Day 2: Spatial Autocorrelation

Short Lecture & Computer Practical

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

Definition [1]

Recall Tobler's 1st Law of Geography

Spatial dependence is the propensity for how nearby objects in geographic space tend to influence each other. It's a reflection of how values observed at one location (e.g., city, region, country) is dependent on the values of neighbouring observations from nearby locations.

- **Spatial autocorrelation**

This describes the degree of how spatial locations (i.e., points, areas, or raster cells) **close** to each other share similar values (i.e., locations that are akin to each other).

Definition [2]

Spatial autocorrelation: This describes the degree of how spatial locations (i.e., points, areas, or raster cells) **close** to each other share similar values (i.e., locations that are akin to each other).

- We can test our data for spatial autocorrelation, with some form of statistical measure of the similarity of attributes of our data.

- We want to distinguish between areas of positively autocorrelated patterns (in which areas with high values are surrounded by areas with high values, and areas with low values that are surrounded by areas with low values to, i.e., **clusters**);

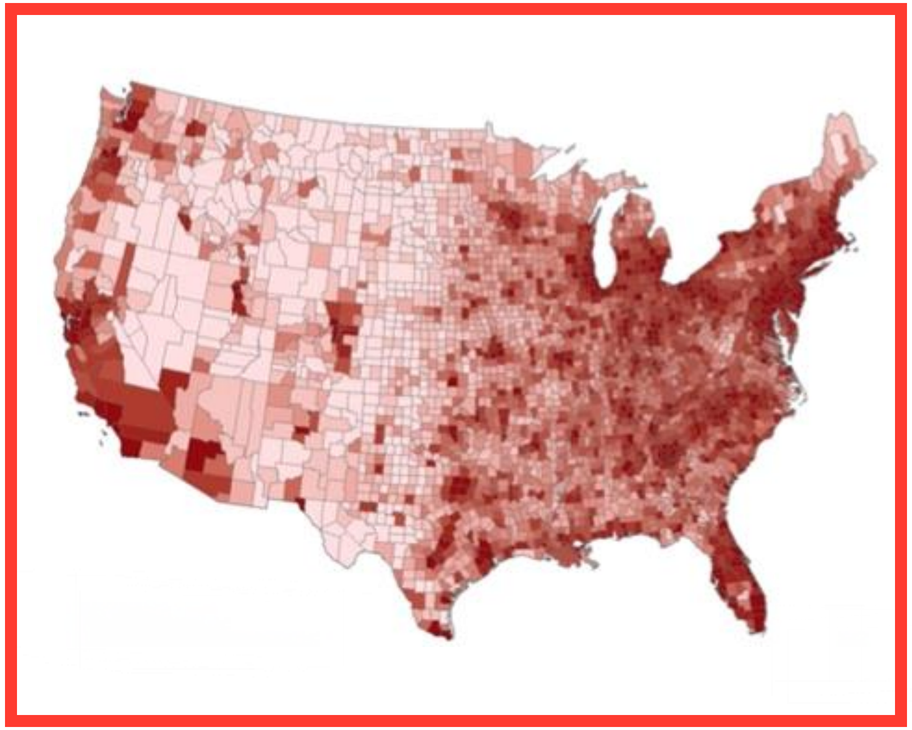
Definition [3]

Spatial autocorrelation: This describes the degree of how spatial locations (i.e., points, areas, or raster cells) **close** to each other share similar values (i.e., locations that are akin to each other).

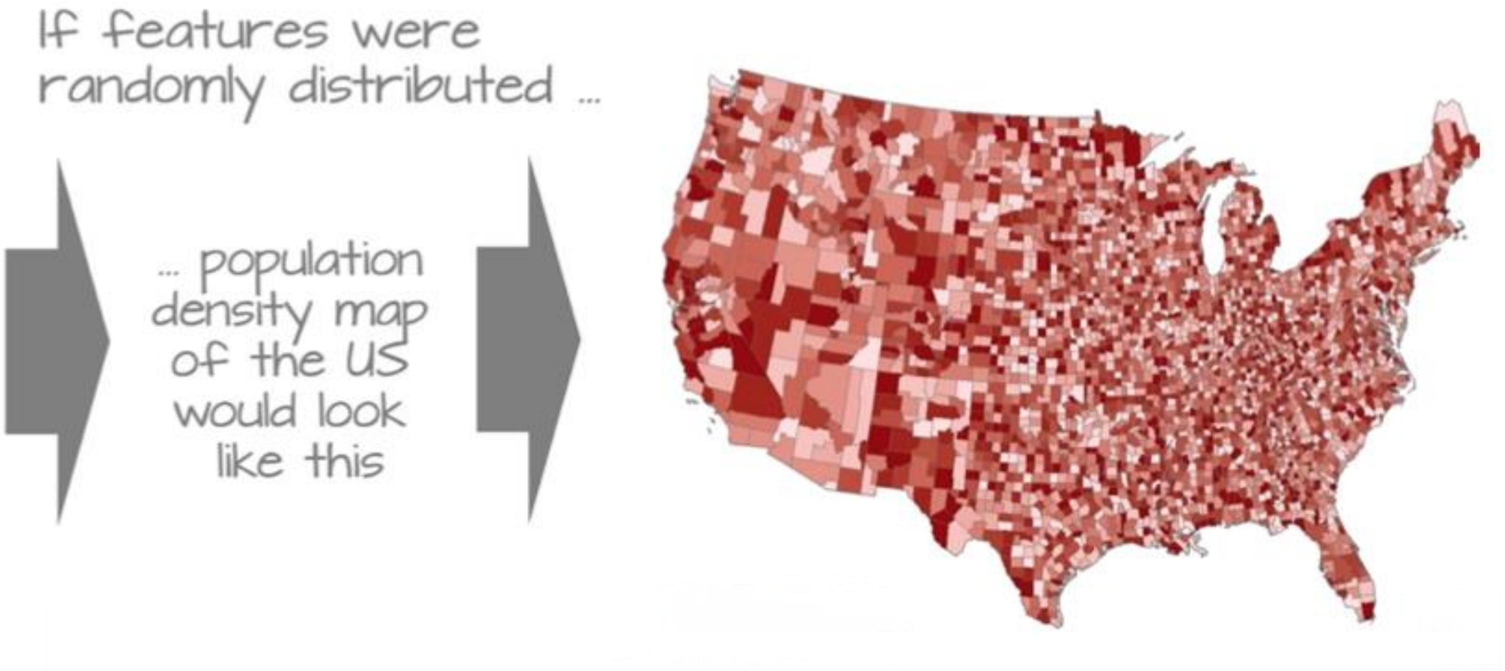
The hypothesis statement for testing evidence for spatial autocorrelation

- **Null hypothesis:** The outcome of interest are spatially independent (i.e., patterns are random)
- **Alternative hypothesis:** The outcome of interest are not spatially independent (i.e., hence, there is evidence of **clustering** or dispersion)

We can apply these hypotheses tests on these scenarios

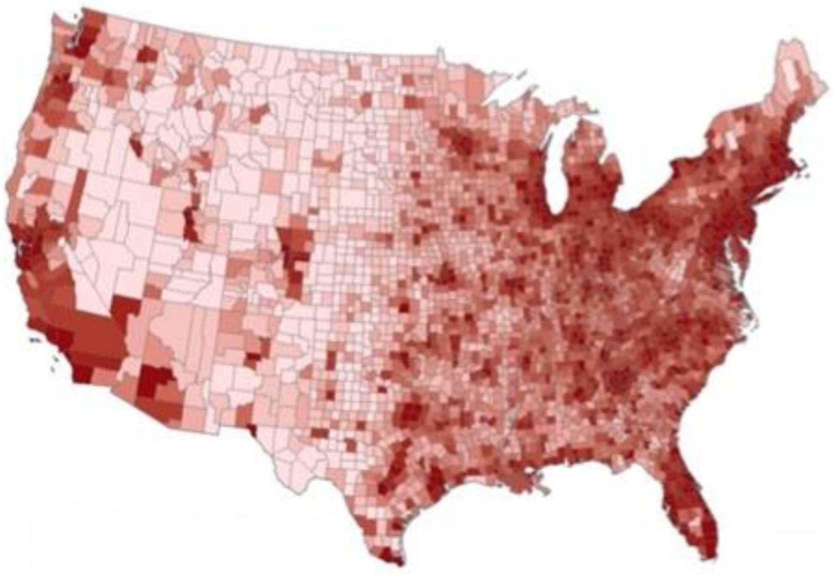


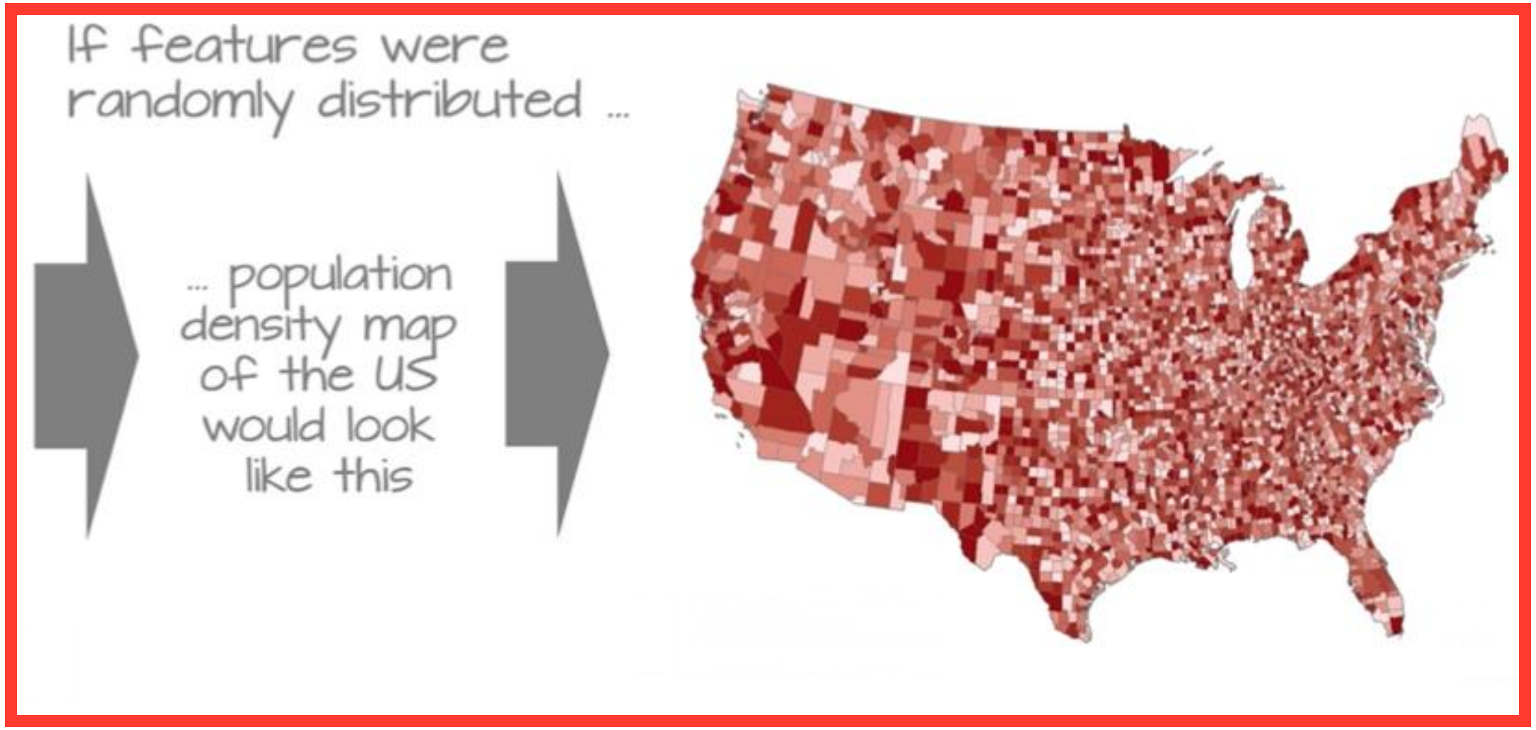
Scenario 1: US Population Density


If features were randomly distributed ...
... population density map of the US would look like this

- **Null hypothesis:** The spatial patterns for the US Population Density are independent. They are random. **[Here, we would reject the null hypothesis]**
- **Alternative hypothesis:** The patterns for the US Population Density are not random. They are indeed clustered. **[Here, we would accept the alternative hypothesis]**

We can apply these hypotheses tests on these scenarios



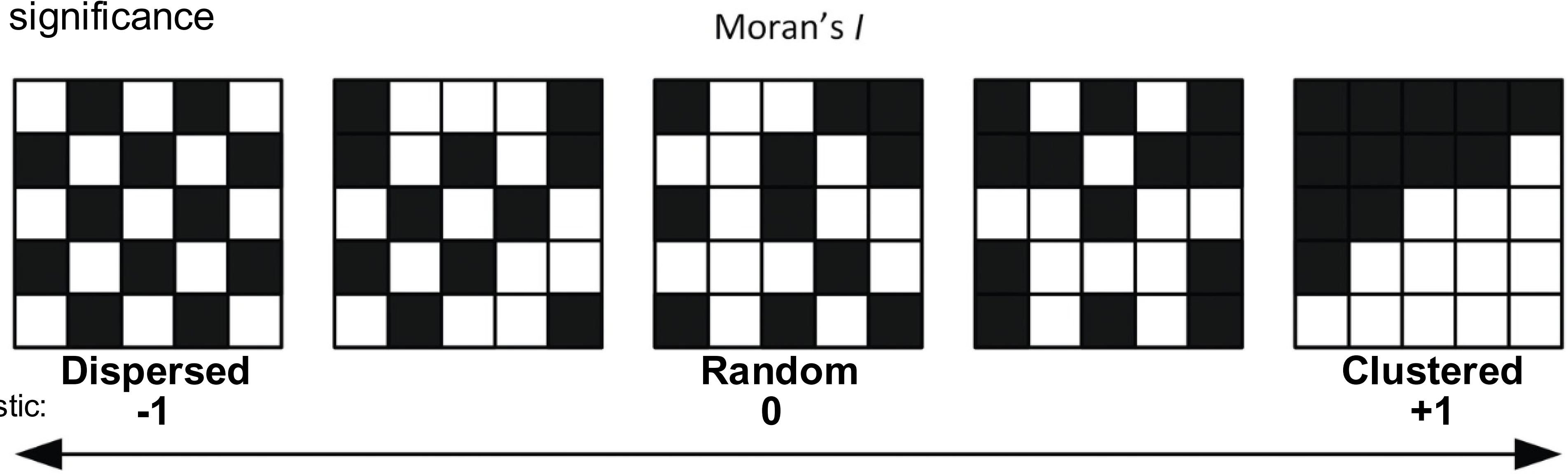

If features were randomly distributed ...
... population density map of the US would look like this

Scenario 2: US Population Density

- **Null hypothesis:** The spatial patterns for the US Population Density are independent. They are random. **[Here, we would accept the null hypothesis]**
- **Alternative hypothesis:** The patterns for the US Population Density are not random. They are indeed clustered. **[Here, we would reject the alternative hypothesis]**

Statistical analysis: Moran’s I statistic [1]

In testing for evidence of spatial autocorrelation, the **Moran’s I statistic,** which is a weighted correlation coefficient, is the statistical test used to detect departures from spatial randomness. Departures from randomness indicate spatial patterns such as clusters (+) or dispersed (-). This statistic is accompanied with a p-value to determine significance



Moran’s I statistic:

Dispersed
-1

Random
0

Clustered
+1

- Positive Moran’s I value is when spatial autocorrelation generally indicates that nearby area have similar values, indicating spatial clustering. **This is a spatial pattern!**

- Negative Moran’s I spatial autocorrelation generally indicates that nearby area have dissimilar values, this is dispersion of values is indeed a **spatial pattern!**

- A Moran’s I value closer to 0 means no evidence of spatial autocorrelation. **No discernible spatial pattern!**

Statistical analysis: Moran’s I statistic [2]

In testing for evidence of spatial autocorrelation, the **Moran’s I statistic,** which is a weighted correlation coefficient, is the statistical test used to detect departures from spatial randomness. Departures from randomness indicate spatial patterns such as clusters (+) or dispersed (-). This statistic is accompanied with a p-value to determine significance

There are two ways for statistically measuring spatial autocorrelation:

- **Global Moran’s I statistic**: What is the overall spatial dependence across the entire data set area? Studying at a global level will tell you how clustered, dispersed or random the data is distributed over the entire area studied.

- **LISA (Local Indicators of Spatial Association)**: What is the difference between each unit of analysis (e.g., areal unit) and its neighbours? We use it for studying at the local level, you can find areas of greater contrast by seeing if places are quantifiably more like or dislike with their neighbours than expected on average.

If the p-value is less than 0.05, we can reject the null hypothesis in favour of the alternative hypothesis and arrive at the conclusion that the patterns are significantly clustered. If the p-value is above 0.05, it means that the clustering/dispersed patterns are not significant, thus we can conclude patterns are random.

Cluster type and categorisation scheme from LISA analysis

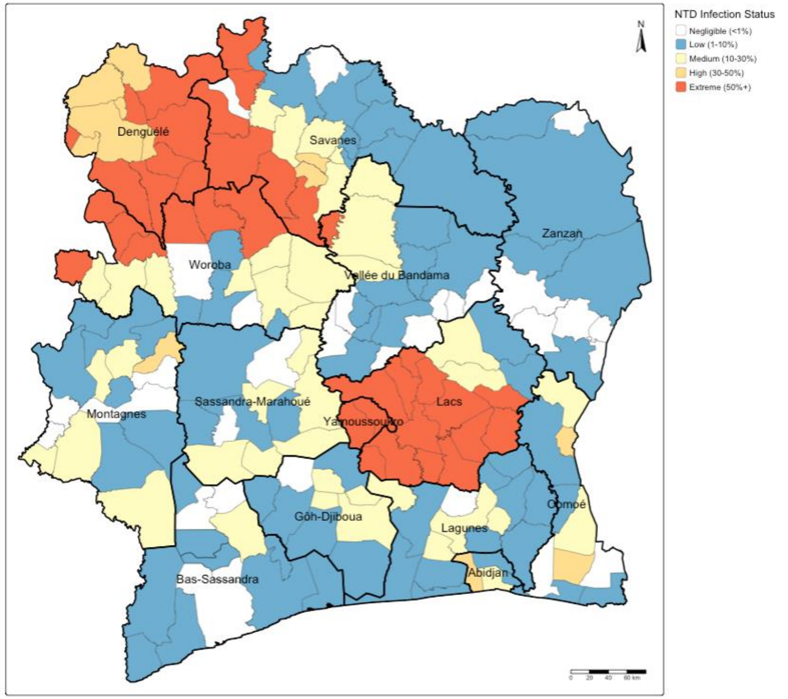
Category	Condition 1		Condition 2		Condition 3	Interpretation
High-High	$z_i > 0$	AND	$z_i^l > 0$	AND	$p_i < 0.05$	Hotspot cluster: Index area is a high-value, which is also surrounded by neighbours with high-values
Low-Low	$z_i < 0$	AND	$z_i^l < 0$	AND	$p_i < 0.05$	Cold-spot cluster: Index area is a low-value, which is also surrounded by neighbours with low-values (observation but lagged)
High-Low	$z_i > 0$	AND	$z_i^l < 0$	AND	$p_i < 0.05$	Mixed cluster: Index area is a high-value, which is also surrounded by neighbours with low-values (i.e., from high to low – spatial outlier)
Low-High	$z_i < 0$	AND	$z_i^l > 0$	AND	$p_i < 0.05$	Mixed cluster: Index area is a low-value, which is also surrounded by neighbours with high-values (i.e., from low to high – spatial outlier)
Not significant					$p_i > 0.05$	No local spatial clustering detected

z_i = Standardised observed value in location i

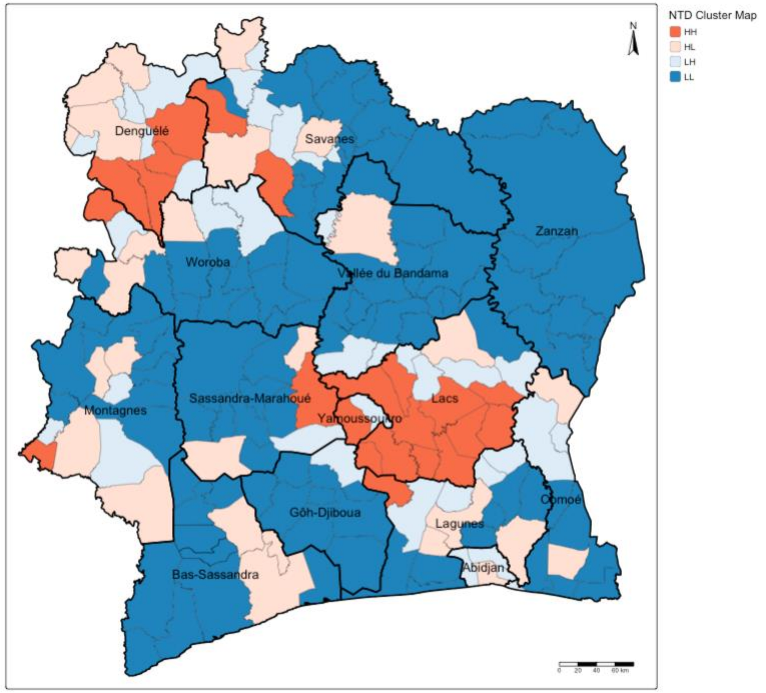
z_i^l = Corresponding standardised observed value(s) in neighbours for location i

p_i = P-value for statistical significance

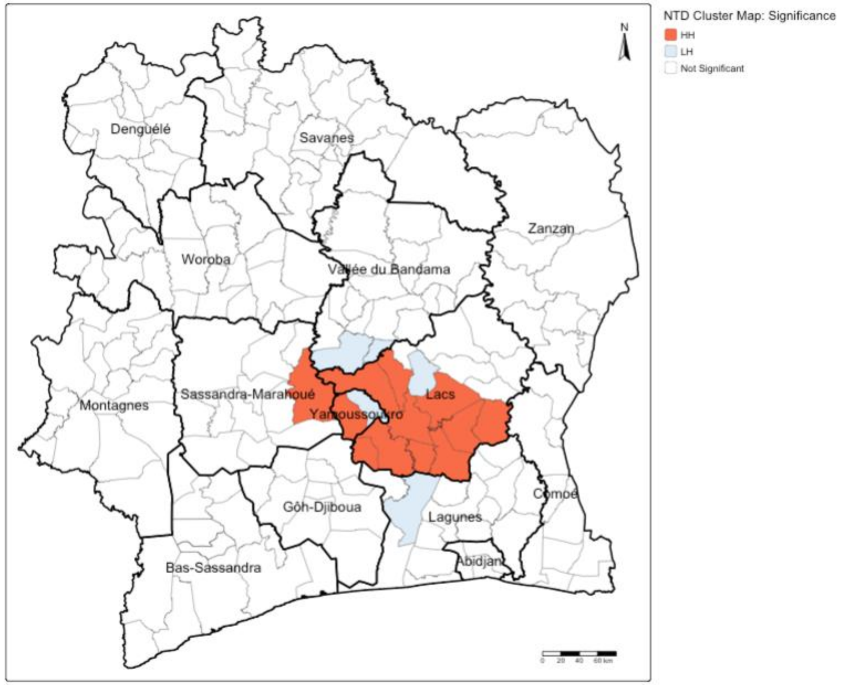
Example: Detection of NTDs clusters in Cote d'Ivoire



A. Distribution of NTDs by Endemicity status



B. Shows where the cluster of NTDs are concentrated – classed as 'High-High', 'High-Low', 'Low-High', 'Low-Low'.



C. Shows which of the cluster classes where NTDs are concentrated are statistically significant

We will generate this output in the computer practicals

Let’s go to the practicals!