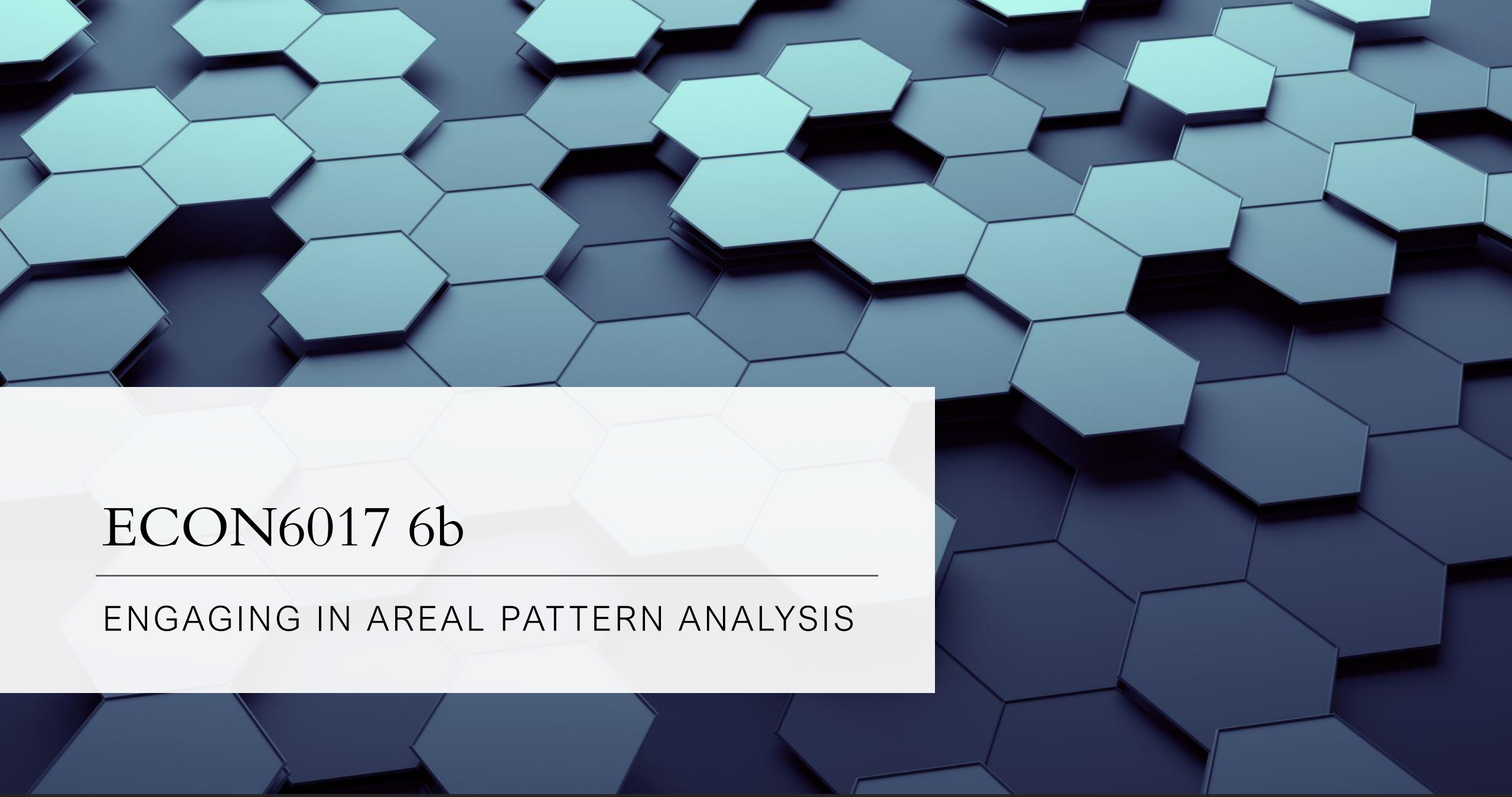


ECON6017 6b

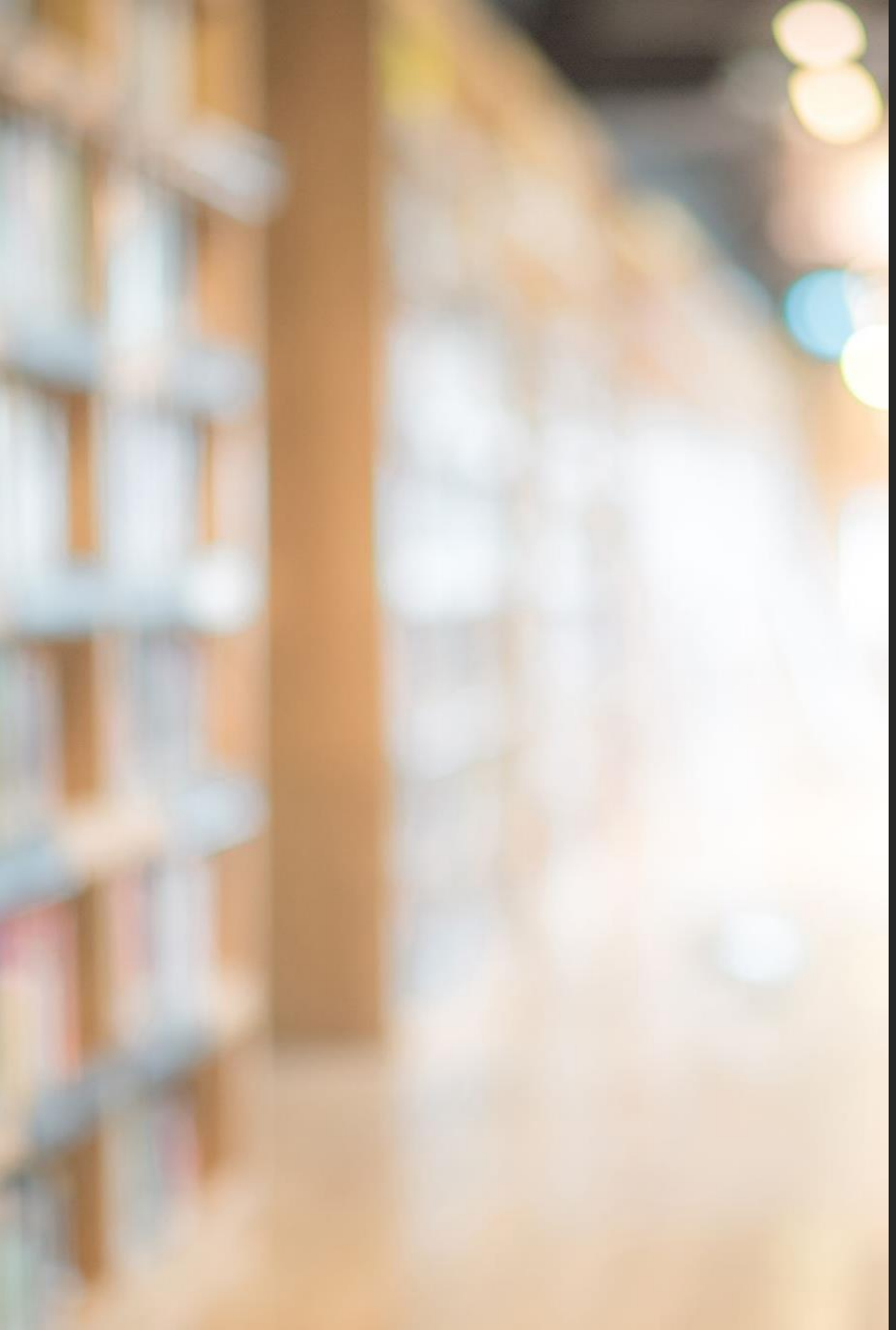
ENGAGING IN AREAL PATTERN ANALYSIS

Packages you need

sf

spdep

tmap (optional for maps)



Rationale for studying areal patterns

As with point patterns, data aggregated at areal levels can be

- clustered
- dispersed or
- random

Our task is to formulate a statistical hypothesis of complete spatial randomness and then validate this based on the empirical observations.

If the pattern is non-random, we then proceed to uncover the processes that underlie the observed pattern (which is the official party line of "spatial econometrics", where observations are not selected independently from one another…)

Spatial autocorrelation and neighbours

Spatial relationships are best modelled based on the principle of spatial neighbours.

Spatial autocorrelation, which is the correlation of a random variable with itself in ”space”, can be measured for both points and areal spatial patterns.

Strong spatial autocorrelation means attributes of adjacent geographic units are strongly related.

- Positive spatial autocorrelation indicates that similar values appear close to each other, or cluster.
- Negative spatial autocorrelation indicates that neighbouring values are dissimilar (or similar values are dispersed).
- Null spatial autocorrelation indicates that the spatial pattern is a CSR pattern.

It is good to start by examining the ***degree of spatial autocorrelation*** before trying to uncover and model the processes that underlie the observed pattern.

Quantifying spatial autocorrelation effects in areal patterns

We will use the following methods to quantify **spatial autocorrelation coefficient**:

- Join count analysis (for binary data)
- Global and local Moran's I statistic
- Global and local Geary's C statistic
- Global and local Getis-Ord's G statistic

These measure capture the **degree of dependence of one variable at a given location to similar variables at neighbouring locations.**

If these measures are similar and statistically significant, then we conclude that **positive spatial autocorrelation** exists in the spatial distribution.

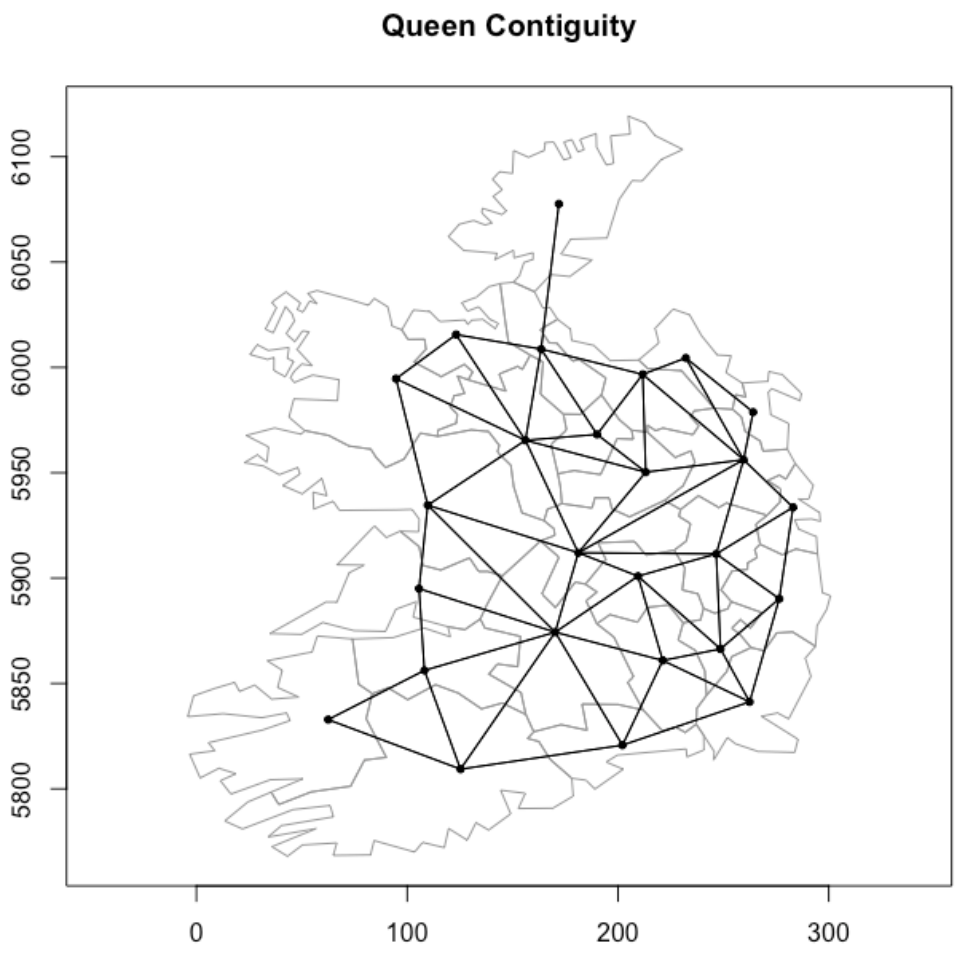
When values in neighbouring locations exhibit different characteristics, then we conclude spatial autocorrelation is weak.

Eire Dataset

Data for the counties of the Irish Republic but omit Dublin from analyses. This data frame contains the following columns:

- A: Percentage of sample with blood group A
- towns: Towns/unit area
- pale: Beyond the Pale 0, within the Pale 1
- size: number of blood type samples
- ROADACC: arterial road network accessibility in 1961
- OWNCONS: percentage in value terms of gross agricultural output of each county consumed by itself
- POPCHG: 1961 population as percentage of 1926
- RETSALE: value of retail sales British Pound000
- INCOME: total personal income British Pound000
- names: County names

https://nowosad.github.io/spData/reference/eire.html



Prep the data and create nb object

```
(eire =
st_read(system.file("shapes/eire.shp",
package="spData")))

(eire_nb = poly2nb(eire))

coords =
st_point_on_surface(st_geometry(eire))

plot(st_geometry(eire), border="grey60",
axes=T,  main="Queen Contiguity")

plot(eire_nb, coords, pch=19, cex=0.6,
add=T)
```

Initial observations

```
> table(st_is_valid(eire))

> plot(st_geometry(eire))

> st_crs(eire)$proj
```

All contiguous polygons that are valid and projected.

We are good to go!

Create weights list objects

row normalised weights

```
> eire_lw = nb2listw(eire_nb)
```

Binary weights for join count analysis

```
> eire_lwB = nb2listw(eire_nb,style="B")
```

Join count analysis

Join count statistics for binary attributes (categorical)

The main methodology to quantify the relationship between similar or dissimilar attributes in adjacent areas of a binary variable (1/0) is done by assigning two colours: black (B) for attribute 1 and white (W) for attribute 0.

If attribute 1 occurs in one area, then the area will be assigned B, if attribute 0 occurs then the area will be assigned W.

Two neighbouring areas are considered “joined”. There are 3 possible **types of joins**:

- BB
- BW
- WW

Join counts tally the numbers of BB, BW and WW in the study area.

Join count statistics can show

Positive spatial autocorrelation (clustering) if the number of BW joins is significantly lower than what we would expect by chance.

Negative spatial autocorrelation (dispersion) if the number of BW joins is significantly higher than what we would expect by chance,

CSR if the number of BW joins is approximately the same as what we would expect by chance.

Join counts

If we have n_B Black units and $n_W = n - n_B$ White, the respective probabilities of observing the two types of units are:

$$P_B = \frac{n_B}{n} \qquad \text{and} \qquad P_W = \frac{n-n_B}{n} = 1 - P_B$$

Probability of BB, WW and BW in adjacent cells are (assuming independence/CSR):

$$P_{BB} = P_B P_B = P_B^2$$

$$P_{WW} = (1 - P_B)^2$$

$$P_{BW} = P_B(1 - P_B) + (1 - P_B)P_B = 2P_B(1 - P_B)$$

Expected join counts

The expected join counts are:

$$E(BB) = \frac{1}{2}\sum_i \sum_j w_{ij} P_B^2$$

$$E(WW) = \frac{1}{2}\sum_i \sum_j w_{ij} (1 - P_B)^2$$

$$E(BW) = \frac{1}{2}\sum_i \sum_j w_{ij} 2P_B (1 - P_B)$$

Division by 2 is because neighbourhood relationship is 2-way. Thus, $\frac{1}{2}\sum_i \sum_j w_{ij}$ is the total number of joins/links (of any type), **assuming a binary weights** matrix (basically, we are adding all the 1's in the W matrix!).

Observed join counts

$$BB = \frac{1}{2}\sum_i \sum_j w_{ij} x_i x_j$$

$$WW = \frac{1}{2}\sum_i \sum_j w_{ij} (1 - x_i)(1 - x_j)$$

$$BW = \frac{1}{2}\sum_i \sum_j w_{ij} \left(x_i - x_j\right)^2$$

Where $x_i = \begin{cases} 1 & if\ i = B \\ 0 & if\ i = W \end{cases}$ and w_{ij} is coming from the spatial weights matrix.

We compare the observed join counts to the expected join counts from a CSR pattern.

The variance of join counts

The variance of BW is,

$$\sigma_{BW}^2 = E(BW^2) - E^2(BW)$$

$$= \frac{1}{4}\left(\frac{2S_2 n_B(n-n_B)}{n(n-1)} + \frac{(S_3 - S_1)n_B(n-n_B)}{n(n-1)} + 4\frac{(S_1^2 + S_2 - S_3)n_B(n-n_B)(n_B-1)(n-n_B-1)}{n(n-1)(n-2)(n-3)}\right) - E^2(BW)$$

where,

$S_1 = \sum_i \sum_j w_{ij}$,

$S_2 = \sum_i \sum_j \left(w_{ij} - w_{ji}\right)^2$,

$S_3 = \sum_i \left(\sum_j w_{ij} + \sum_j w_{ji}\right)^2$.

Z test for join counts

Z test for each join count is calculated as:

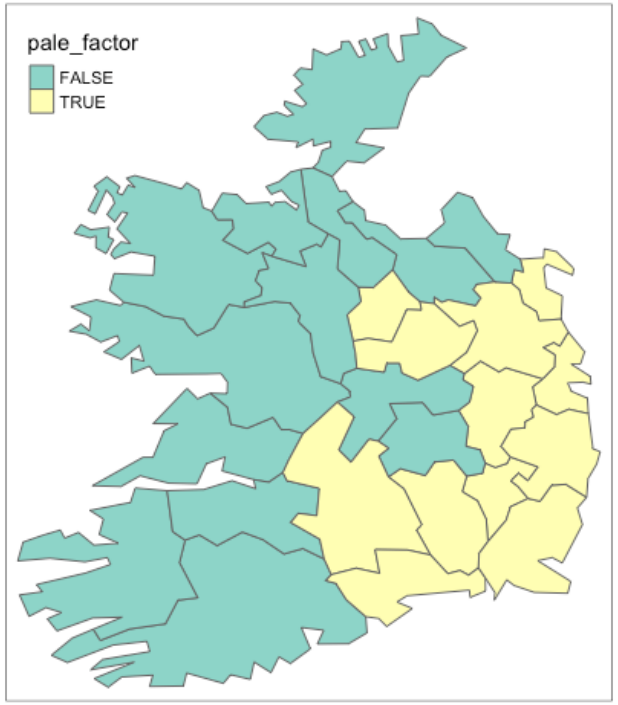
$$Z_{BW} = \frac{BW - E(BW)}{\sigma_{BW}} \sim N(0,1)$$

The join count statistic is asymptotically normally distributed under the null hypothesis of no spatial autocorrelation.

Join count analysis in R

Pale is a binary variable. The Pale or the English Pale was the part of Ireland directly under the control of the English government in the late middle ages.

```
> eire$pale # pale is binary

> eire$pale_factor = as.factor(eire$pale ==1 )

> summary(eire$pale_factor)

FALSE  TRUE

   14    12

> tm_shape(eire) + tm_polygons(col="pale_factor")
```



Join count test

The derivation of the test assumes that the weights matrix is symmetric. **Binary weights are symmetric.**

```
> joincount.multi(eire$pale_factor, listw=eire_lwB)
```

The observed BW count (T:F = 21) is significantly less than the expected value under CSR (T:F = 29.5). Sufficient evidence to suggest significant clustering at 5% level of significance.

Conclusion: Reject H0 at 5% level of significance.

Jtot is the count of all different-colour joins.

```
                Joincount Expected Variance z-value
FALSE:FALSE   18.0000  15.9600   8.4175  0.7031
TRUE:TRUE     18.0000  11.5754   6.8634  2.4523
TRUE:FALSE    21.0000  29.4646  11.9130 -2.4524
Jtot          21.0000  29.4646  11.9130 -2.4524
```

Let's verify the numbers in the output

1. Total number of neighbourhood relationships, $\sum_i \sum_j w_{ij}$ is given by:

```
> eire_lwB
Number of nonzero links: 114 (=2*(18+18+21))
```

2. Number of TRUE and FALSE counts

```
> summary(eire$pale_factor)
FALSE  TRUE
   14    12
```

3. Expected number of TRUE:FALSE counts (adjusted for finite samples):

$$\text{E(BW)} = \frac{1}{2} \times 114 \times 2 \times \frac{14}{26} \times \frac{12}{25} = 29.46462$$

Try this!

A. Verify the expected counts for FALSE:FALSE and TRUE:TRUE as well?

B. Count the number of TRUE:FALSE counts in the map and verify that it is indeed 21.

Pros and cons

Join count statistics offer an easy way to represent spatial distribution.

However,

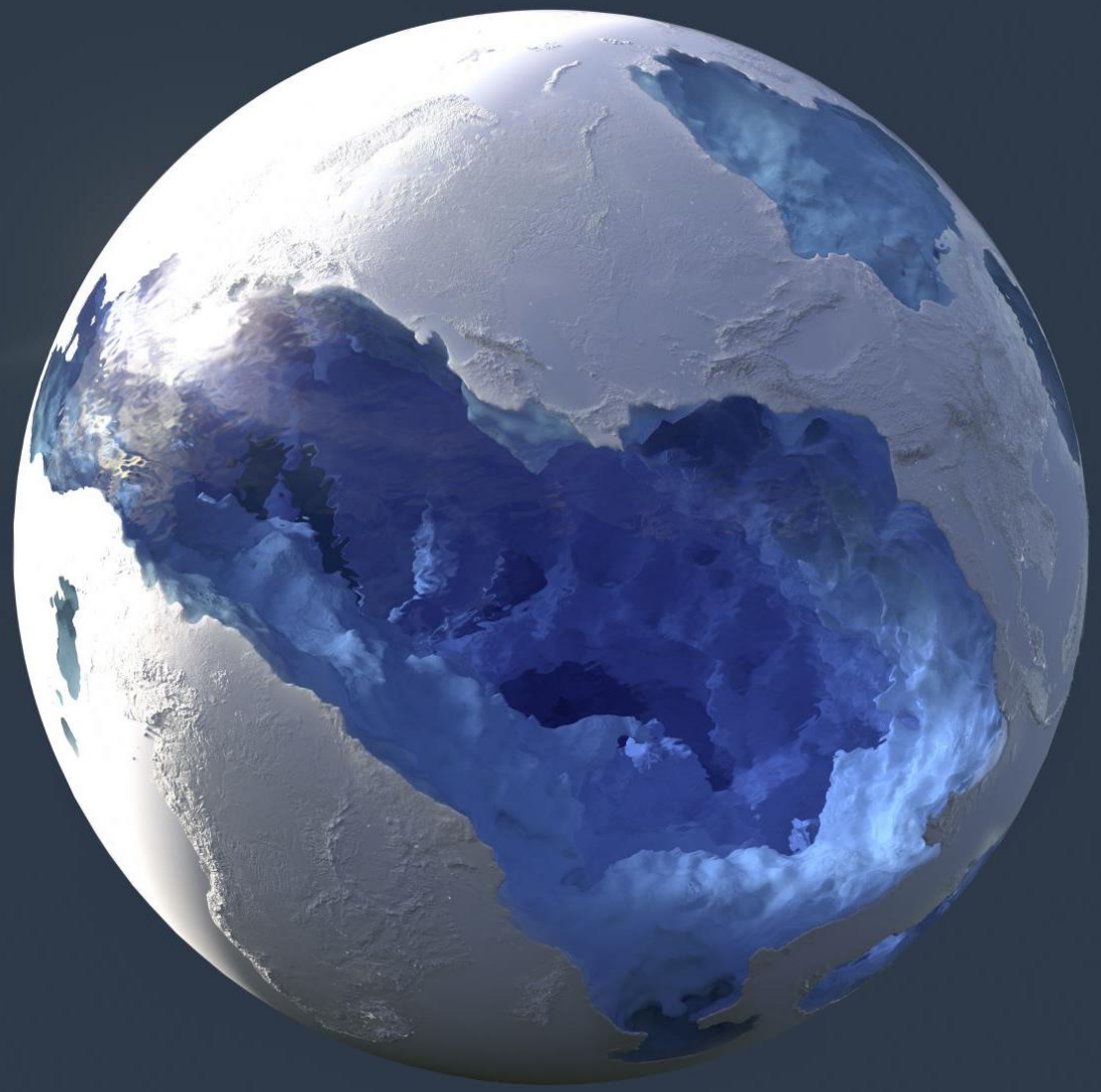
- it can only be applied to nominal data (a rudimentary method for numerical data is to convert to factor and conduct join count analysis)
- does not provide a simple summary measure.

Exercise caution when converting a continuous variable into binary variable and applying join counts to measure spatial autocorrelation.

Moran's I statistic

GLOBAL MORAN'S I

LOCAL MORAN'S I



Moran's I

Moran's I measures the degree of spatial autocorrelation in ratio and interval measured data (a.k.a. numerical data).

Moran's I statistic is arguably the **most used indicator** of global spatial autocorrelation.

The computation of Moran's I is achieved by dividing the spatial covariation by the total variation. In essence, it is a **cross-product statistic between a variable and its spatial lag**, with the variable expressed in deviations from its mean.

The resulting value ranges from -1 (perfect dispersion) to +1 (perfect clustering).

Global Moran's I Statistic

Suppose we have a study region *R*, divided into *n* areal units.

(Global) Moran's I is calculated as follows:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

For a row standardised W matrix, $\sum_i \sum_j w_{ij} = n$ and hence the Moran's I statistic simplifies to

$$I = \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Inference for Moran's I is based on a null hypothesis of spatial randomness.

Global Moran's I test statistic

$$Z = \frac{I - E(I)}{s.e.(I)} \sim N(0,1)$$

where, $E(I) = -\frac{1}{n-1}$ under the null hypothesis of no autocorrelation and

$$Var(I) = \frac{nS_1 - S_2 S_3}{(n-1)(n-2)(n-3)\left(\sum_i \sum_j w_{ij}\right)^2}$$

where,

$$S_1 = \frac{1}{2}(n^2 - 3n + 3)\sum_i \sum \left(w_{ij} + w_{ji}\right)^2 - n\sum_i \left(\sum_j w_{ij} + \sum_j w_{ji}\right)^2 + 3\left(\sum_i \sum_j w_{ij}\right)^2$$

$$S_2 = n\frac{\sum_i (x_i - \bar{x})^4}{\left(\sum_i (x_i - \bar{x})^2\right)^2}$$

$$S_3 = \frac{1-2n}{2}\sum_i \sum \left(w_{ij} + w_{ji}\right)^2 + 6\left(\sum_i \sum_j w_{ij}\right)^2$$

Global Moran's I for Eire data

There are 26 features in this sf object.

Given that $E(I) = -\frac{1}{n-1}$, the null hypothesis is,

$$H_0: E(I) = -\frac{1}{n-1} = -\frac{1}{25} = -0.04$$

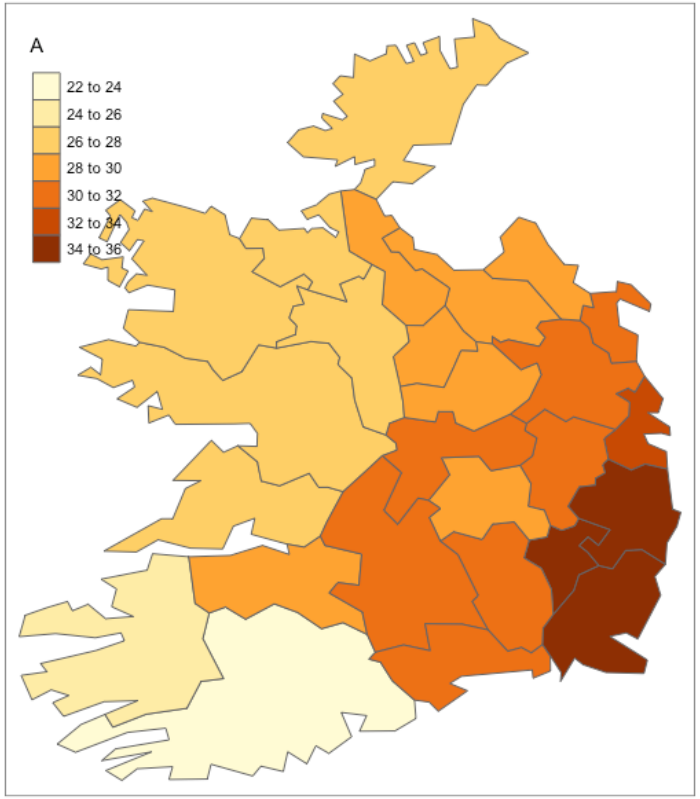
$$H_1: E(I) > -0.04$$

The default alternative hypothesis in *r* is positive autocorrelation (you can change it to negative or two tailed tests)

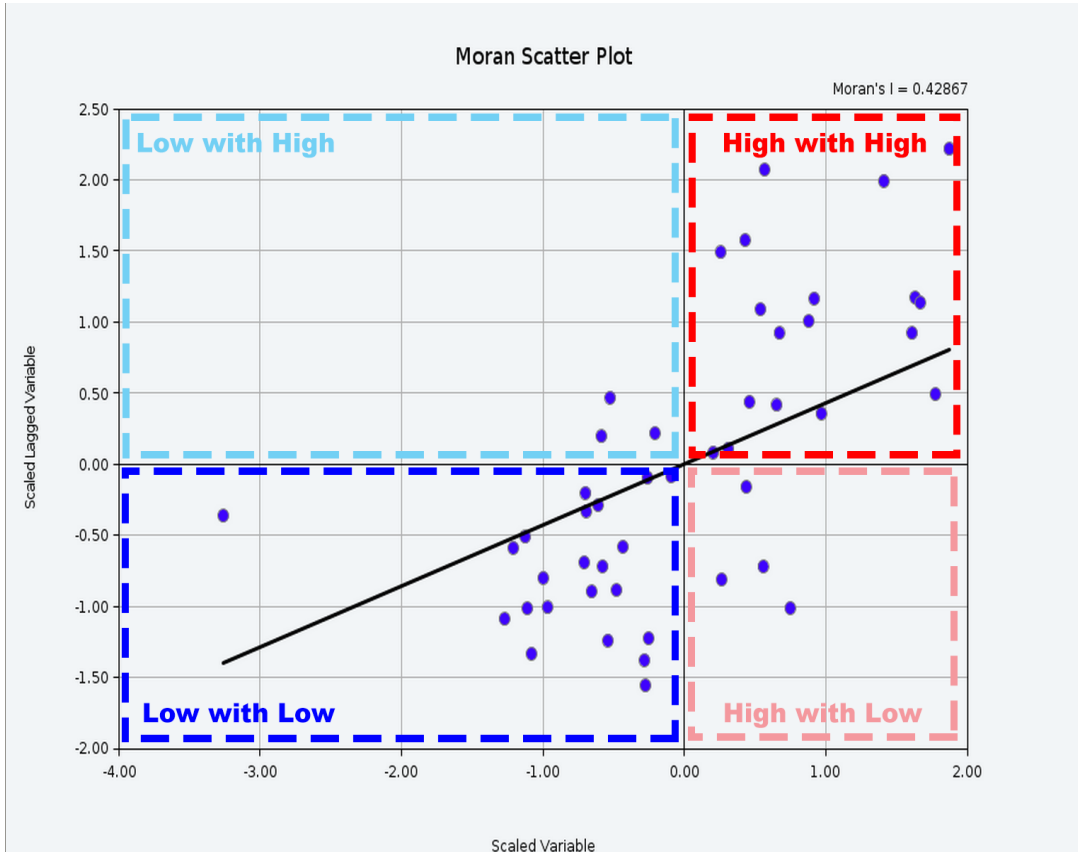
Global Moran's I in R

A is the percentage of sample with blood group A

```
> tm_shape(eire) + tm_polygons(col="A")

> moran.test(eire$A, listw=eire_lw)

Moran I statistic standard deviate = 4.6851, p-value = 1.399e-06

alternative hypothesis: greater

sample estimates:

Moran I statistic       Expectation (H0)     Variance

       0.55412382       -0.04000000        0.01608138
```



Conclusion: Reject the null of no spatial autocorrelation in favour of positive spatial autocorrelation. There seem to be a significant spatial patterning of the in the incidence of group A type blood at county level (for whatever reason).



Moran's scatterplot

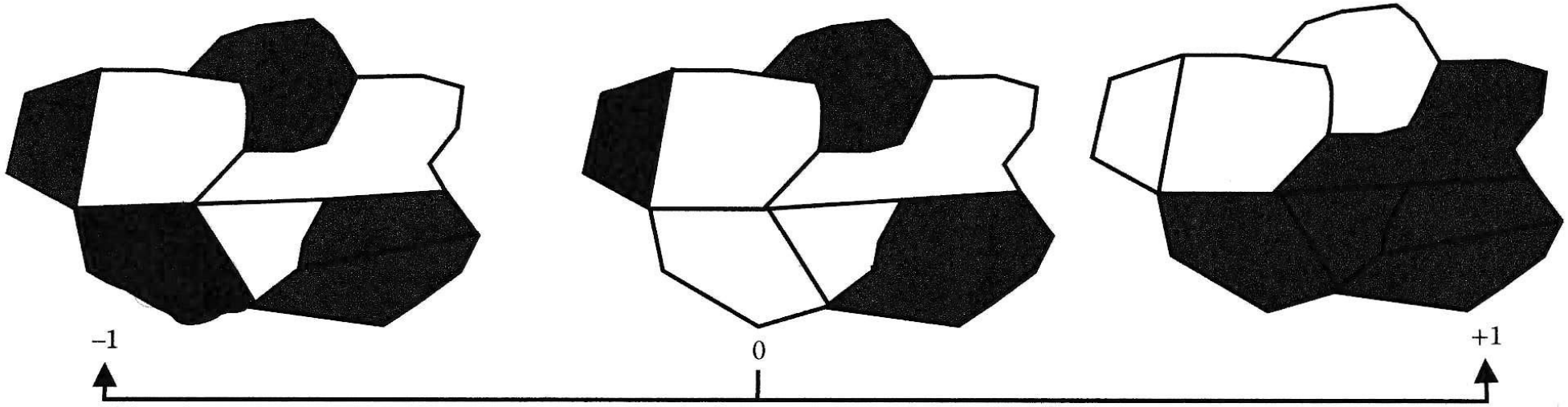
The Moran scatter plot, consists of a plot with the spatially lagged variable on the y-axis and the original variable on the x-axis.

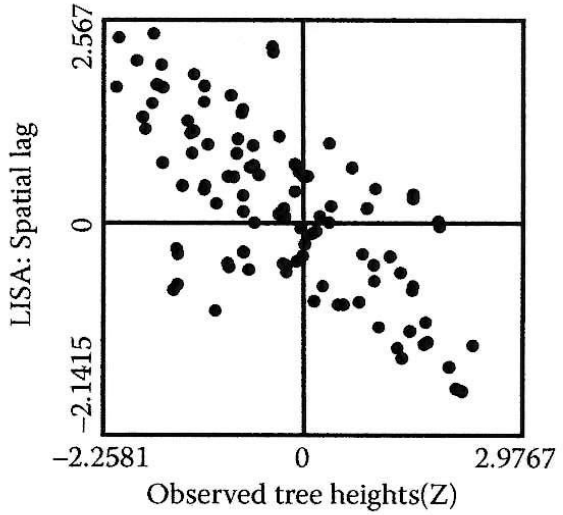
The slope of the linear fit to the scatter plot equals Moran's I.

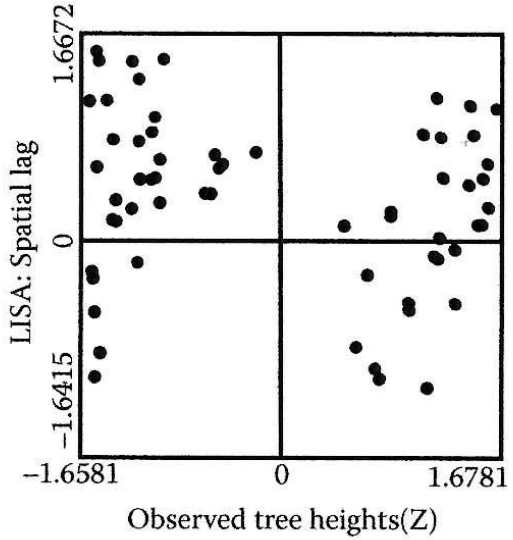
The plot has 4 quadrants centred at $(\bar{x}, \frac{\sum_i w_{ij} x_i}{n})$ which can be re-scaled to have an origin of (0,0).

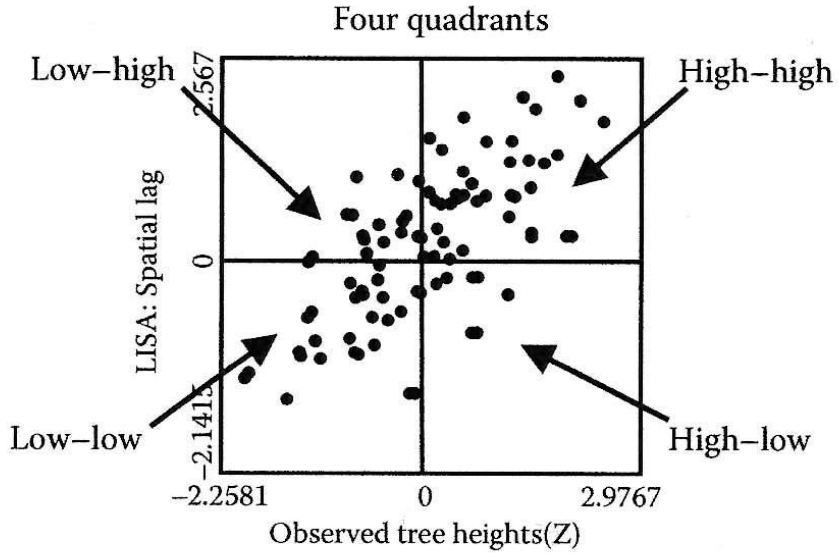
- High-High: neighbours with high values surrounded by those with high values (**hotpots**)
- High-Low/Low-High: neighbours with high values surrounded by those with low values in other words (**spatial outliers)**
- Low-Low: neighbours with low values surrounded by those with low values (**coldspots**)

Understanding spatial autocorrelation with Moran's scatterplot









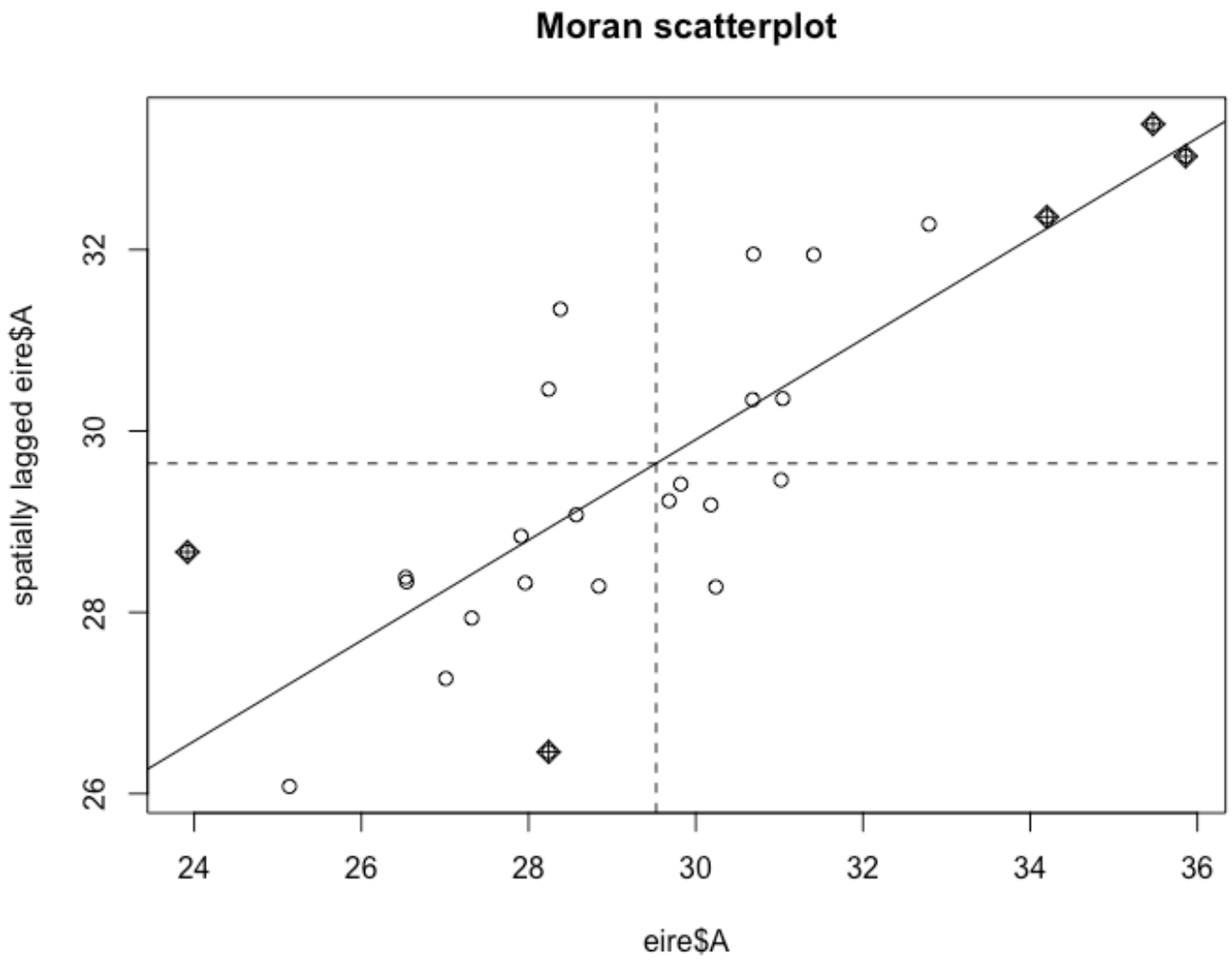
High–high, Low–low = Spatial clusters

High–low, Low–high = Spatial outliers

Moran's scatterplot in R

```
> moran.plot(eire$A,
listw=eire_lw,
main="Moran scatterplot",
labels=F)
```

The significant positive spatial autocorrelation suggested by the test can be visually confirmed here (In statistical analysis the scatterplot comes first which is what you must do when you conduct your own analysis.)



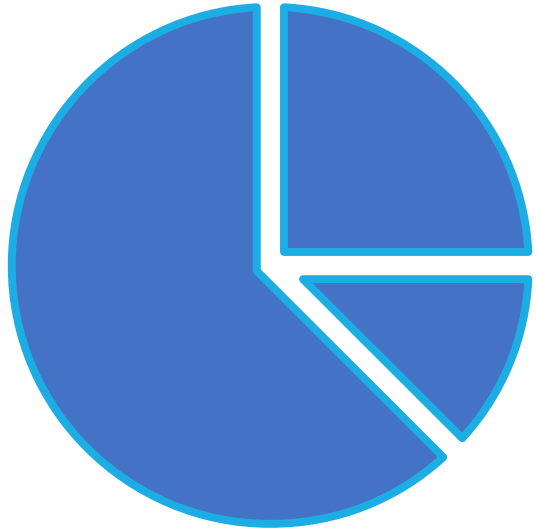
Local Moran's I statistic

Global Moran's I statistic for spatial autocorrelation is calculated from local relationships between observed values at spatial units and their neighbours.

It is possible to break these measures down into their components (at a location specific level), thus constructing local Moran's I statistic for spatial autocorrelation.

This belongs to a family of "local indicators for spatial association (LISA) that can be used to identify the degree to which one areal unit is autocorrelated relative to it's neighbours:

- Detect spatial clusters or outliers at a local level
- Measure spatial autocorrelation at a local level
- Identify spatial patterns or hotpots at a local level

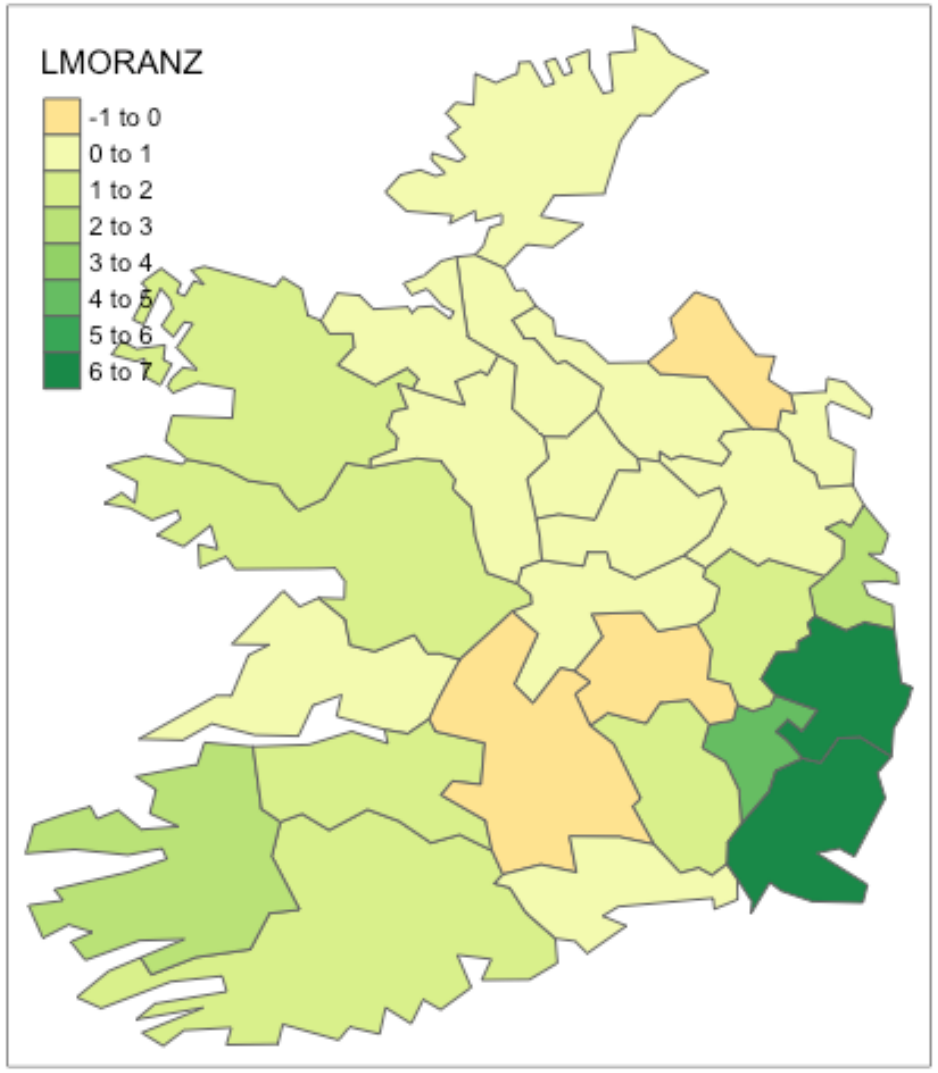


Local Moran's I statistic

Local Moran's I statistic is constructed as one of the n components which comprise the global test:

$$I_i = (x_i - \bar{x}) \frac{\sum_j w_{ij} (x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2 / n}$$

Sum of the local Moran's I statistics is proportional to the global Moran's I statistic.



Local Moran's I in R

```
localm = localmoran(eire$A, listw=eire_lw)

head(localm)

eire$LMORANZ = localm[,4]

tm_shape(eire) + tm_polygons(col="LMORANZ")
```

Global Moran as the sum of Local Morans

```
sum(localm[,1])/
sum(unlist(eire_lw$weights))
```

[1] 0.5541238

Potential drawbacks of Moran's I

Moran's I statistic is highly sensitive to spatial patterning in the variable of interest. Spatial autocorrelation values do not always produce useful insights into the DGP.

Moran's I statistic is highly **sensitive to the choice of spatial weights** matrix. Where the weights do not reflect the "true" structure of spatial interaction, estimated autocorrelation (or lack thereof) may be due to misspecification.

- Example: the computed Moran's I changes when the weight matrix is changed from row standardised to binary. `> moran.test(eire$A, listw=eire_lwB)`
- The computation assumes there are no neighbour-less units in the study area (zero.policy=F).

These drawbacks are in general common to all forms of spatial autocorrelation statistics (including Gear's C, up next).

Drawbacks contd.

Notice that the test is sensitive to the weights "style" being used. **Row-standardisation in general favours outcomes with fewer neighbours (i.e. assign a higher weight for those), thus maybe biased towards the spatial units along the edges**. At present there is **no technique in the literature to make a reliable edge correction** for row standardised weights. However, in many econometric spatial models, using a row standardised W matrix is the default option (which you may observe in the "spreg" and "spdep" function coding as well).

Activity A

Establish the relationship between the global and local Moran's I statistics.

Using the row standardised version of the global Moran's I statistic, to show that the slope of the linear fit to the Moran's scatter plot equals Moran's I. (Hint: https://geodacenter.github.io/workbook/5a_global_auto/lab5a.html#fn4)

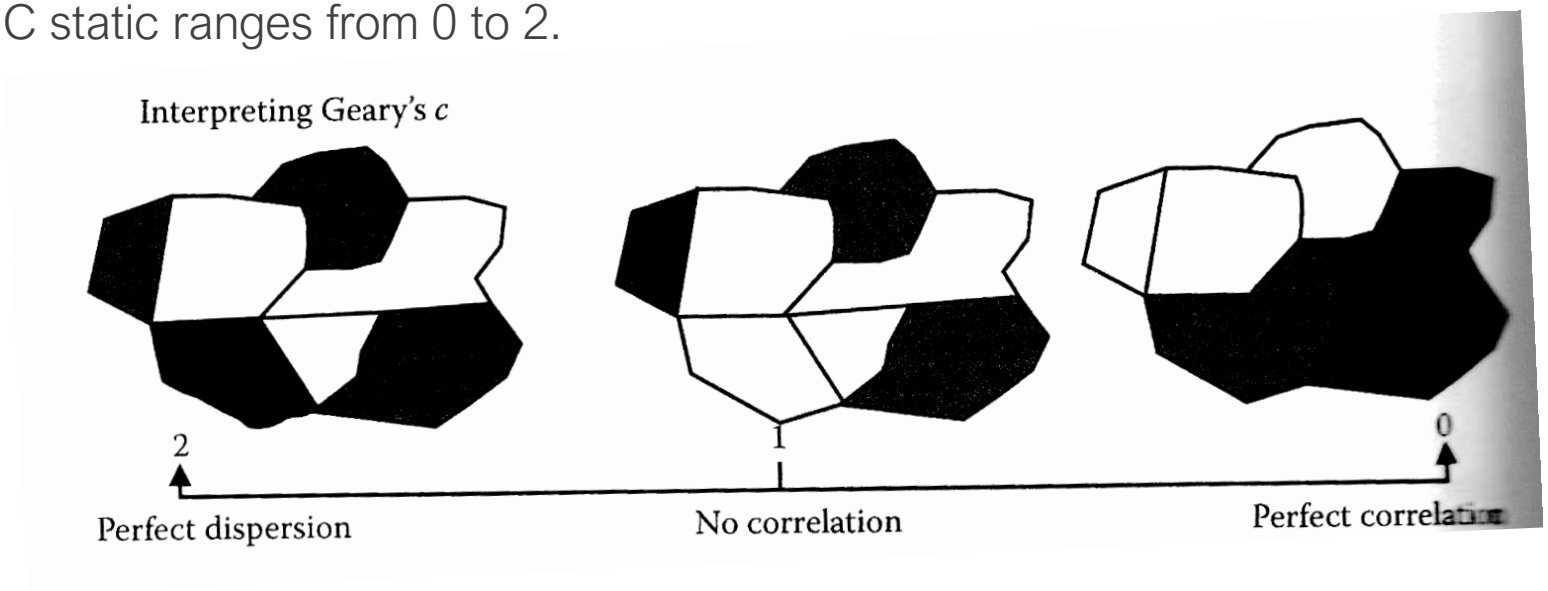
Geary's C statistic

Global Geary's C

Geary's C is an alternative measure of spatial autocorrelation.

It determines the degree of spatial autocorrelation using **sum of squared differences paired data values as its measure of covariance**.

The Geary's C static ranges from 0 to 2.



Global Geary's C statistic

Suppose we have a study region *R*, that is divided into *n* cells, Global Gear's C is computed by,

$$C = \frac{n-1}{2\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}\left(x_i - x_j\right)^2}{\sum_i (x_i - \bar{x})^2}$$

Under the null hypothesis of CSR, $C = 1$.

The statistical significance of Geary's C is also based on a standard normal distribution.

Geary's C vs. Moran's I

Moran's I is based on cross product of deviations from the mean of a variable at a particular unit and a neighbouring unit, while Geary's C is a cross product of actual values of a variable at a particular location and another neighbouring unit. Thus, **Geary's C is less likely to be affected much by extremes.**

Also, Geary's C uses the **unbiased estimator**, $S_x^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$ in the denominator of the estimator compared to the biased estimator $S_x^2 = \frac{1}{n}\sum_i (x_i - \bar{x})^2$ that is used by Moran's I.

Global Geary's C in R

```
> geary.test(eire$A, listw=eire_lw)
```

Geary C statistic standard deviate = 4.5146, p-value = 3.172e-06

alternative hypothesis: Expectation greater than statistic

sample estimates:

Geary C statistic	Expectation	Variance
0.38011971	1.00000000	0.01885309

Conclusion: Reject the null of no spatial autocorrelation (C=1) in favour of positive spatial autocorrelation. (Notice the inverted nature of the alternative hypothesis.)

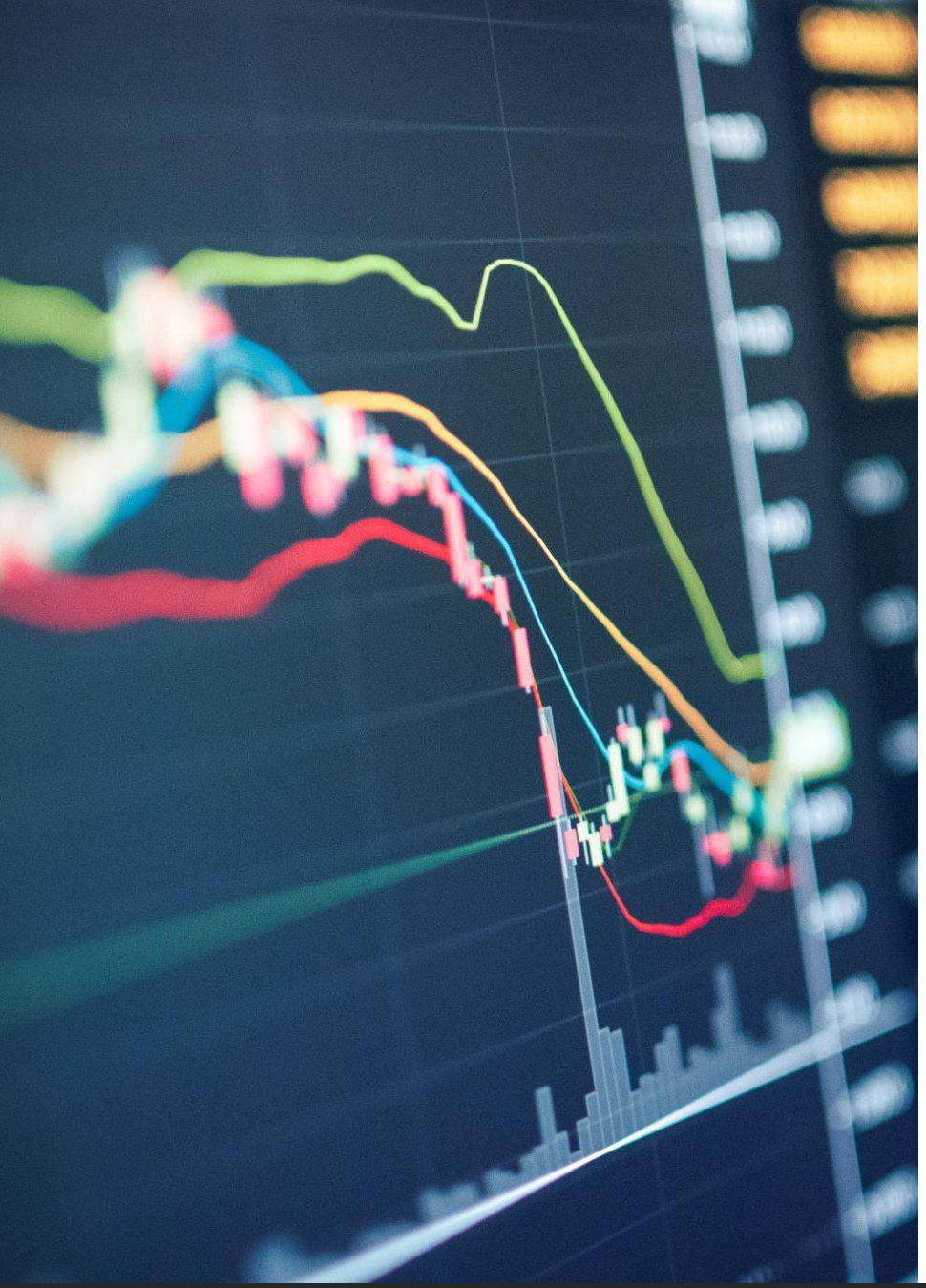
Local Geary's C statistic

Local Gear's C is computed by,

$$C_i = \frac{\sum_j w_{ij}\left(x_i - x_j\right)^2}{\sum_i (x_i - \bar{x})^2/(n-1)}$$

Sum of the local Gear's C statistics is proportional to the global Gear's C statistic, however it has no direct correspondence with the slope in a scatter plot.

Currently there is **no methodology for implementation** of the the local Gear's C in R.



Local Geary's C in R

spdep package does not have a function to compute local Gear's C statistic.

If needed, you may consider the following function from the geocomander package.

local_gearys {geomander}

https://search.r-project.org/CRAN/refmans/geomander/html/local_gearys.html

Getis-Ord's G statistic

Global Getis-Ord's G statistic

Global Getis-Ord's G statistic measures the **concentration of high or low values** for a given study area. It is computed as,

$$G = \frac{\sum_i \sum_j w_{ij} x_i x_j}{\sum_i \sum_j x_i x_j} \quad for\ i \neq j$$

The statistical significance of Getis-Ord's G is also based on a standard normal distribution.

$\mathrm{E}(G) = \frac{\sum_i \sum_j w_{ij}}{n(n-1)} = \frac{1}{n-1}$ when W is row normalised.
The range for General G will be between 0 and 1.

A binary weighting scheme is recommended for this statistic.

Getis-Ord's G statistic in R

```
> globalG.test(eire$A, listw=eire_lwB)

standard deviate = 1.3199, p-value = 0.09344

alternative hypothesis: greater

sample estimates:

Global G statistic  Expectation   Variance

      1.787800e-01  1.753846e-01  6.617796e-06
```

Conclusion: Do not reject the null of no spatial autocorrelation at 5% level of significance.

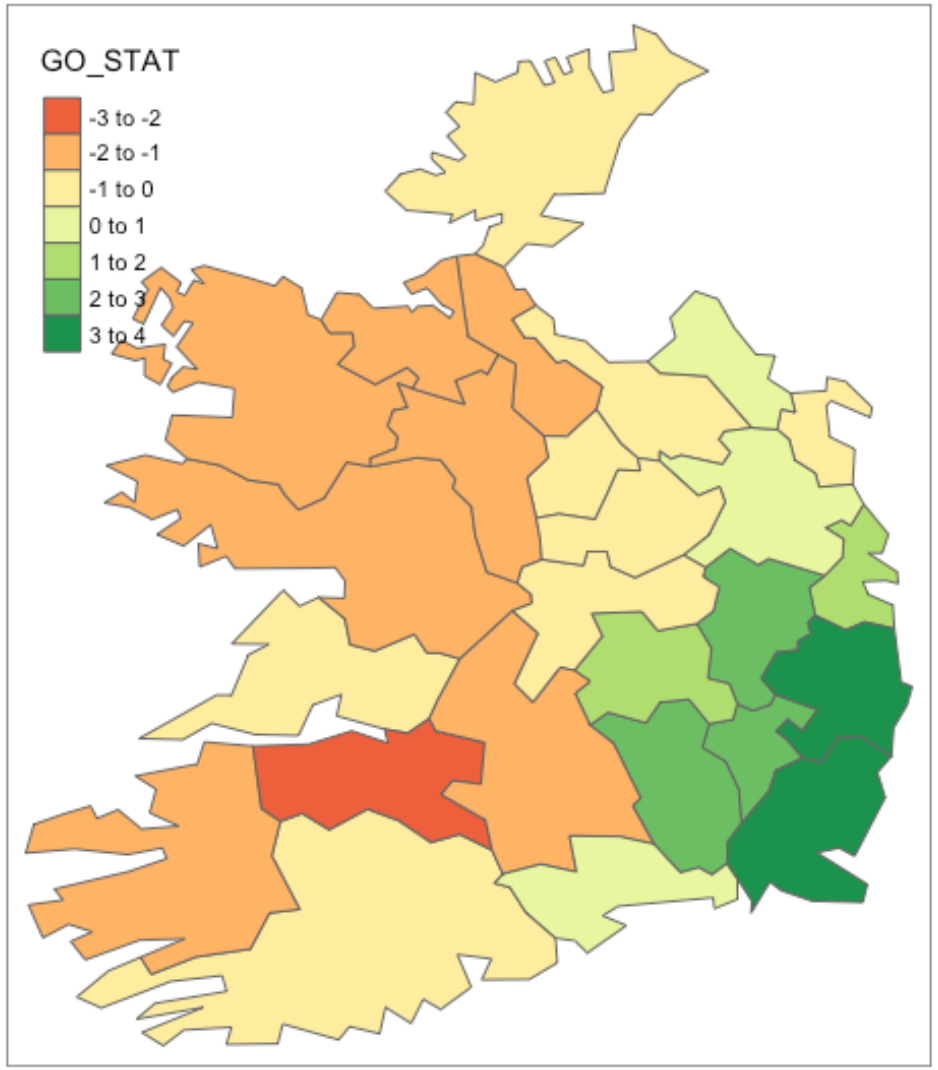
Local Getis-Ord's G statistic

The local Getis-Ord's G statistic consist of a ratio of the weighted average of the values in the neighbouring locations, to the sum of all values, **not including the value at the location** (i).

$$G_i = \frac{\sum_{j \neq i} w_{ij} x_j}{\sum_{j \neq i} x_j}$$

The interpretation of the Local Getis-Ord statistics is very straightforward:

- a value larger than the mean (or, a positive value for a standardized z-value) suggests a high-high cluster or hot spot,
- a value smaller than the mean (or, negative for a z-value) indicates a low-low cluster or cold spot.
- In contrast to the Local Moran and Local Geary statistics, the Getis-Ord approach does not consider spatial outlier



Local Getis-Ord's G statistic in R

```
localG = localG(eire$A, listw=eire_lwB)

eire$GO_STAT = localG

tm_shape(eire) + tm_polygons("GO_STAT")
```

Correlogram

MEASURING THE DEGREE OF SPATIAL CONTAMINATION

Spatial correlogram

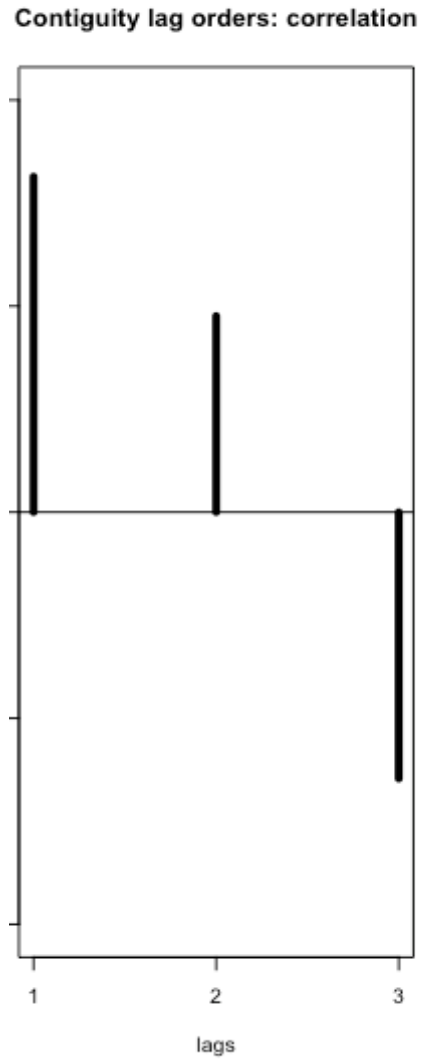
In time series a correlogram (also called Auto Correlation Function ACF Plot or Autocorrelation plot) is a visual way to show *serial correlation* in data that changes over time. Basically it shows for how long (number of lags), temporal memory remains.

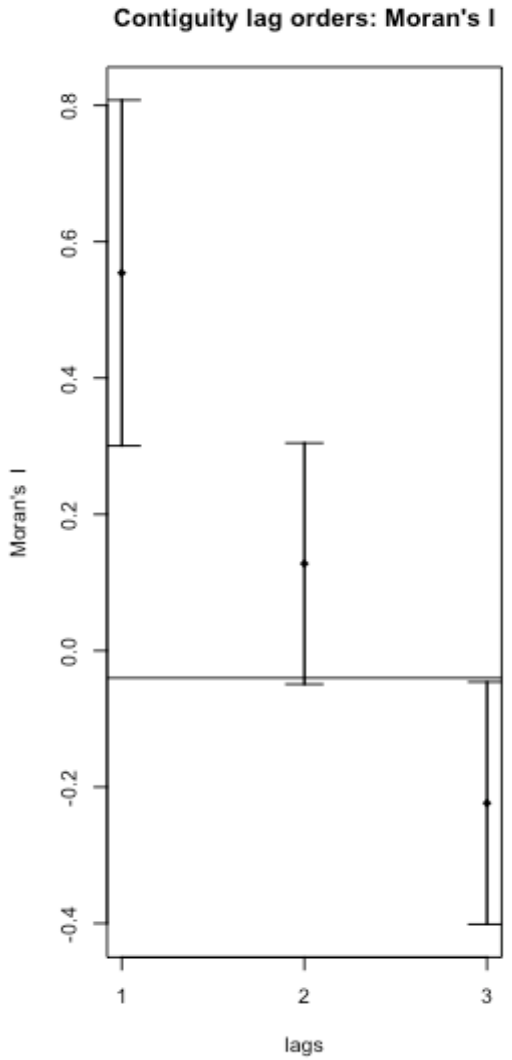
A spatial correlogram is the spatial version of the same plot.

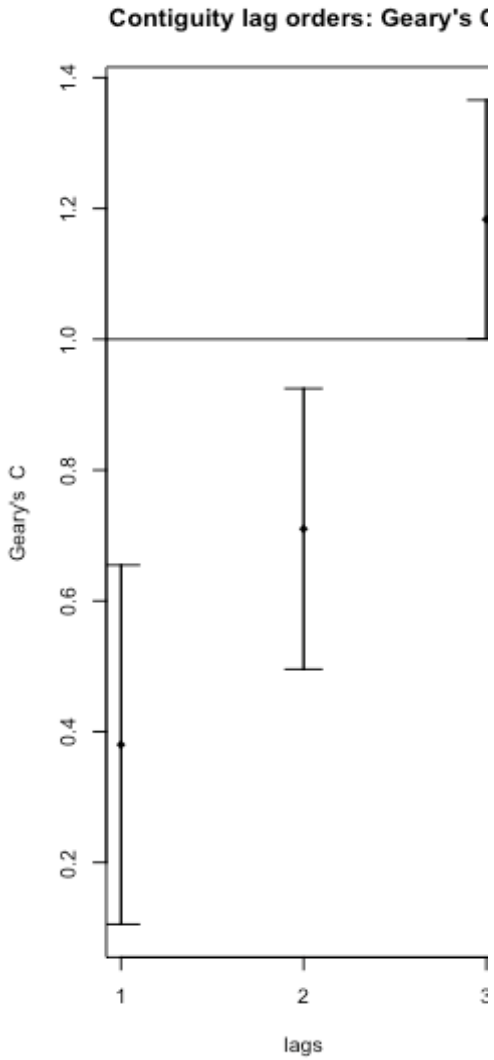
It gives us important information on *how many lags* we should consider when considering spatial regression models (should I only consider my immediate neighbour or should I also consider my neighbour's neighbour as well and so on…!).

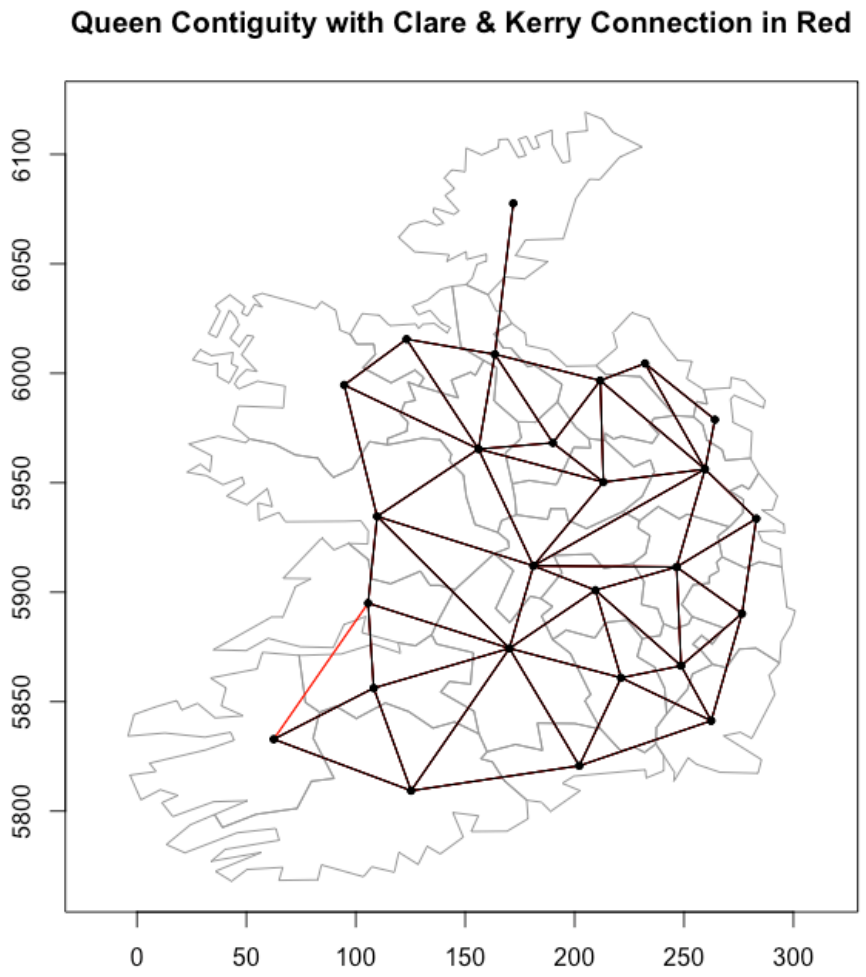
Spatial correlogram in R

```
correl =
sp.correlogram(neighbours=eire_
nb, var=eire$A, order=3,
method="corr", style="W")

correlI =
sp.correlogram(neighbours=eire_
nb, var=eire$A, order=3,
method="I", style="W")

correlC =
sp.correlogram(neighbours=eire_
nb, var=eire$A, order=3,
method="C", style="W")
```









Activity B

The eire_nb object you created uses Queen contiguity form where two units are neighbours if they share a common *land* border. Re-do the above analysis by adjusting the eire_nb object to include a connectivity between the regions “Clare” and “Kerry”. (Hint: Use the “snap” argument in “poly2nb or create a GAL file from eire_nb and manipulate it.)

Take home points

Different mesures spatial autocorrelation

- Join count analysis (for binary data)
- Moran's I statistic
- Geary's C statistic
- Getis-Ord's G statistic

Local vs. global measures of spatial autocorrelation

- Hypothesis tests based on global measures
- Derive the global measures using local measures
- Visualise local measures

Spatial correlogram

Important R functions

- joinocunt.multi()
- moran.test()
- moran.plot()
- localmoran()
- geary.test()
- globalG.test()
- localG()
- sp.correlogram()

References

Spatial Analysis by Tonny Oyana, 2nd edition, Chapter 7.

Applied Spatial Data Analysis with R by Roger S. Bivand, Edzer Pebesma, and Virgilio Gómez-Rubio, 2nd edition, (2013), Chapter 9.

https://cran.r-project.org/web/packages/spdep/spdep.pdf

https://geodacenter.github.io/workbook/5a_global_auto/lab5a.html#fn4