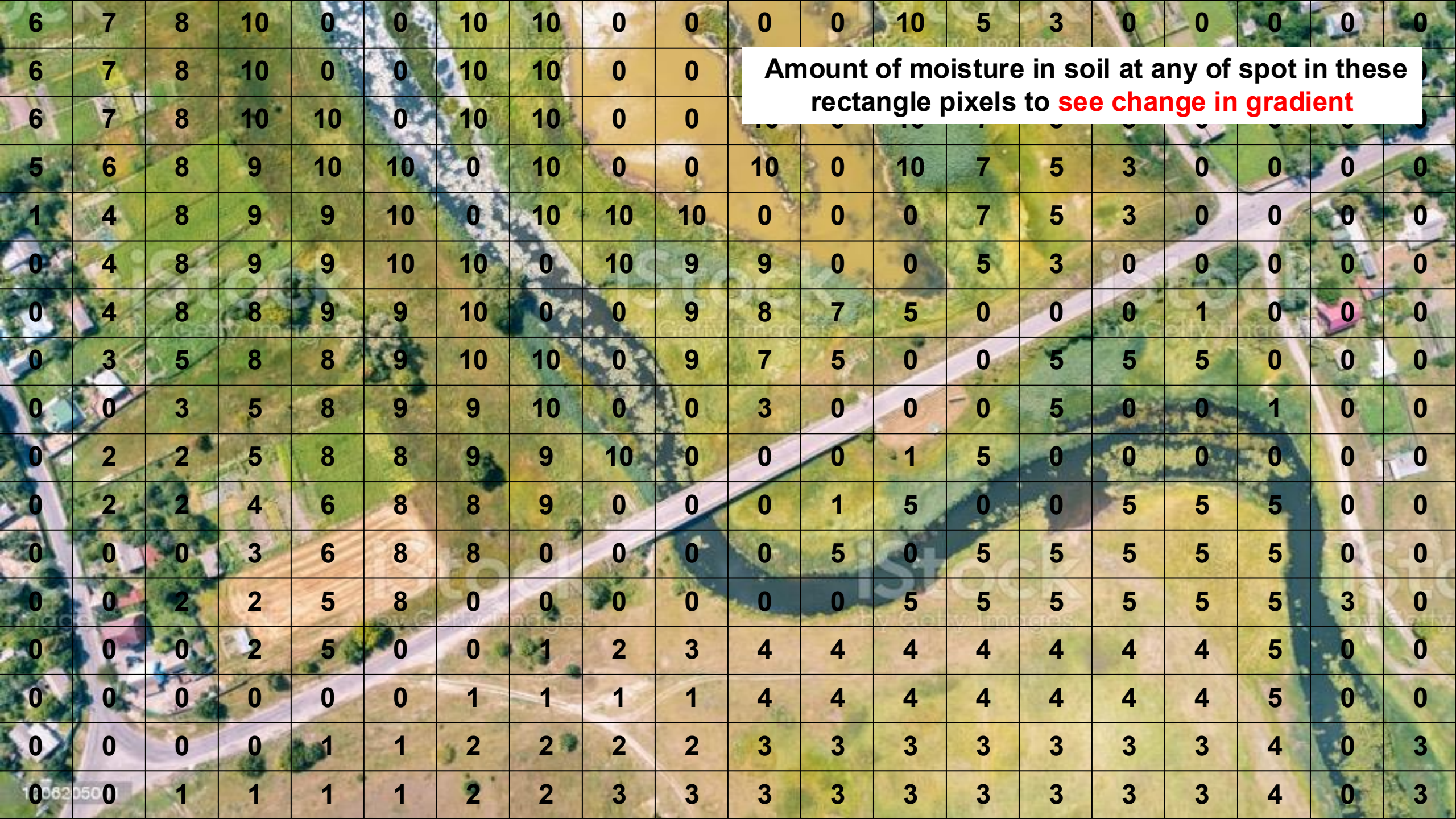
For example:

- A building is a **point feature** on this map, the number of people living a building is a **numeric attribute** describing this **feature**. Type of building (i.e., Victorian or modern) is a **text attribute**

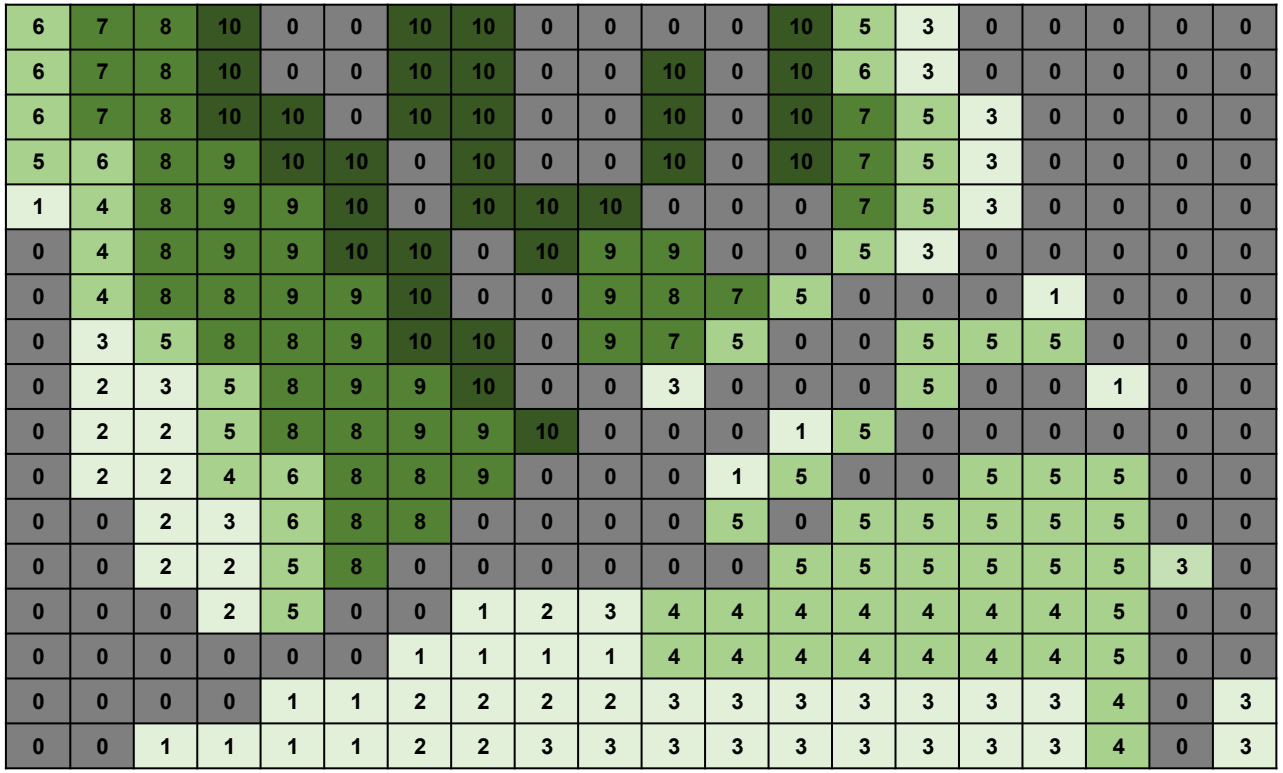
- The road network is a **polyline feature**, the length (or distance (m)) of the road is a **numeric attribute** describing the road

- Land-use plot is a **polygon (or area) feature**, the type of land-use (i.e., farming) is the **text attribute** describing what that polygon is etc.


What about the soil moisture



6	7	8	10	0	0	10	10	0	0	0	0	10	5	3	0	0	0	0	0
6	7	8	10	0	0	10	10	0	0										
6	7	8	10	10	0	10	10	0	0										
5	6	8	9	10	10	0	10	0	0	10	0	10	7	5	3	0	0	0	0
1	4	8	9	9	10	0	10	10	10	0	0	0	7	5	3	0	0	0	0
0	4	8	9	9	10	10	0	10	9	9	0	0	5	3	0	0	0	0	0
0	4	8	8	9	9	10	0	0	9	8	7	5	0	0	0	1	0	0	0
0	3	5	8	8	9	10	10	0	9	7	5	0	0	5	5	5	0	0	0
0	0	3	5	8	9	9	10	0	0	3	0	0	0	5	0	0	1	0	0
0	2	2	5	8	8	9	9	10	0	0	0	1	5	0	0	0	0	0	0
0	2	2	4	6	8	8	9	0	0	0	0	1	5	0	0	5	5	5	0	0
0	0	0	3	6	8	8	0	0	0	0	0	5	0	5	5	5	5	5	0	0
0	0	2	2	5	8	0	0	0	0	0	0	0	5	5	5	5	5	5	3	0
0	0	0	2	5	0	0	1	2	3	4	4	4	4	4	4	4	5	0	0
0	0	0	0	0	0	1	1	1	1	4	4	4	4	4	4	4	5	0	0
0	0	0	0	1	1	2	2	2	2	3	3	3	3	3	3	3	4	0	3
0	0	1	1	1	1	2	2	3	3	3	3	3	3	3	3	3	4	0	3



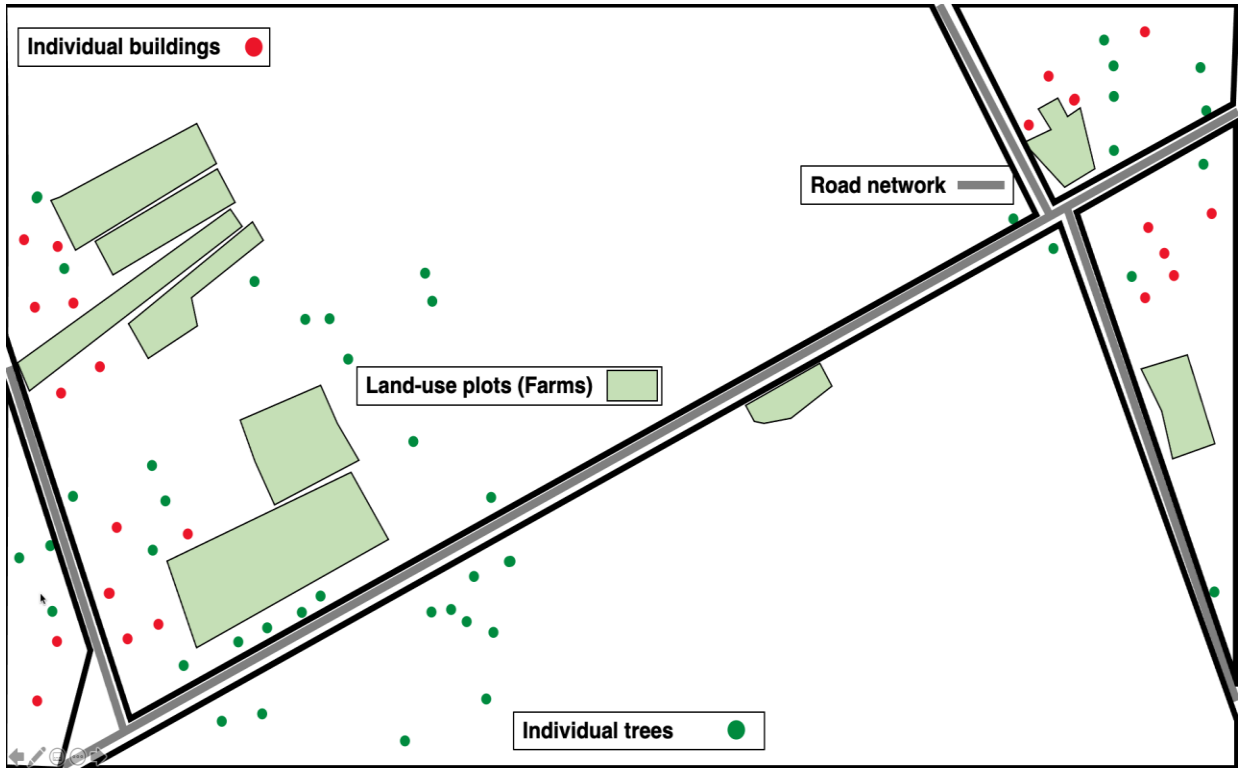
Soil moisture index
10
7 to 9
4 to 6
1 to 3
0

Unlike the vector data. The above feature describes how moisture levels across the surface of the landscape – the feature is is not measured discretely but on a **continuous** surface to show gradient in changes for soil moisture across the landscape

Now, this **Non-discrete** feature is classed a **Raster Data**

What is Raster Data?

- **It is a matrix of pixels or grid-cells** that contains a numeric or text value for a feature its representing
- It is composed of **rows and columns**
- Each pixel or grid-cell has a **resolution** (or **size for height and width**)



6	7	8	10	0	0	10	10	0	0	0	0	10	5	3	0	0	0	0	0
6	7	8	10	0	0	10	10	0	0	10	0	10	6	3	0	0	0	0	0
6	7	8	10	10	0	10	10	0	0	10	0	10	7	5	3	0	0	0	0
5	6	8	9	10	10	0	10	0	0	10	0	10	7	5	3	0	0	0	0
1	4	8	9	9	10	0	10	10	10	0	0	0	7	5	3	0	0	0	0
0	4	8	9	9	10	10	0	10	9	9	0	0	5	3	0	0	0	0	0
0	4	8	8	9	9	10	0	0	9	8	7	5	0	0	0	1	0	0	0
0	3	5	8	8	9	10	10	0	9	7	5	0	0	5	5	5	0	0	0
0	2	3	5	8	9	9	10	0	0	3	0	0	0	5	0	0	1	0	0
0	2	2	5	8	8	9	9	10	0	0	0	1	5	0	0	0	0	0	0
0	2	2	4	6	8	8	9	0	0	0	1	5	0	0	5	5	5	0	0
0	0	2	3	6	8	8	0	0	0	0	5	0	5	5	5	5	5	0	0
0	0	2	2	5	8	0	0	0	0	0	0	5	5	5	5	5	5	3	0
0	0	0	2	5	0	0	1	2	3	4	4	4	4	4	4	4	5	0	0
0	0	0	0	0	0	1	1	1	1	4	4	4	4	4	4	4	5	0	0
0	0	0	0	1	1	2	2	2	2	3	3	3	3	3	3	3	4	0	3
0	0	1	1	1	1	2	2	3	3	3	3	3	3	3	3	3	4	0	3

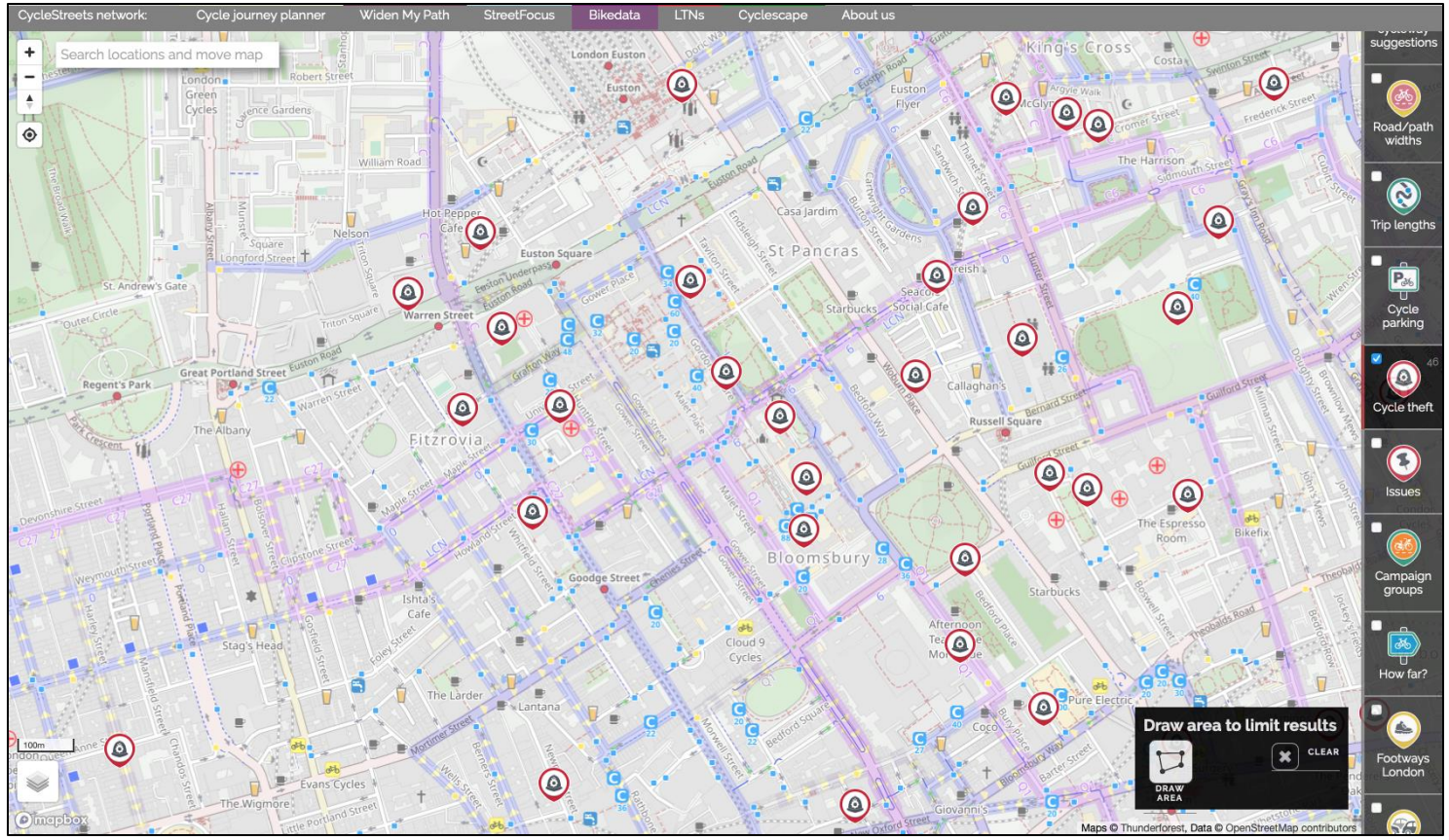
Vector data

Raster data

These are discrete spatial data that are stored as a shapefile format (.shp)

These are continuous gridded spatial data that are stored as a GEOTIFF format (.tiff)

Point Pattern Data (PPD)



Represents point locations of bicycle thefts in Central London area
Source: BikeData.CycleStreets network https://bikedata.cyclestreets.net/

Key Characteristics

The main interest is the occurrences of an event at a points (or points). These events occur at “random” at any given geographic space and time.

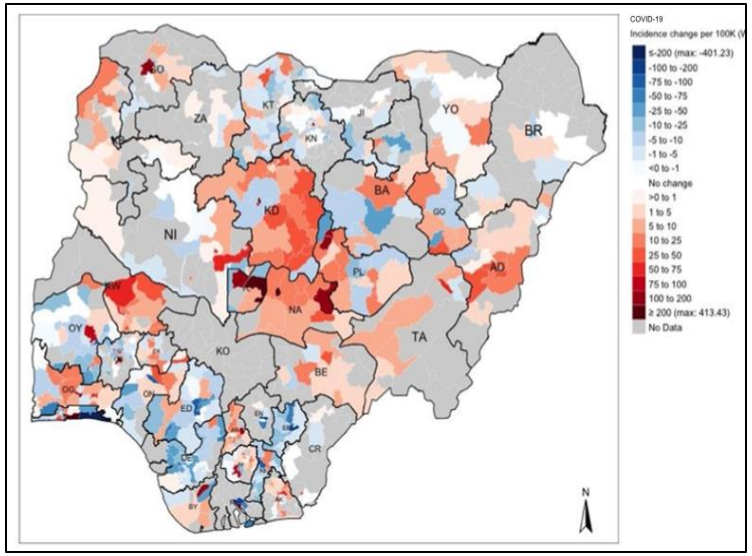
Examples of PPDs events (or outcomes):

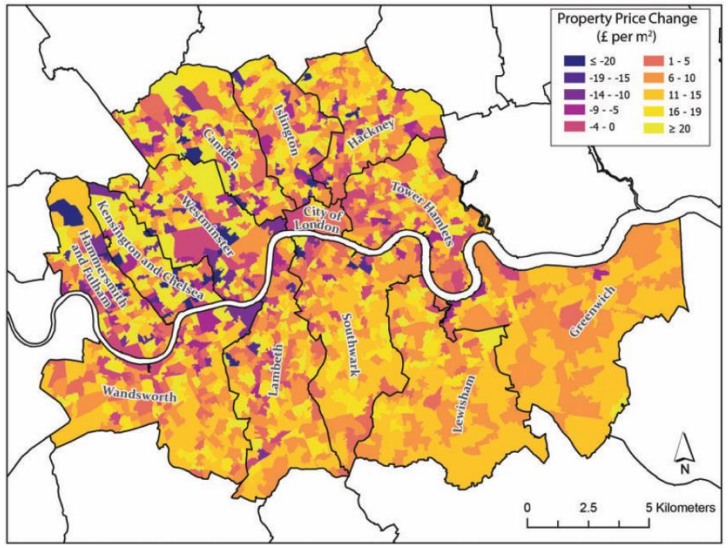
- Point locations of burglaries
- Riots
- Locations of car collisions etc.,
- Locations of where an adult tree needs to be replanted

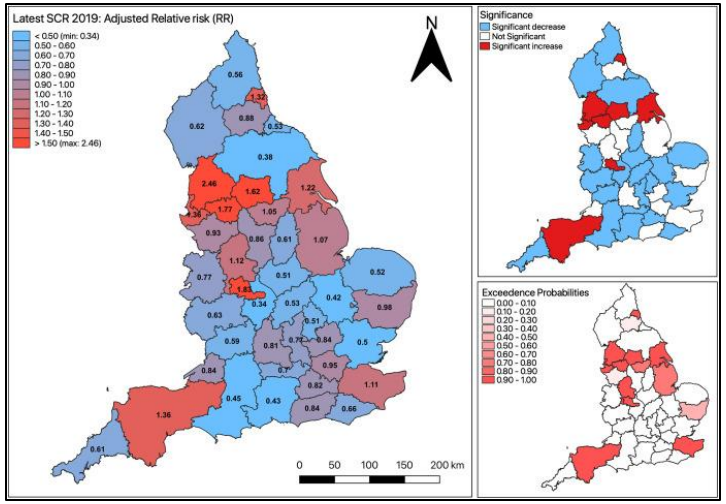
Some PPDs events may carry additional information that may describe the occurrence of an observed event (or outcome):

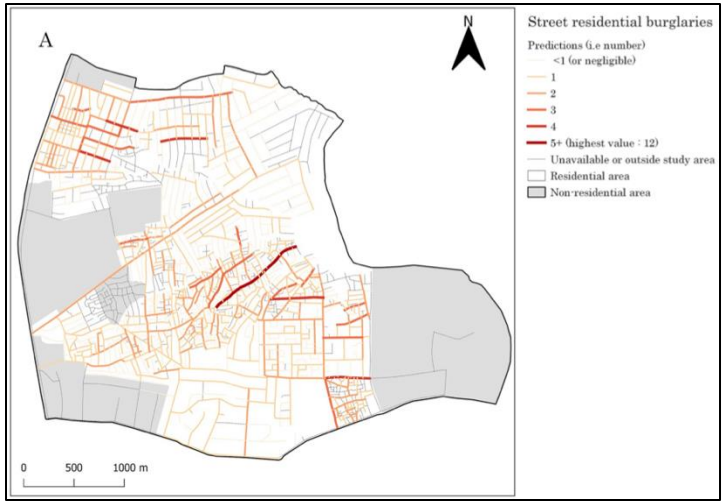
- Burglary: Type of premise that was burgled, time of day the burglary occurred etc.,
- Car collision: type of road, weather condition etc.

Aggregated Data









Key Characteristics

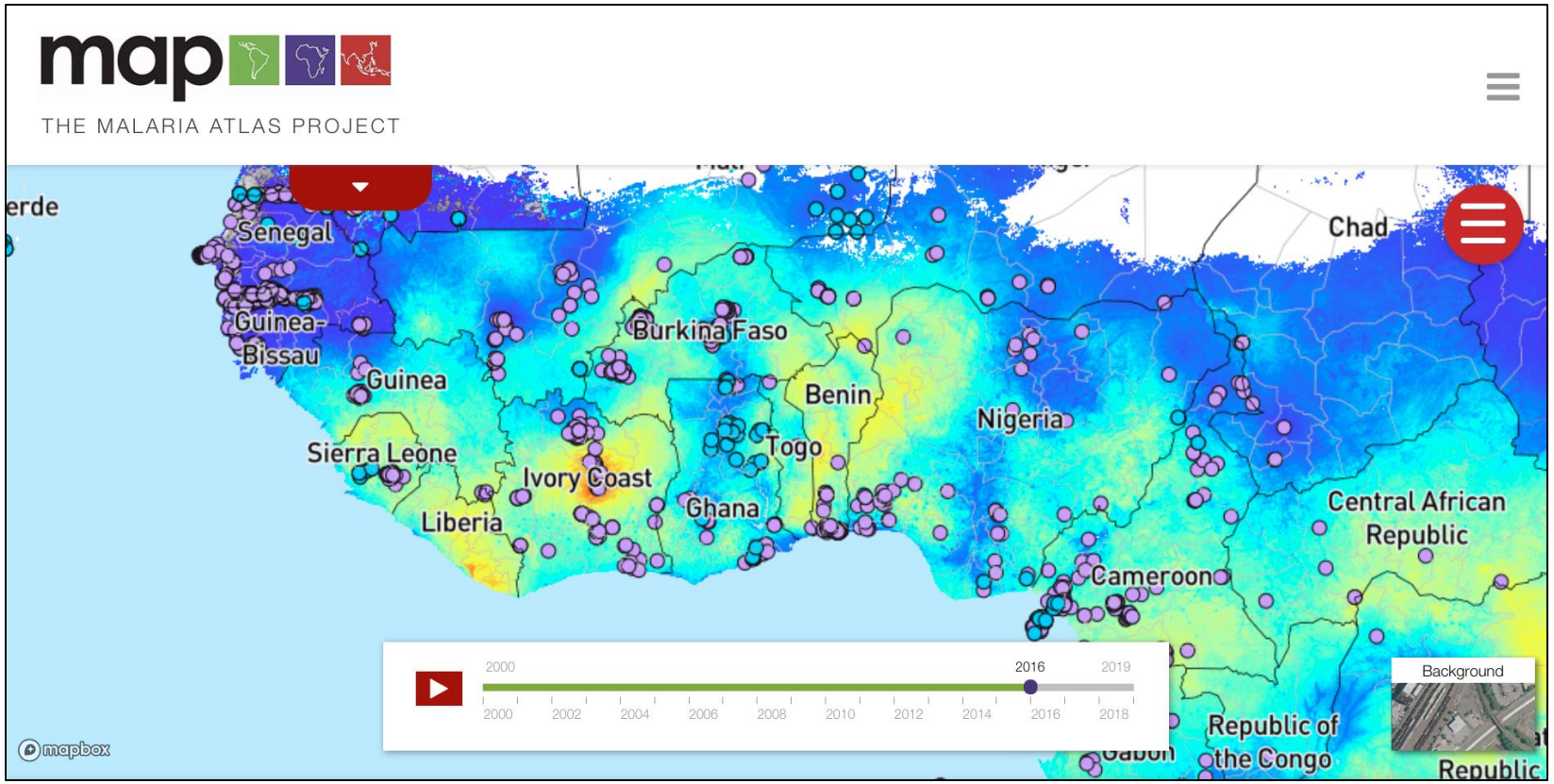
The main interest the quantity of interest defined for line segments, areas, or regions.

Events (or outcomes) that are aggregated measures to areas:

- Prevalence of a disease in areas
- Population density in a county
- Regional unemployment rates
- Risk of an outcome

Sources:
Musah, A., *et al*. (2020): https://doi.org/10.1016/j.apgeog.2019.102126
Todd, J., *et al*. (2021): https://doi.org/10.1177/23998083211001836
Li, L., *et al*. (2022): https://doi.org/10.1016/j.apgeog.2022.102718
Elimian, K., *et al*. (2022): http://dx.doi.org/10.1136/bmjopen-2022-063703

Geostatistical Data



Sampled points are surveys on prevalence of malaria, which were used to make survey predication of prevalence at unsampled areas in Sub-Saharan Africa

Source: The Malaria Atlas Project https://malariaatlas.org

Key Characteristics

- The quantity of interest has a value at any location across a given area.
- These are values over a grid/raster

Events are usual statistical prediction determined from "**sampled**" points with continuous data values:

- Land surface elevation
- Diffusion of ambient air pollutants
- Environmental suitability for breeding habitats of mosquitoes
- Community surveys pertained to disease burden

Non-spatial context & data structure:

	ATTRIBUTE			
	Variable 1	**Variable 2**	$\cdots$	**Variable n**
Entity 1	$attribute_{11}$	$attribute_{12}$	$\cdots$	$attribute_{1n}$
Entity 2	$attribute_{21}$	$attribute_{22}$	$\cdots$	$attribute_{2n}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Entity m	$attribute_{m1}$	$attribute_{m2}$	$\cdots$	$attribute_{mn}$

To apply some spatial analysis to data – you must have some variable that defines the entity's geographic location. This can be **GPS coordinates**, **spatially referenced geometries** of an area

NOTES: It is not enough to have the just the name of the area(s). It must be some geometric entry!

- Attributes that defines an entity's location are typically excluded from the analysis
- The conventional statistical methods, that assumes independence, are used for analyzing such dataset
- Results churned from this dataset are completely independent from "**spatial arrangement**" of the entities.

Spatial context & data structure:

	Geographical Coordinate		ATTRIBUTE			
	X	Y	Variable 1	Variable 2	⋯	Variable n
Entity 1	X_1	Y_1	$attribute_{11}$	$attribute_{12}$	⋯	$attribute_{1n}$
Entity 2	X_2	Y_2	$attribute_{21}$	$attribute_{22}$	⋯	$attribute_{2n}$
⋮	⋮	⋮	⋮	⋮	⋱	⋮
Entity m	X_m	Y_m	$attribute_{m1}$	$attribute_{m2}$	⋯	$attribute_{mn}$

In this example, what defines the entity's geographic location are **X, Y GPS coordinates, or some geometry**

Definition of an entity's location are not limited to coordinates, you can have spatially reference areas with their associated boundaries with geometries.

This instance illustrates an example of geostatistical data.

- Attributes that defines an entity's location are typically explicitly incorporated into analysis
- Spatial statistical methods, that assumes dependence, are used for analyzing such geographically referenced dataset
- Results churned from this dataset are completely dependent from "**spatial arrangement**" of the entities.

Various Spatial Analytical Techniques

[1] Areal Data: Spatial Autocorrelation & Dependence

Spatial analytical technique used for detecting clusters of an outcome.

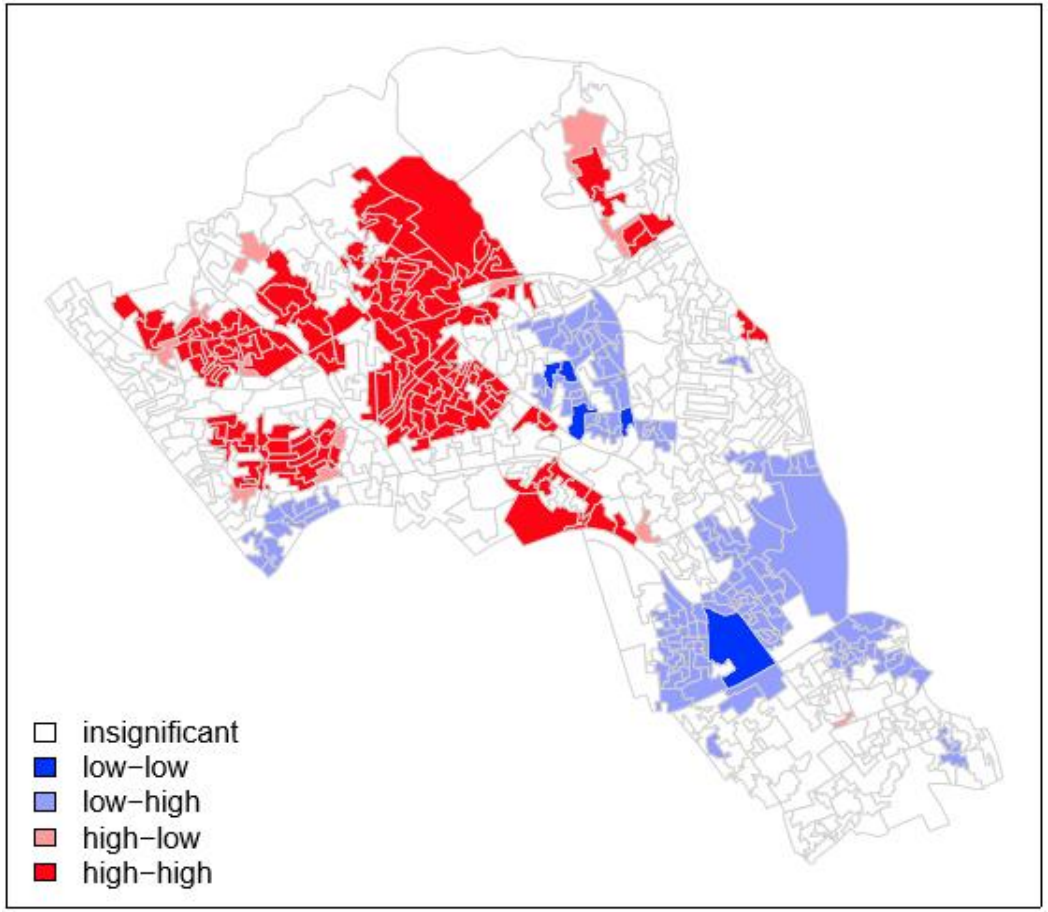
Global Test

- **Generate the hypothesis**
  - Null hypothesis: Are the patterns residential burglaries in Camden (London) are random?
  - Alternative hypothesis: The patterns of residential burglaries in Camden are not random, and are indeed either spatially clustered, or dispersed.

How to perform Global Moran's I test to get a p-value to accept, or reject the null hypothesis

- **Result (Global Moran's I = 0.5448 [p=0.0001 < 0.05])**

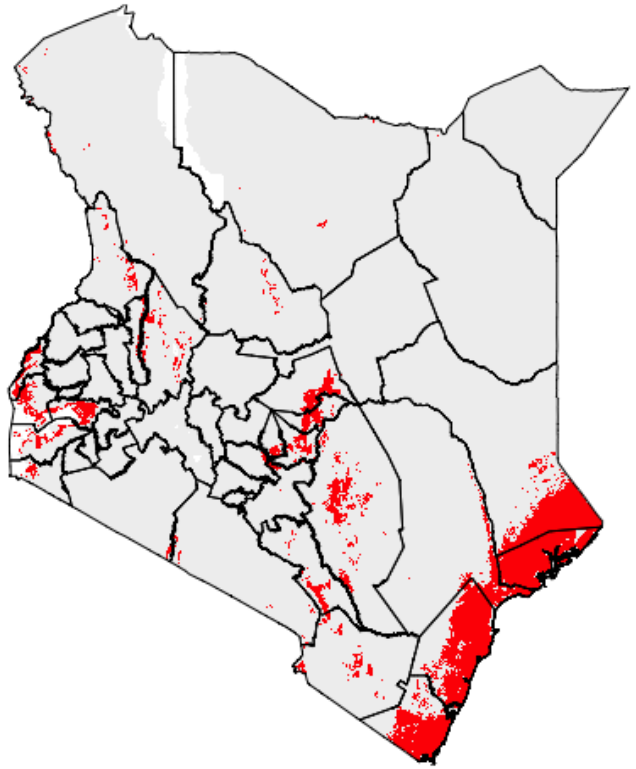
Local Test

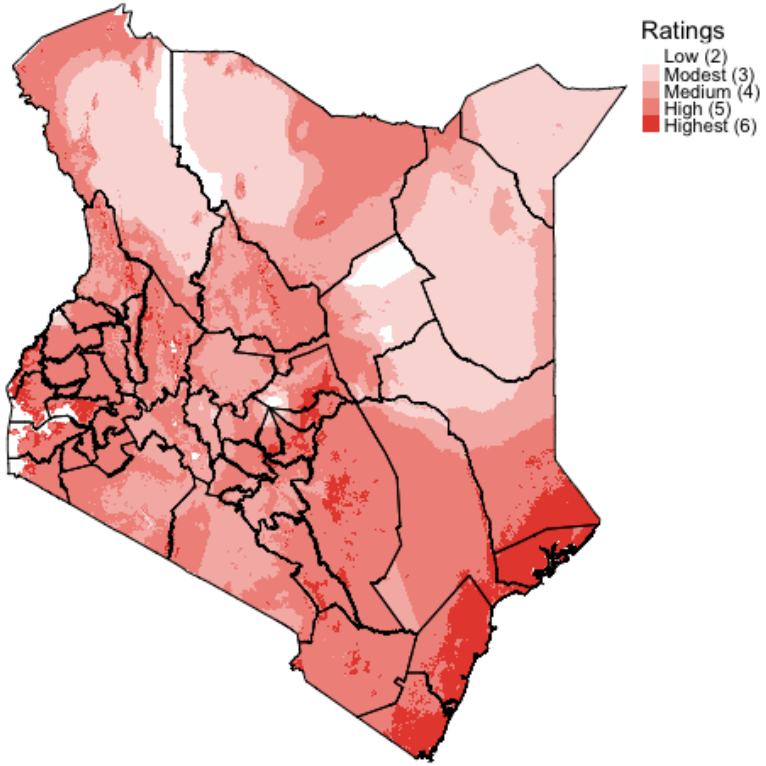


This is what we want. The p-values helps us to know whether the LISA estimates are statistically significant or not.

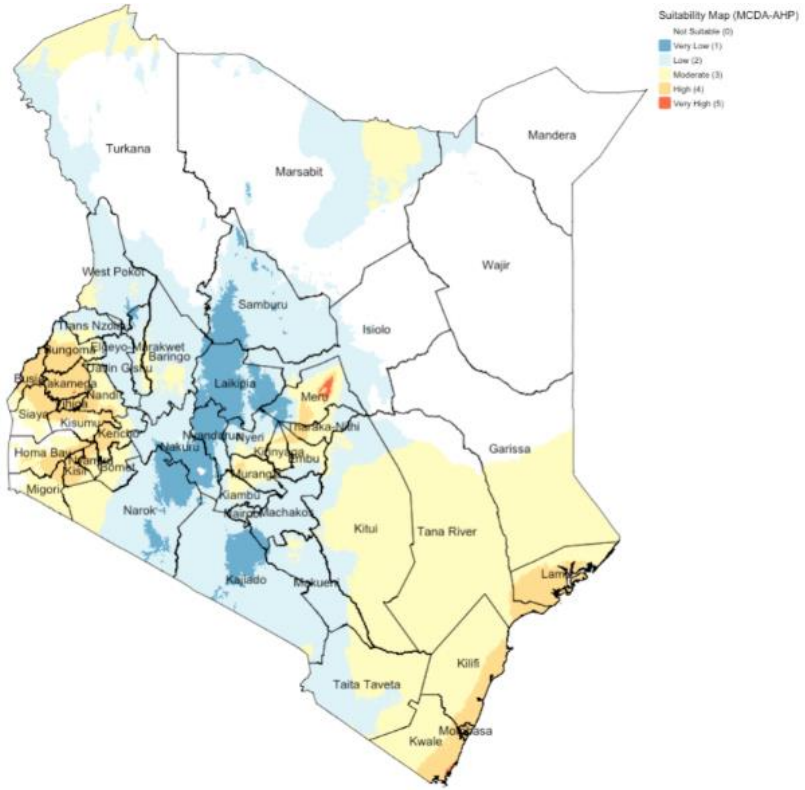
[2] Gridded Data: Multi-Decision Criteria Analysis

Simple Overlays: Booleans & Rankings





Analytical Hierarchy Process



CASE STUDY: GIS-MCDA applied to identify suitable areas for Lymphatic Filariasis (LF) transmission in Kenya.

NOTES

Description: There are no available up-to-date models to explain the occurrence of Lymphatic Filariasis (LF) in Kenya and **geospatial empirical data are scarce.** The Kenyan Ministry of Health (K-MoH), through its LF control programme, is planning to launch a public health intervention by introducing mass drug administration (MDA) of albendazole (combined with ivermectin) to infected people with LF and to remove microfilaria in their bloodstream. Mapping of suspected areas for LF must be carried out, however, due to financial constraints and limited resources, the K-MoH wishes to first **identify areas that are highly suitable for LF transmission** before spending this limited resources to survey, map and apply MDAs to these areas.

African man with heavy & chronic LF microfilaria infection, resulting in a swollen leg called 'Elephantiasis'

One many vectors, i.e., the Culex mosquito, that spreads LF by injecting microfilaria (microscopic worms) into their source of food (i.e., human) before taking its bloodmeal

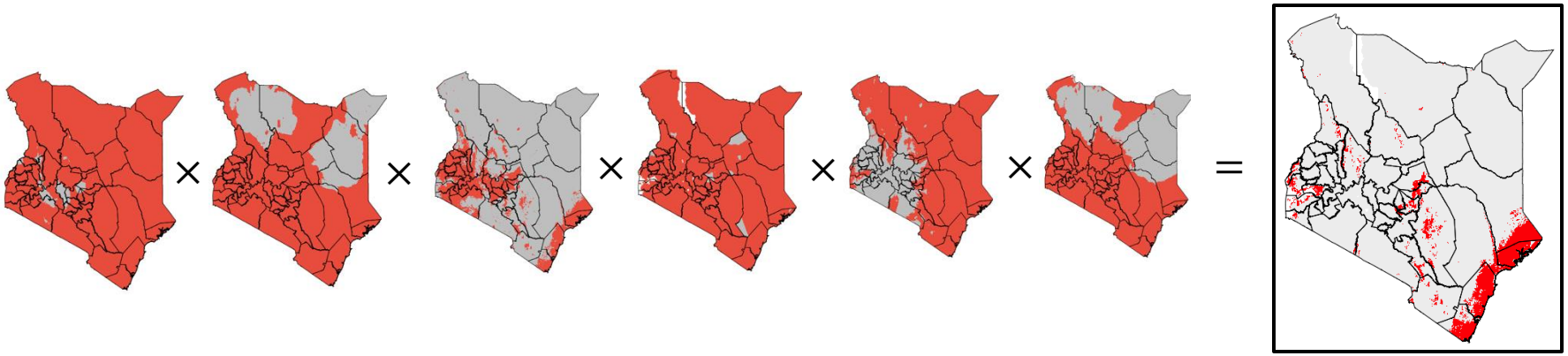
[a] Binary (or Boolean) maps with the factor layers

NOTES

We have these data as raster – for this approach requires simply reclassification of the data into binary raster and multiplication

- Temperature (> 15 degree Celsius) (**F**) (↑)
- Precipitation (> 350 mm) (**F**) (↑)
- Vegetation Index (> 0.5) (**F**) (↑)
- Population Density (> 0) (**F**) (↑)
- Elevation (<1,200m above sea-level) (**F/C**) (↓)
- Aridity (> 0.2 (Semi-humid & dry environment)) (**F/C**)(↑)

e.g., if temperature > 15 then change pixel value to 1 (good condition), else change to zero (bad condition)



Binary suitability is determined simply by multiplying the six individual layers that were reclassified to 0's and 1's.

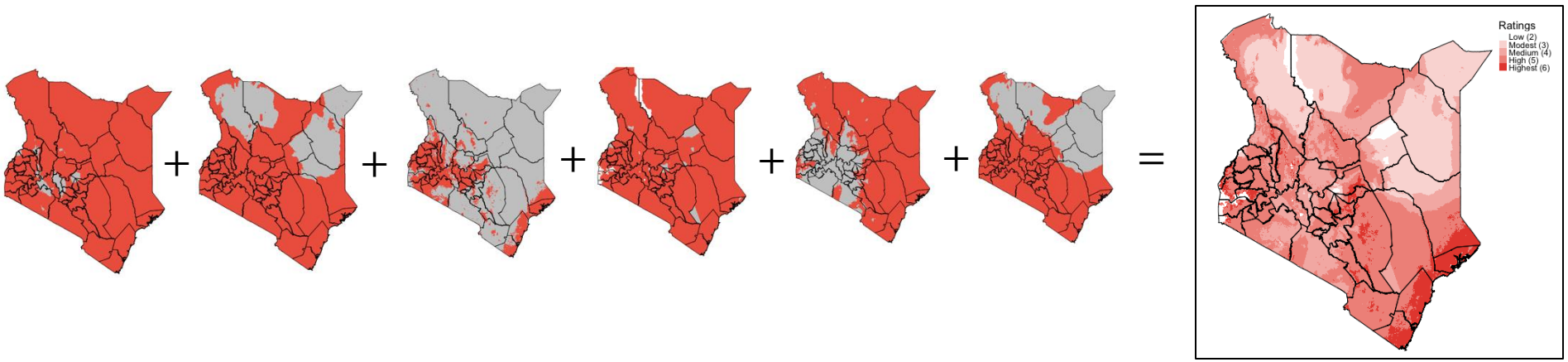
[b] Rankings/Ratings maps with the factor layers

NOTES

We have these data as raster – for this approach requires simply reclassification of the data into binary raster and summation to get scores

- Temperature (> 15 degree Celsius) (**F**) (↑)
- Precipitation (> 350 mm) (**F**) (↑)
- Vegetation Index (> 0.5) (**F**) (↑)
- Population Density (> 0) (**F**) (↑)
- Elevation (<1,200m above sea-level) (**F/C**) (↓)
- Aridity (> 0.2 (Semi-humid & dry environment)) (**F/C**)(↑)

e.g., if temperature > 15 then change pixel value to 1 (good condition), else change to zero (bad condition)



The ratings or suitability scores are determined simply by summing the values across the six individual layers that where reclassified (minimum value = 2, and maximum = 6).

[c] Analytical Hierarchy Process (AHP)

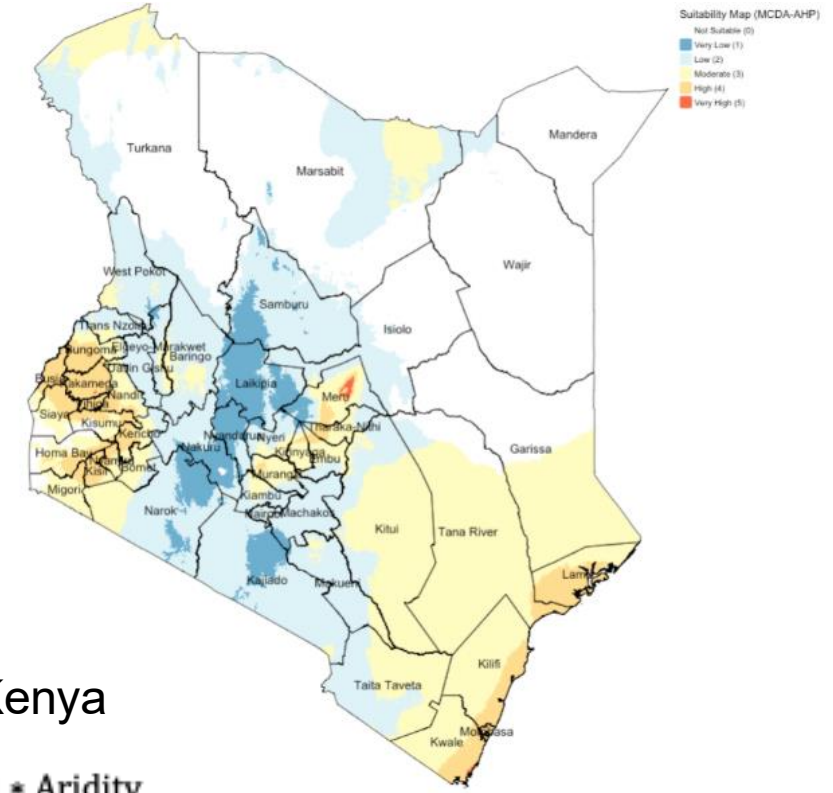
This involves building a decision table based on expert opinion

	Decision table using Saaty's scale									
	Extreme Favours	Very Strong Favours	Strongly Favours	Slightly Favours	Equal	Slightly Favours	Strongly Favours	Very Strong Favours	Extreme Favours	
	9	7	5	3	1	1/3	1/5	1/7	1/9	
Factor(s)										Factor(s) [versus]
Precipitation					1					Temperature
Precipitation				3						Population Density
Precipitation			5							Elevation
Precipitation	9									NDVI
Temperature				3						Population Density
Temperature			5							Elevation
Temperature		7								NDVI
Population Density				3						Elevation
Population Density			5							NDVI
Elevation				3						NDVI

Priority weight matrix is generated from the decision table

	Precipitation	Temperature	Population Density	Elevation	NDVI	Priority Vector (or Weights)
Precipitation	0.378151261	0.37366548	0.398230088	0.348837209	0.36	0.371776808
Temperature	0.378151261	0.37366548	0.398230088	0.348837209	0.28	0.355776808
Population Density	0.12605042	0.12455516	0.132743363	0.209302326	0.2	0.158530254
Elevation	0.075630252	0.074733096	0.044247788	0.069767442	0.12	0.076875716
NDVI	0.042016807	0.053380783	0.026548673	0.023255814	0.04	0.037040415
						1

Resulting Map from Decisions



This is used to build the equation to make the prediction about where LF is in Kenya

$$S = (0.372 * \text{precipitation} + 0.356 * \text{temperature} + 0.159 * \text{population} + 0.077 * \text{elevation} + 0.037 * \text{NDVI}) * \text{Aridity}$$

[3] Point Pattern Process: Ecological Niche Models

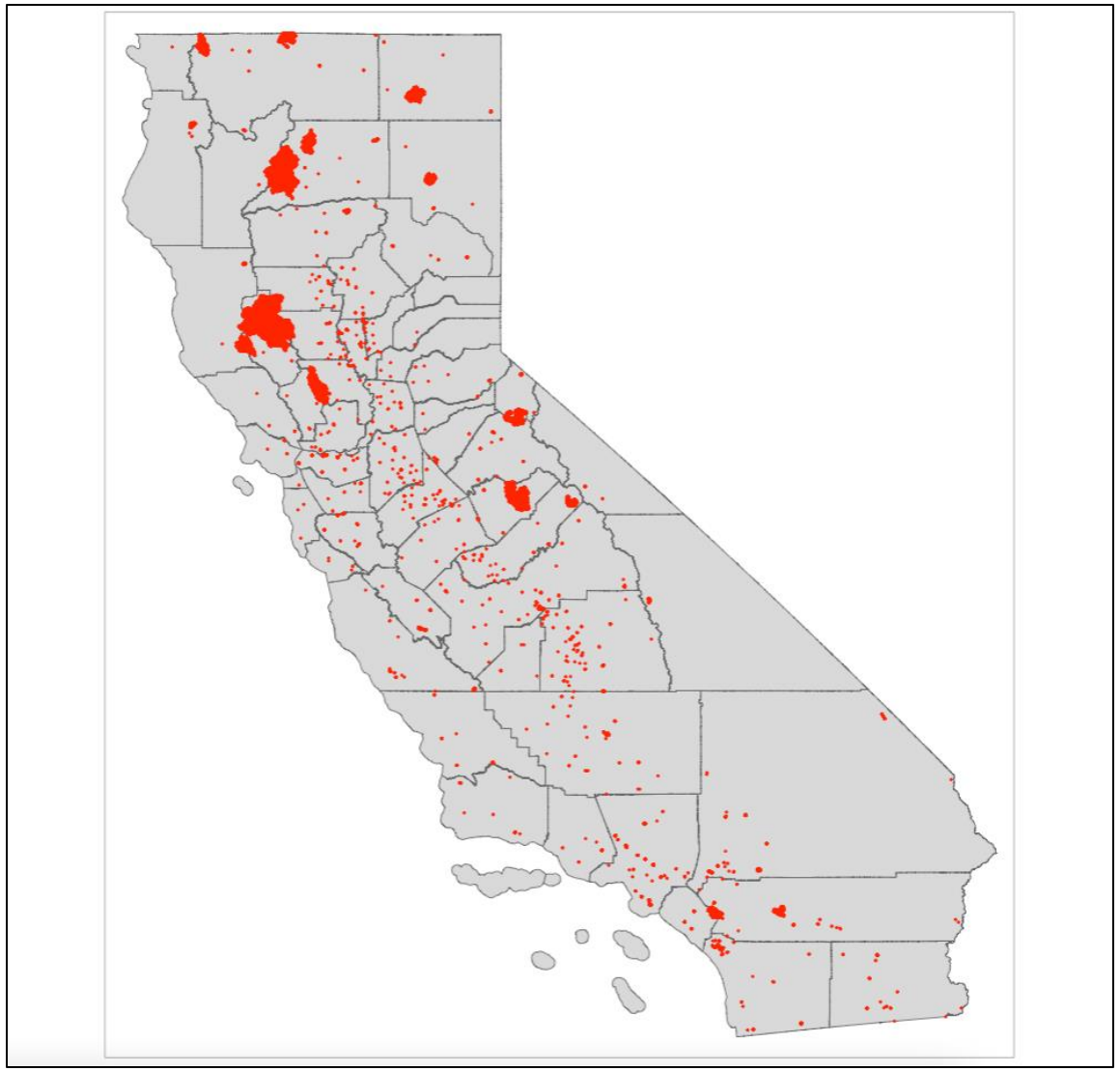


Figure shows the study area of California (counties) and points are occurrence of wildfires during the summer period of 2018. Presence-only points.

The social side of fires: assessing the inclusion of human social factors in fire prediction models

The objectives are to determine the occurrence of wildfires, as well as infer the extent (or zones) for such environmental hazard in California given a set of predictor variables (i.e., climate, vegetation, anthropogenic and socioeconomic risk factors which are raster)

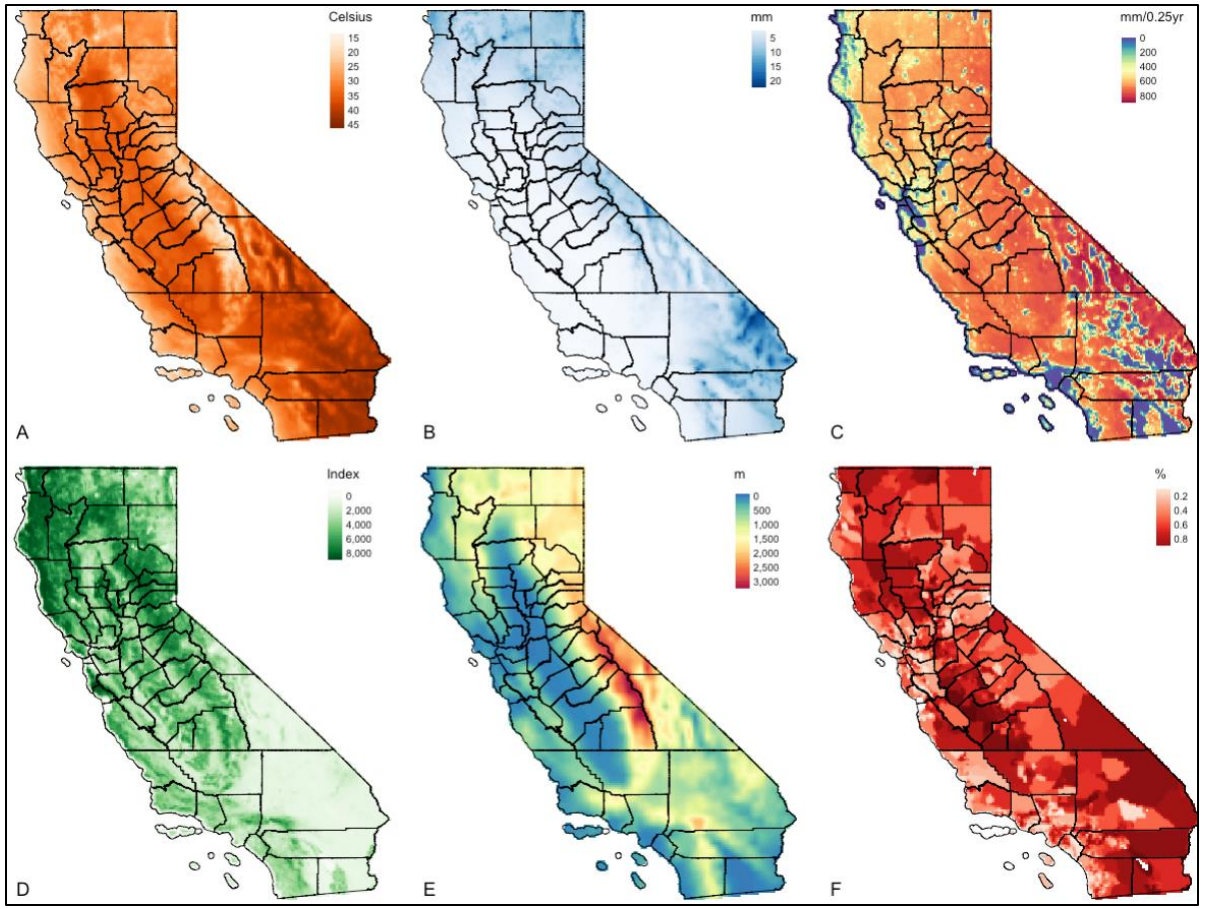
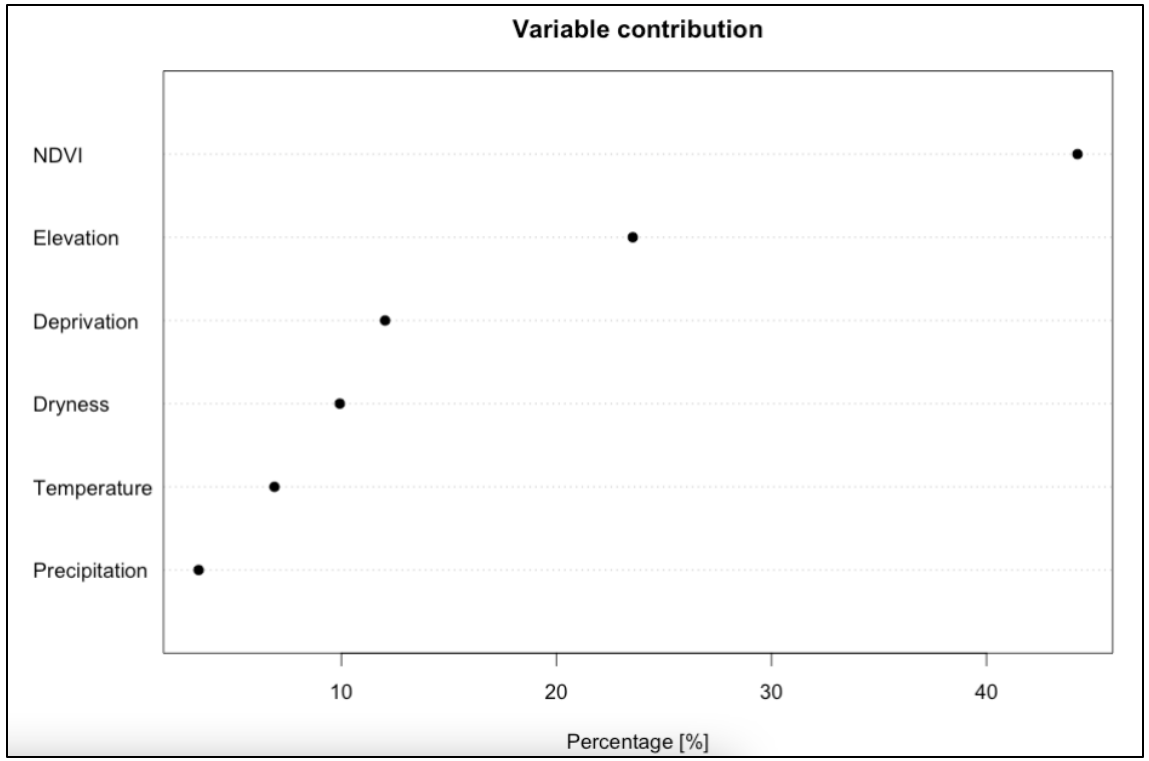


Figure panel shows A: Temperature (degree Celsius); B: Precipitation (mm); C: Dryness (Evapotranspiration) (mm/0.25 year); D: Vegetation (NDVI); E: Elevation (meters [m]); & F: Socioeconomic vulnerability index (%)

Using MAXENT

Result 1: Variable Contribution & Response curves



Interpretation: Here, we can see the following contribution estimates: NDVI (44.2321%); Elevation (23.5530%); Deprivation (12.0339%); Dryness (9.9266); Temperature (6.8892%); and Precipitation (3.3653%). The contribution estimates should sum up to 100%. From this plot, we can see that the model is most sensitive to variation in NDVI, followed with additional contributions from land surface elevation, and from increased levels of socioeconomic deprivation (reporting top three).

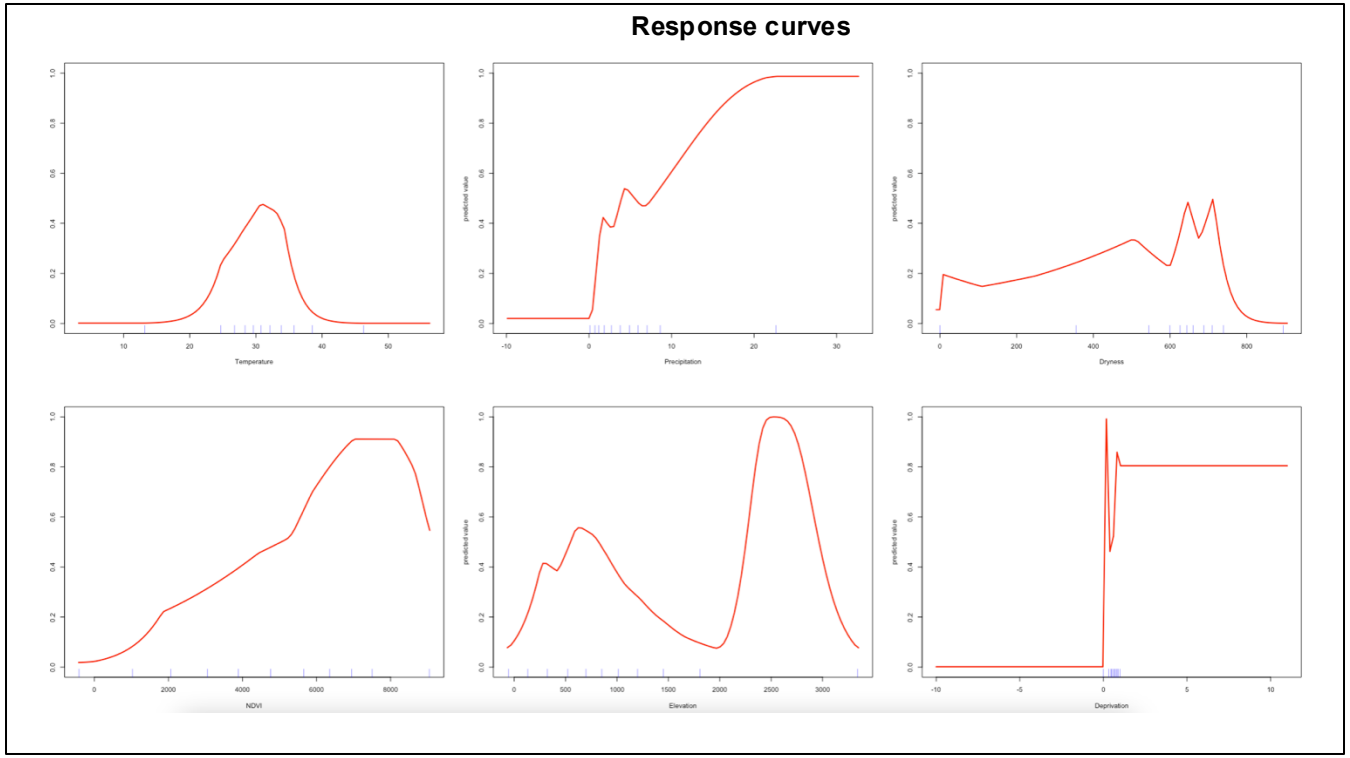
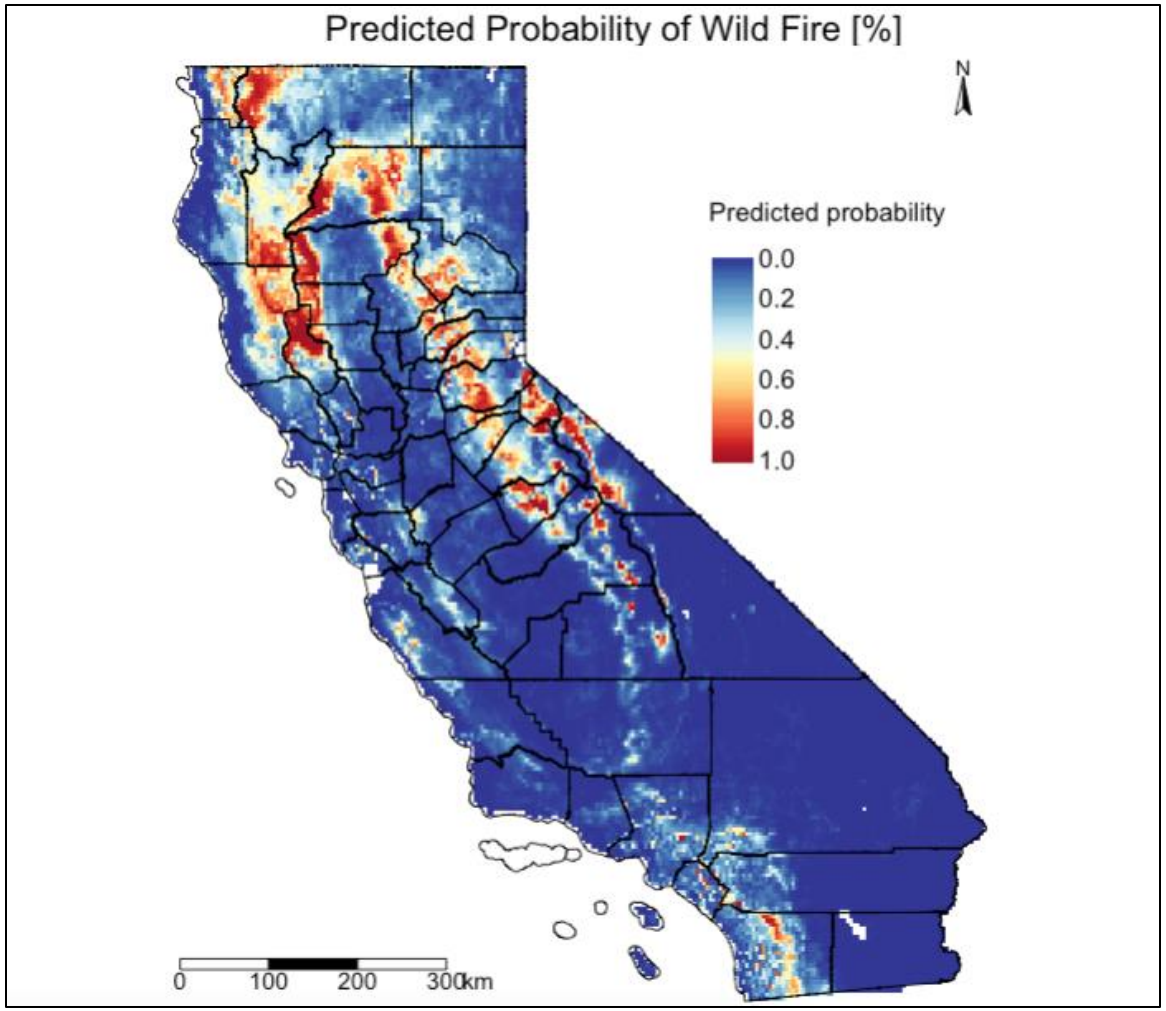


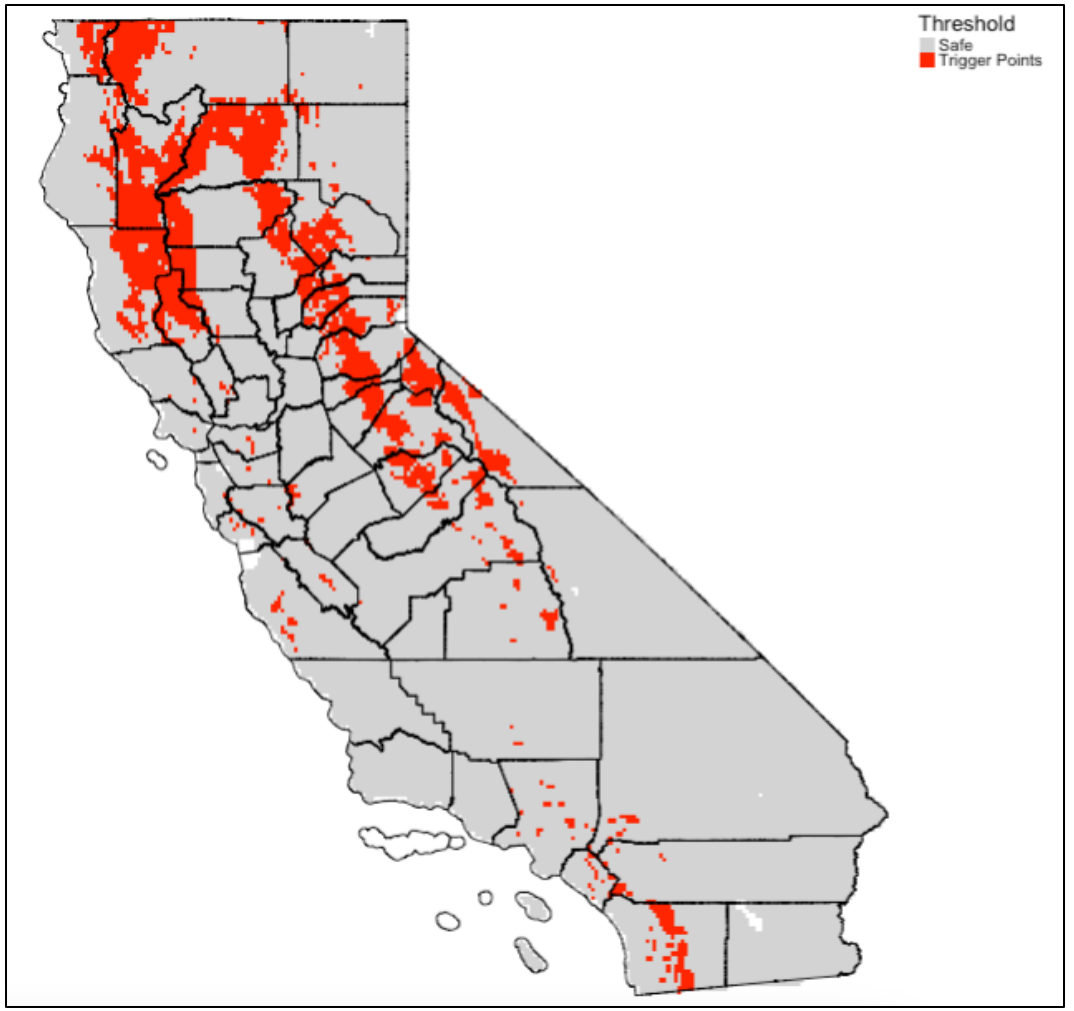
Figure panel shows from top left: Temperature (degree Celsius); Precipitation (mm); Dryness (Evapotranspiration) (mm/0.25 year); Vegetation (NDVI); Elevation (meters [m]); & Socioeconomic vulnerability index (%)

Interpretation: In the response plots, we are looking at how the probability of fire occurrence (Y-axes, from zero to one) varies with each the environmental predictors (X-axes). From these plots, we can see that the MAXENT models can include complex environmental responses including plateau, linear, and nonlinear shapes, and some which are utterly unclear. For example, if we look at mean temperature during the summer, we can see that the probability for fire occurrence peaks around 0.60 when temperatures are around 30 degrees Celsius. We can also see that the probability of such outcome increases with more and more vegetation during the summer period. Probability in terms of fires in relation to deprivation is a flat line. For precipitation, dryness and elevation - the patterns are unclear.

Result 2: Prediction of high-risk locations & where it is suitable (output is in gridded format)



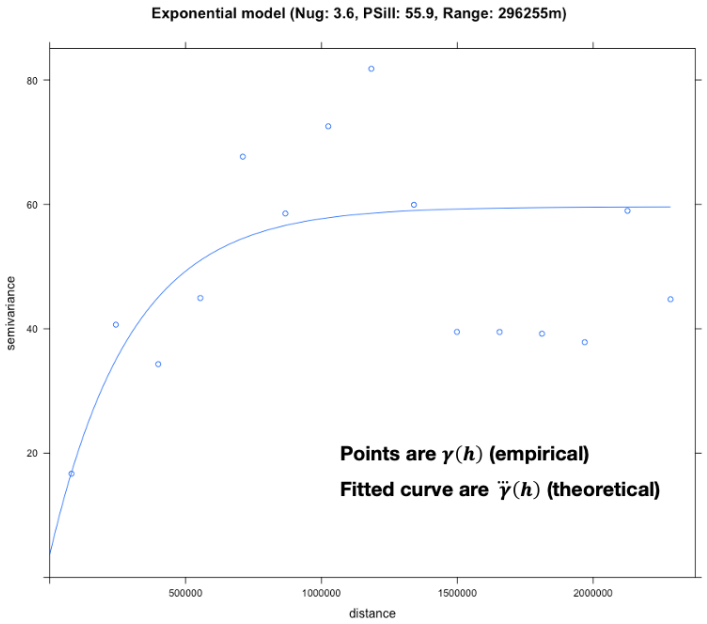
We mapped predicted probability of fires using the trained model after making sure its valid. The multi-band raster is fed to the trained model to make full scale predictions.



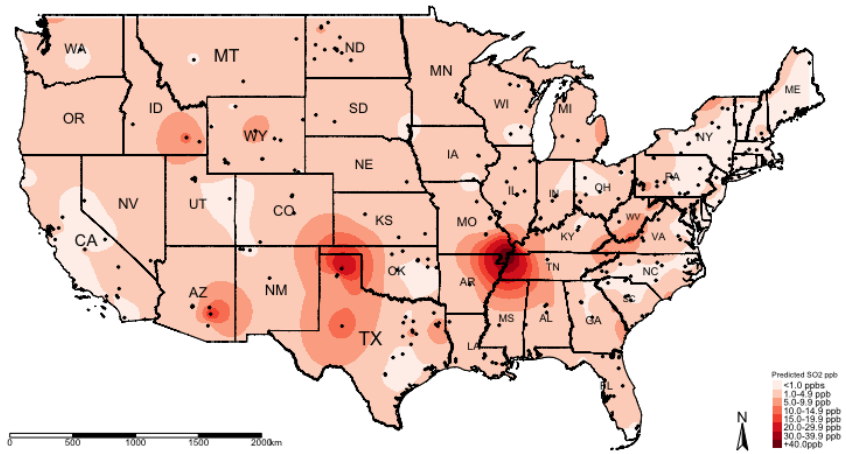
This is based on the optimized estimate obtained after model validation i.e., maximizes the True Positive Rate and the True Negative Rate is 0.4054474 (40.55%). Here, we mapped **predicted probability > 0.4054** as a reclassified raster.

[4] Geostatistical Data: Kriging

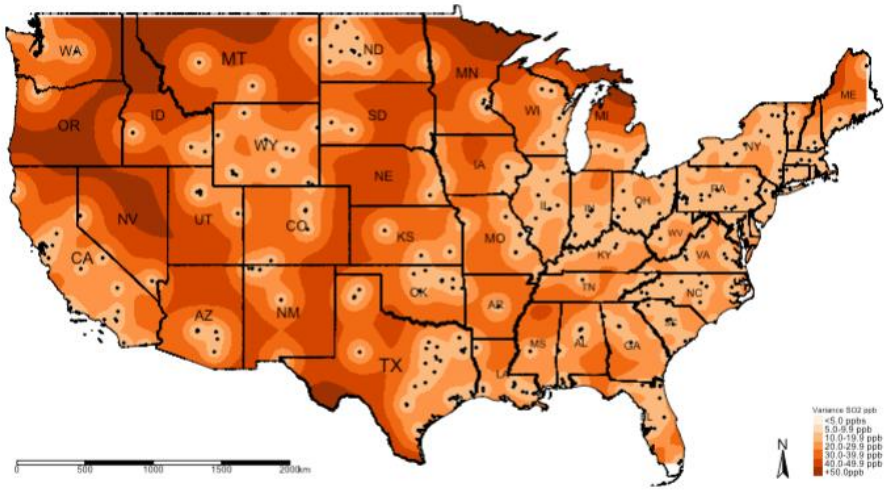
Variogram analysis



Predicted air SO_2 level from the Kriging model



Uncertainty for the predicted air SO_2 levels from the Kriging model

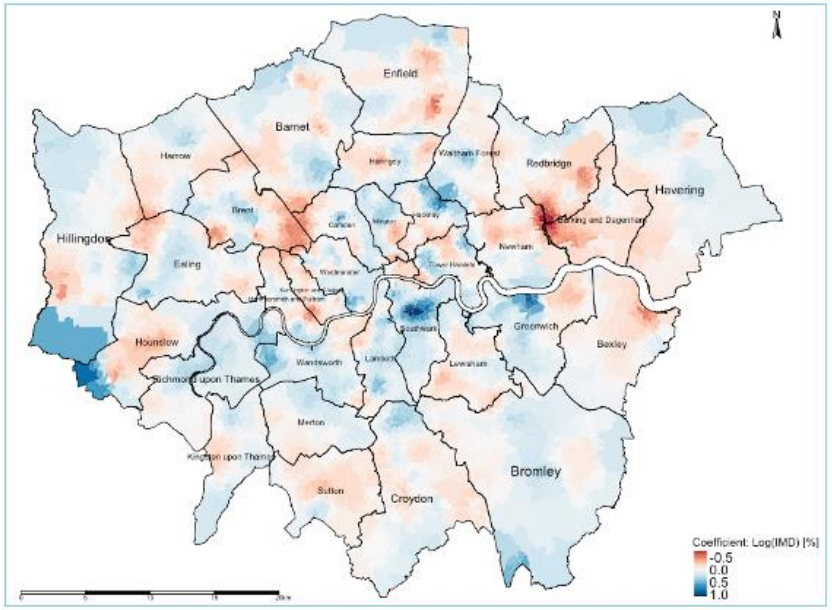


Interpretation: The **nugget** is a small value of 3.6, which is an indication for evidence of spatial variability in the concentrations for SO_2 across sampling sites in USA. The **range** is 296,255m, which indicates that any separation distance above this value means that spatial autocorrelation in the observed levels of SO_2 between points are no longer similar. However, points with a separation distance less than 296,255m indicated that their SO_2 values are similar. For the **partial sill**, within this range for the Semivariance i.e., 3.6 and 55.9 – is the values are spatially autocorrelated.

Along the belt of the following states – Missouri, Tennessee, Kentucky and Illinois, the predicted concentration of SO_2 levels exceeds +40ppb, whereas there are pockets in Texas where concentrations of SO_2 are a cause for concern i.e., 30-39.9ppb.

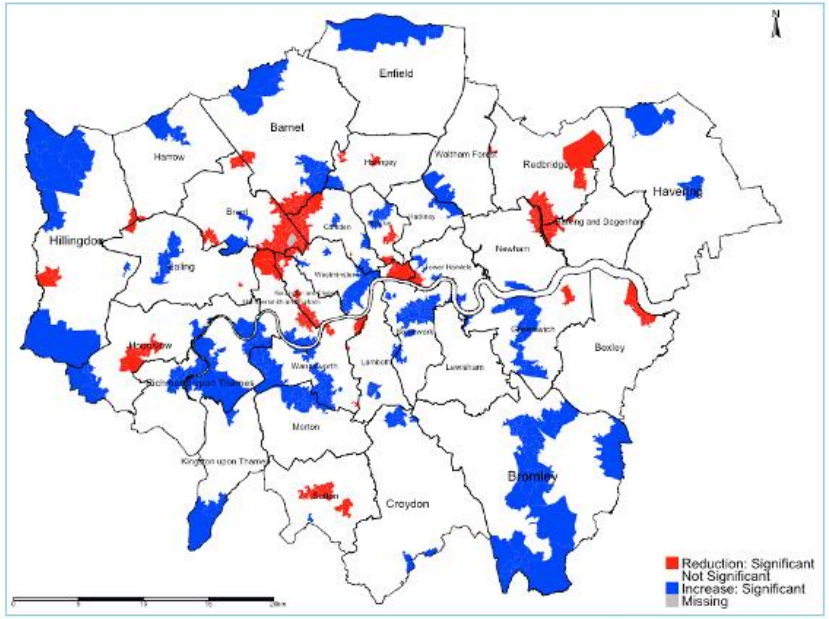
[5] Areal Data: Geographically Weighted Regression

[1] Local regression coefficients



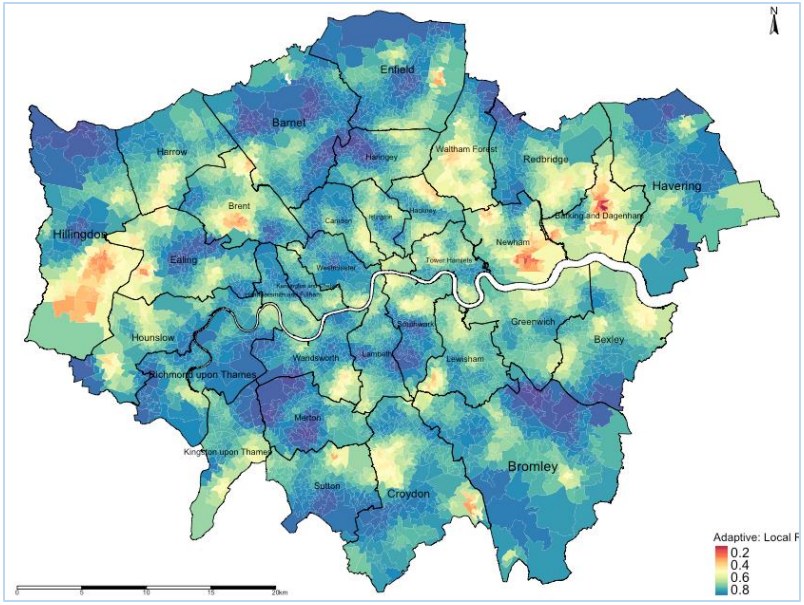
Interpretation: There is spatial variability in the relationship between our variable socioeconomic deprivation (transformed) and averaged house price (transformed) in London. The GWR outputs reveals that local coefficients range from a minimum value of -0.946 to a maximum value of 1.085, indicating that one percentage point increase in the levels of deprivation in LSOAs of London is associated with a reduction of 0.946% in house prices in some LSOAs and (weirdly) an increase of 1.085% in others. Broadly, the relationship are opposing.

[2] Statistical Significance



Interpretation: For instance, in the **Borough of Hounslow**, we can see a significant reduction in house prices in relation to increased levels of socioeconomic deprivation (adjusted for income and accessibility). Such reduction are clustered in the mid-section of Borough of Hounslow which were coloured red. Note that in far northeastern section of the Borough of Hounslow with pockets of LSOA's coloured blue shows a significant increase in house price in relationship to IMD which is difficult to explain and thus can be interpreted as a chance finding. All sections that are coloured white are not significant.

[3] Local R^2 : Model Diagnostics



Interpretation: The areas that are going towards the shade of dark reds (i.e., value of 0) are local regression models that have broadly performed poorly in its prediction for house price and its association with the three variables (income, deprivation and PTAL). Likewise, the areas that are going towards the shade of dark blues (i.e., value of 1) are local regression models that have broadly performed very well in its prediction for house price and its association with the three variables (income, deprivation and PTAL).

Note: These results are essential as the local R2 values of each area show the model's ability to predict the explained variance in house prices caused by deprivation, income and accessibility for specific areas.

[6] Areal Data: Spatial Bayesian Models for Risk Assessment

Report the overall risk results

2017	LIRAa 2	
	RR (95% CrI)	Pr(RR>1)
Intercept	1.64 (95% CrI: 0.14 to 7.07)	0.51
Temperature	0.93 (95% CrI: 0.74 to 1.12)	0.23
Precipitation	1.01 (95% CrI: 0.96 to 1.07)	0.73
NDVI	1.09 (95% CrI: 0.71 to 1.60)	0.63
Urbanisation	1.18 (95% CrI: 0.37 to 2.90)	0.52

RR: Relative risks; Pr(RR > 1): Exceedance probabilities (the probability that RR being greater than 1)

Interpretation (examples):

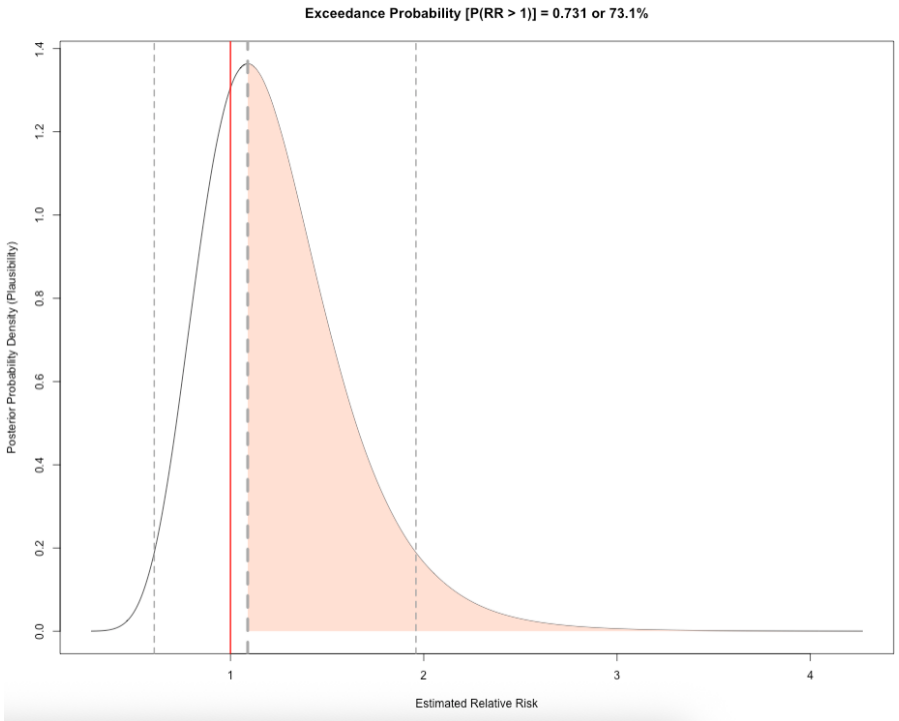
- **Intercept**: The overall baseline risk of mosquito-borne infestation is 1.64 times (or 64%) **greater** in the population of Campina Grande. The overall probability that there's excess risk of infestation (i.e., RR > 1.00) is 51%.
- **Temperature**: In relation to temperature, the risk of mosquito-borne infestation is 0.93 times (or 7%) **lower** in Campina Grande. The probability of observing an excess risk of infestation (i.e., RR > 1.00) in relation to temperature is 23%.
- **Urbanisation index**: In relation to urbanisation, the risk of mosquito-borne infestation is 1.18 times (or 18%) **higher** in Campina Grande. The probability of observing an excess risk of infestation (i.e., RR > 1.00) in relation to urbanisation is 52%.

NOTE: All relative risk estimates have the null value (1) between its lower and upper 95% credibility intervals. While the results, excluding temperature, show an increased risk of infestation – **these are all statistically not significant**.

Describe the distribution of the relative risks

Here, in this distribution image, we show all the possible values of how temperature is related to mosquito infestation in terms of impact and risk given the dataset we have. The relative risk for mosquito infestation given temperature can be anywhere from 0.278 (reduced risk) to 4.271 (increased risk). But from the graph, it indicates that the most plausible relative risk estimate is 1.089 because it has the highest density in our posterior distribution.

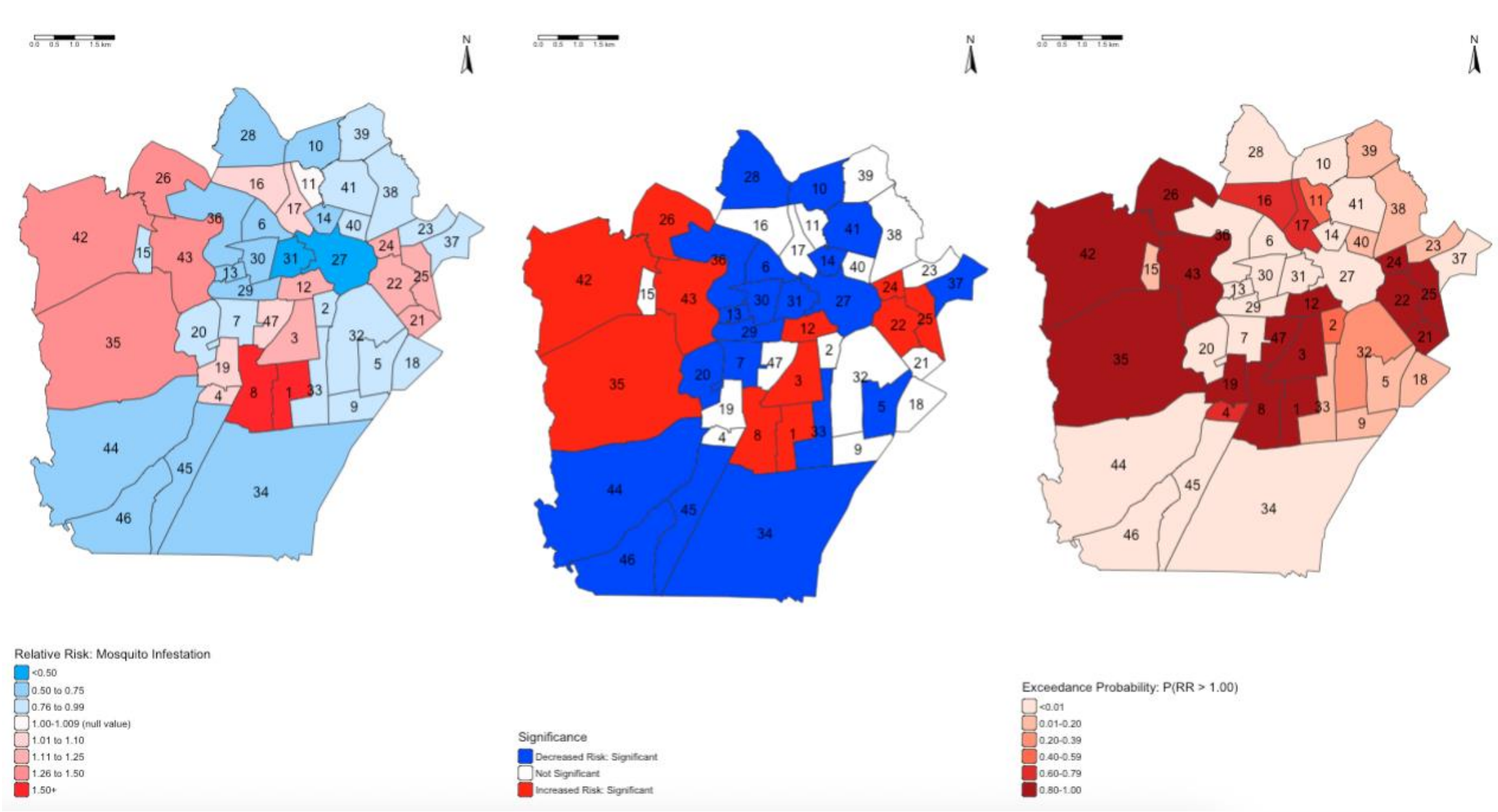
To gauge levels of certainly about this estimated risk, we can calculate the probability that the relative risk is 1.089 or above. The probability of observing a relative risk of mosquito infestation being 1.089 or above, given the environmental levels of temperature, is 0.731 or 73.1%, which is quite high



[1] Area-specific risks

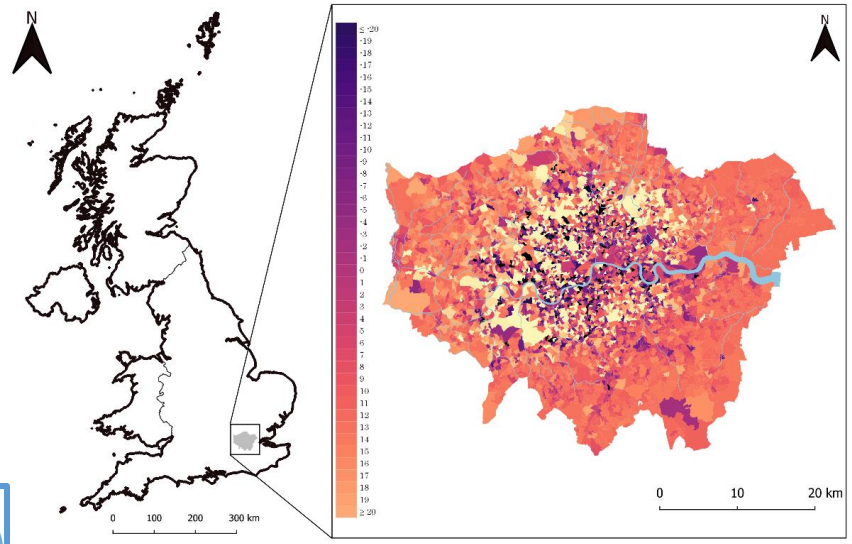
[2] Statistical Significance

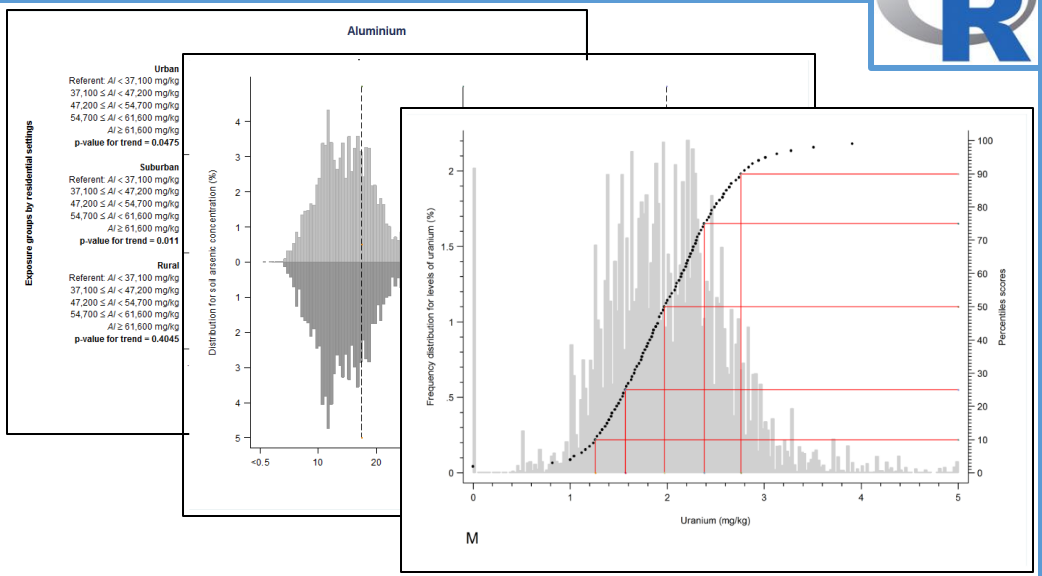
[3] Exceedance Probabilities

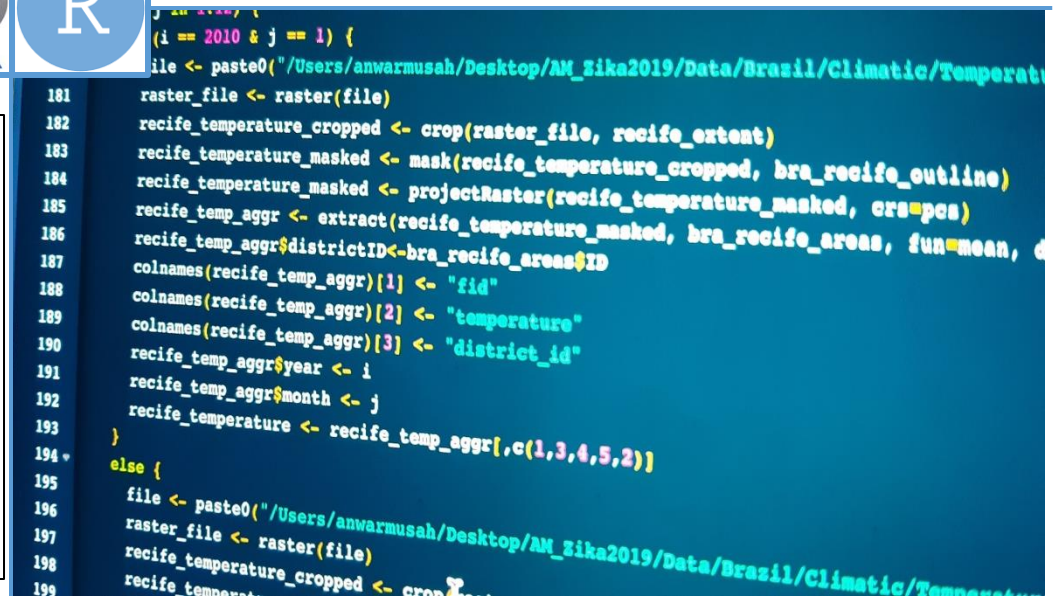


RStudio as a GIS Software

What is R/RStudio?







R (Standard)

RStudio (Best)

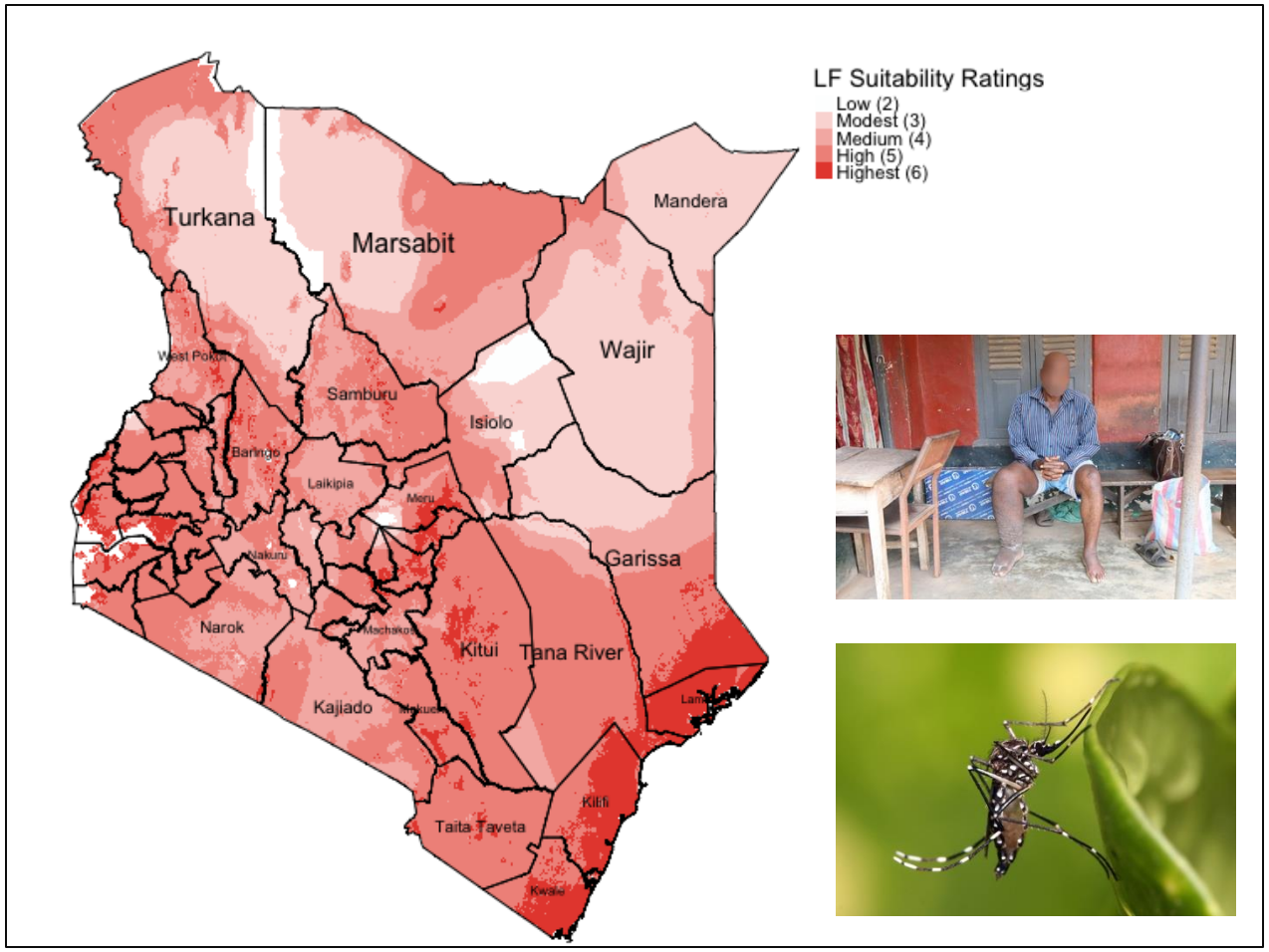
There are two version of the software: 1.) R, and 2.) RStudio; The second is much preferred as its straightforward and intuitive.

Why are we teaching RStudio?

1. Flexible and provides access to powerful packages for analysis

2. Impressive graphs, visualizations and maps

3. Excellent statistical capabilities too



Example: Map generated in R to illustrate areas that are environmentally suitable for the spread of neglected tropical disease called 'Lymphatic Filariasis (LF)' in Kenya.

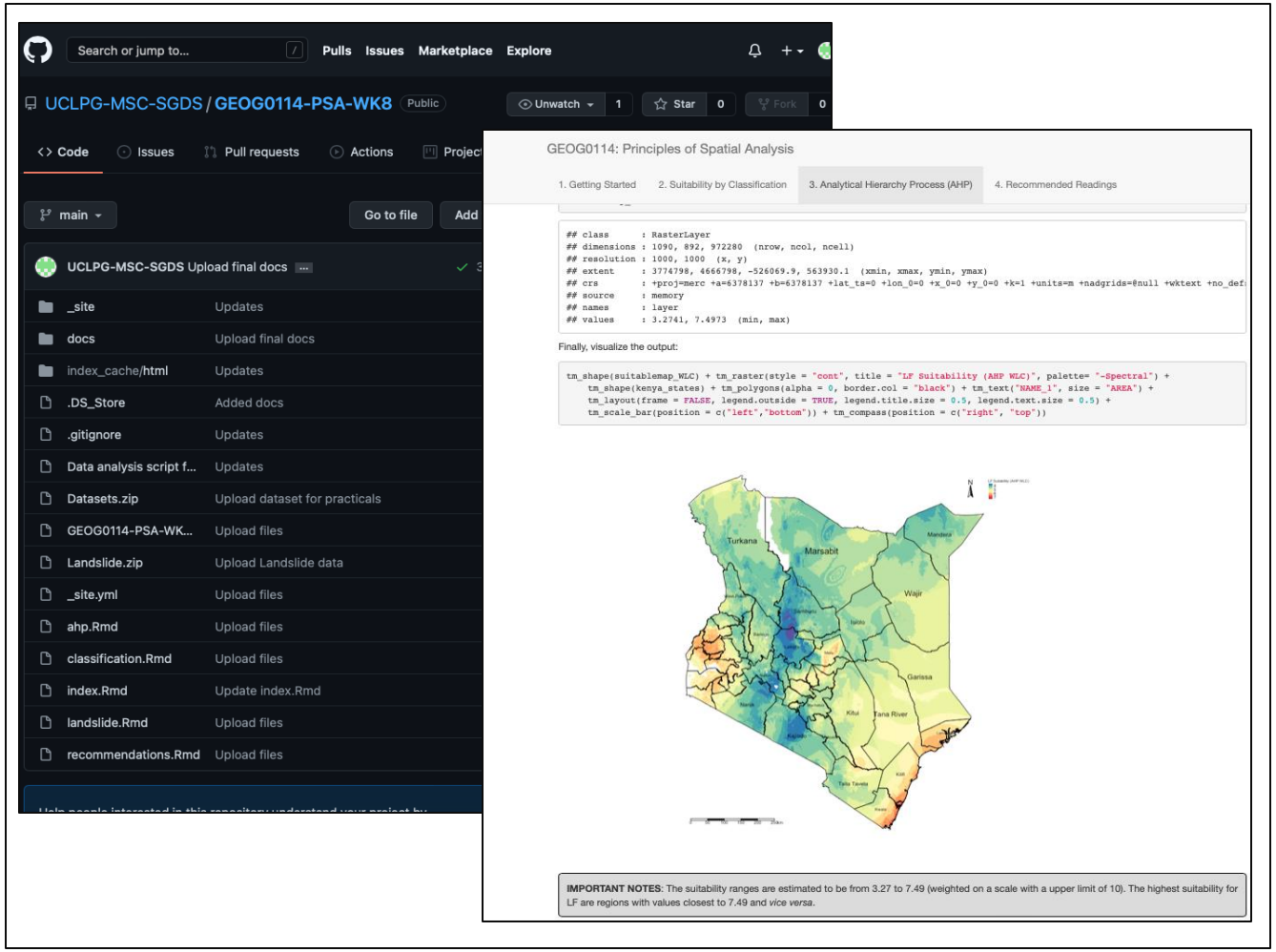
Sources:
1. Global Atlas for Helminths Infection (http://www.thiswormyworld.org)
2. ESPEN (https://espen.afro.who.int)

… and why learn how to code in RStudio?

1. **Efficiency**
   - Automated tasks and data managing
   - Can recycle & reuse code scripts for new projects

2. **Fosters good scientific practice**
   - Transparency and replication (AKA reproducible research)
   - Creates log so anyone can follow in your footstep (i.e., github, gitlab etc.,)

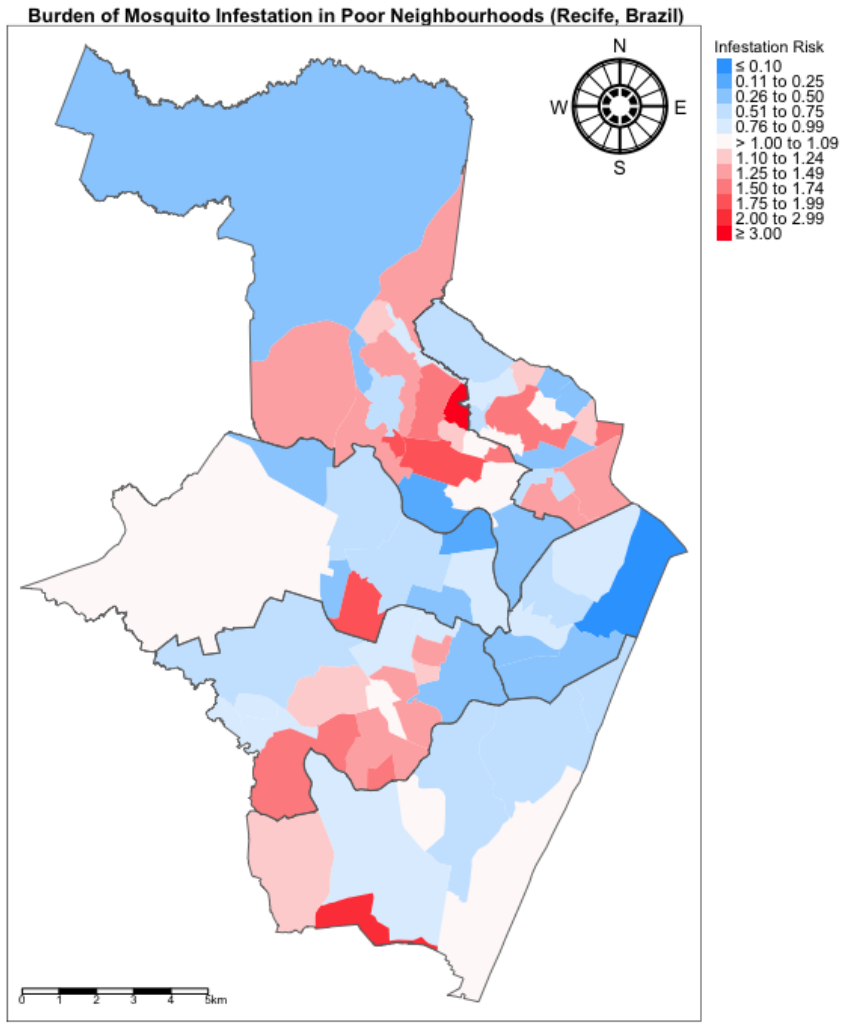
You can literally pull-off some really creative stuff like generating websites, accessing tools via APIs etc.



Example: Working in RStudio and synchronising it with GitHub to not only use as a cloud back-up, but to generate a website through RStudio and GitHub for teaching MSc Students.

Sources:
1. GitHub (https://github.com)

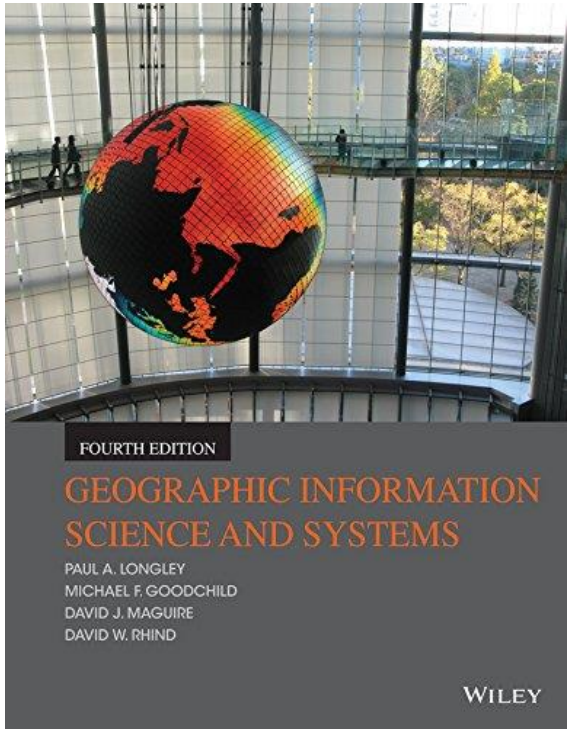
Example of a basic code structure in RStudio

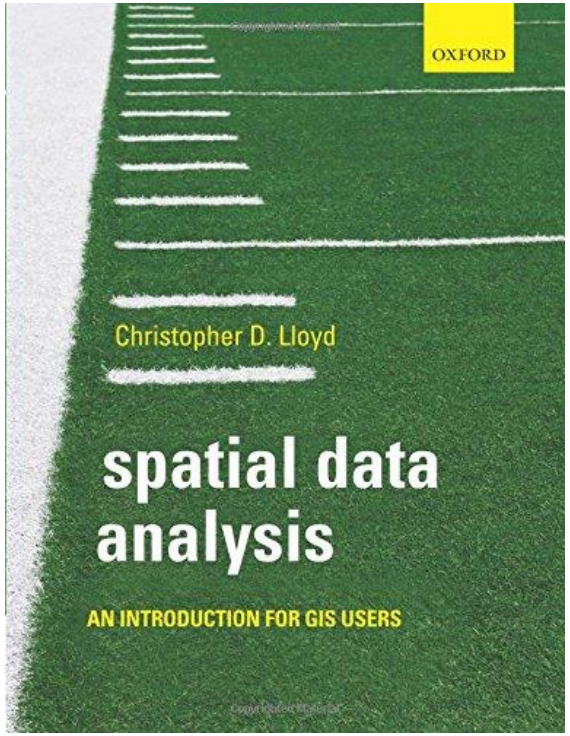



```
# comment: activate packages for performing GIS in R
library("sf")
library("tmap")

# comment: add neighbourhood shapefile w/mosquito infestation data using read_sf()
recife.neighbourhoods <- read_sf("Recife_neighb_epsg3857_fixed.shp")
recife.healthzone <- read_sf("Recife_regions_epsg3857_fixed.shp")

# comment: assigning labels for the risk estimate legends
RiskCategorylist <- c("\u2264 0.10", "0.11 to 0.25", "0.26 to 0.50", "0.51 to 0.75",
"0.76 to 0.99",">1.00 to 1.09", "1.10 to 1.24", "1.25 to 1.49", "1.50 to 1.74", "1.75
to 1.99", "2.00 to 2.99", "\u2265 3.00")

# comment: generating the divergent color scheme from Blues to Red spectrum
RRPalette <- c("#33a6fe", "#65bafe", "#98cffe", "#cbe6fe", "#dfeffe", "#fef9f9",
"#fed5d5", "#feb1b1", "#fe8e8e", "#fe6a6a", "#fe4646", "#fe2424", "#fe0000")

# comment: map of risk of infestation
tm_shape(recife.neighbourhoods) +
  tm_fill("RelativeRiskCat",
          style = "cat",
          title = "Infestation Risk",
          palette = RRPalette,
          labels = RiskCategorylist) +
tm_shape(recife.healthzone) +
  tm_polygons(alpha = 0, border.alpha = 0.90) +
  tm_layout(frame = TRUE,
            main.title = "Mosquito Infestation in Neighbourhoods (Brazil)",
            main.title.size = 0.8,
            main.title.position = 0.02,
            main.title.fontface = 2,
            legend.outside = TRUE,
            legend.outside.position = "right",
            legend.title.size = 0.8,
            legend.text.size = 0.7) +
  tm_scale_bar(position = c("left", "bottom")) +
  tm_compass(type = "radar", show.labels = 2, position = c("right", "top"))
```

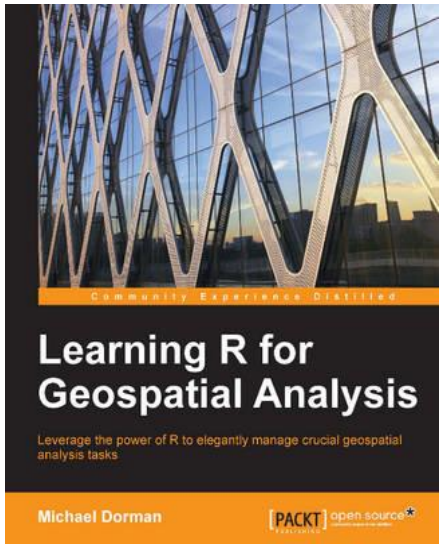
Access to tomorrow's practical session: https://uclpg-msc-sgds.github.io/BISemWeb2025/00-index.html

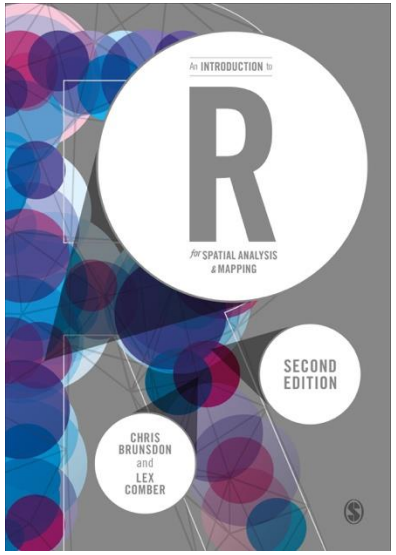
Recommended Books for learning Spatial Data Science

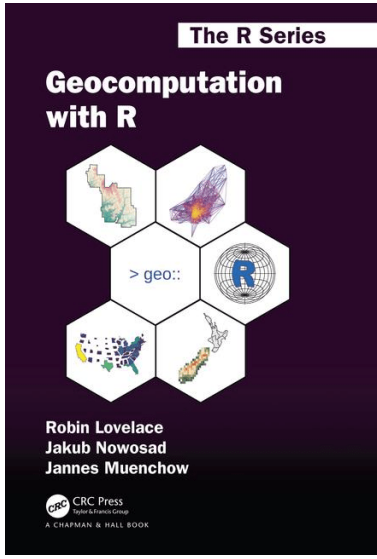




High recommendation for the mastery of basic theory and principles of spatial analysis







High recommendation for the coding experience and execution of spatial analysis in R

Any questions?