Spatial Analysis Part 1

Geog 210B Winter 2018

Kostas Goulias

Outline

- Introduction & Motivation
- Spatial Autocorrelation
- Spatial Weights
- Geostatistics (part 1)
- Spatial Data and Basic Visualization in R
- Moran I and Geary C
- Local Moran I

Motivation for spatial analysis

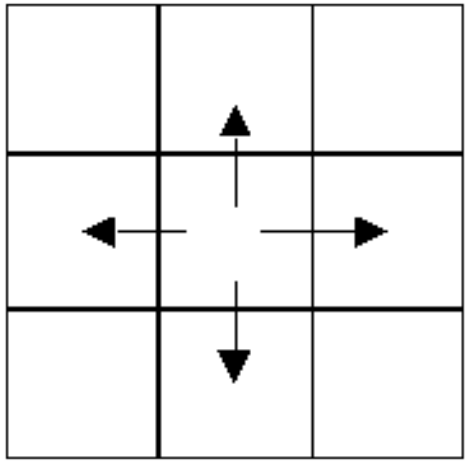
- What happens at one place influences what happens at another place
  - Neighborhoods
  - Relationships among units
  - Closeness (Tobler)
- The impact of a variable on an outcome is different among different places (heterogeneity)
- Something related to space is excluded from the model specification (unobserved latent trait) – model is misspecified

Spatial Correlation

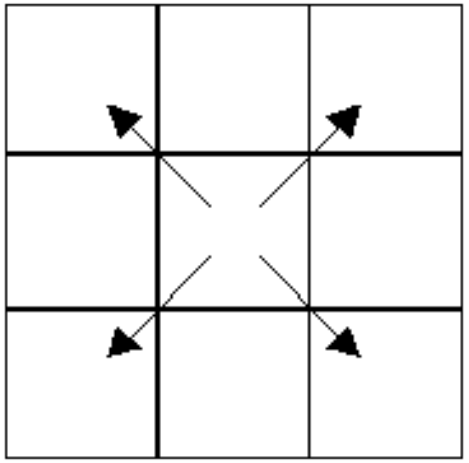
- Spatial autocorrelation measures the degree to which a phenomenon of interest is correlated to itself in space (Cliff and Ord 1973, 1981).
- Tests of spatial autocorrelation examine whether the observed value of a variable at one location is independent of values of that variable at neighboring locations.
- Positive spatial autocorrelation indicates that similar values appear close to each other, or cluster, in space
- Negative spatial autocorrelation indicates that neighboring values are dissimilar or, equivalently, that similar values are dispersed.
- Null spatial autocorrelation indicates that the spatial pattern is random.

Contiguity types

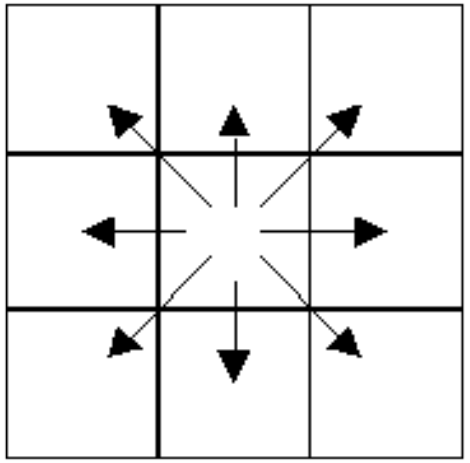
Rooks Case



Bishops Case

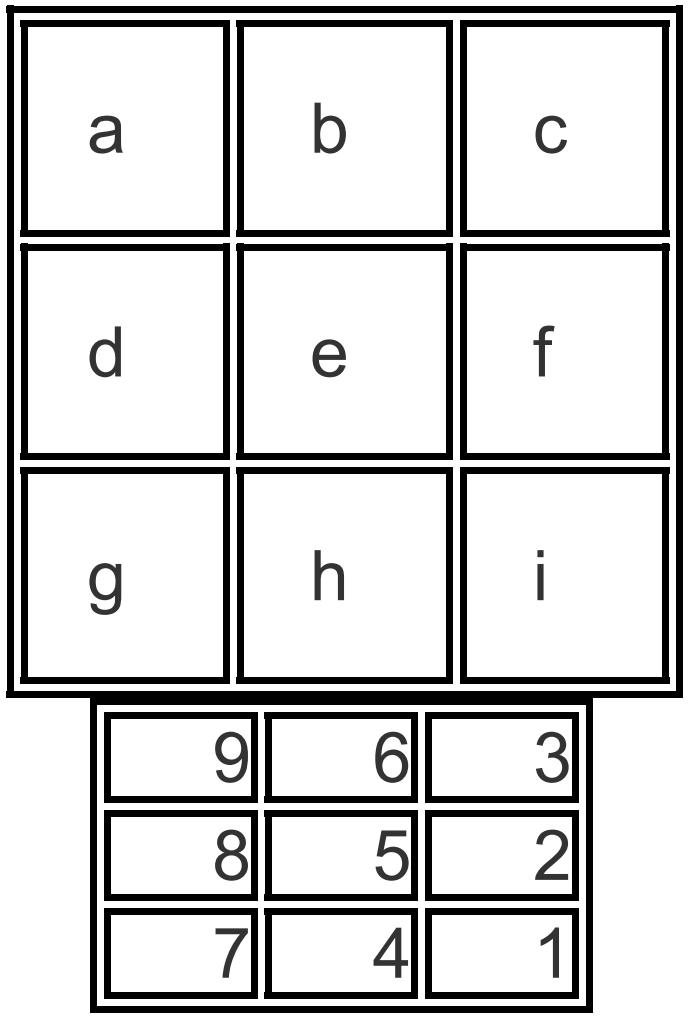


Queen's (Kings) Case



These are all first order (right next "door")
We could also have second order. For now we focus on the first order.

Sawada's example on GauchoSpace



Cell names

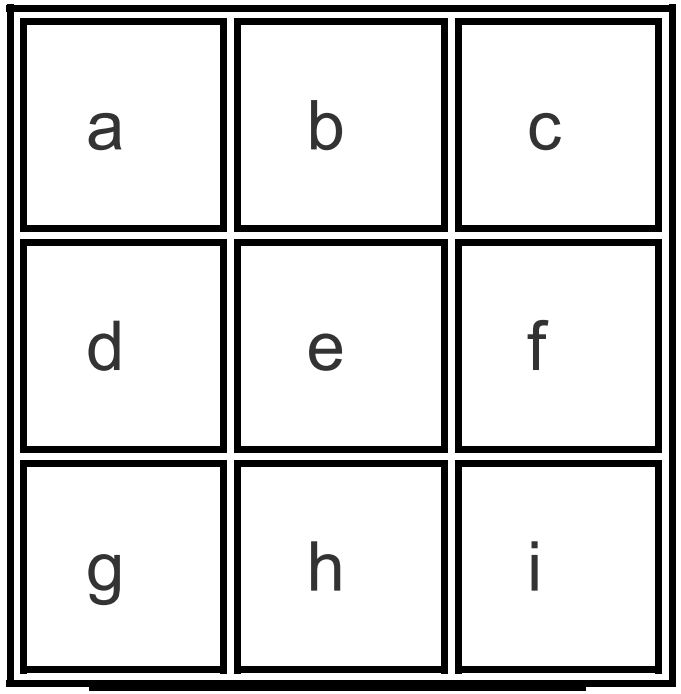
Cell values of the variable we analyze X_i

Sawada's example on GauchoSpace

Neighbors and spatial weights

This is the contiquity or connectivity or spatial weights matrix

	a	b	c	d	e	f	g	h	i
a	0	1	0	1	0	0	0	0	0
b	1	0	1	0	1	0	0	0	0
c	0	1	0	0	0	1	0	0	0
d	1	0	0	0	1	0	1	0	0
e	0	1	0	1	0	1	0	1	0
f	0	0	1	0	1	0	0	0	1
g	0	0	0	1	0	0	0	1	0
h	0	0	0	0	1	0	1	0	1
i	0	0	0	0	0	1	0	1	0



This is the Rook contiguity

Wij = 1 if I can reach the neighboring cell when the cells share a side; otherwise is zero

Compare cell e to cell f to cell a

	a	b	c	d	e	f	g	h	i
a	0	1	0	1	0	0	0	0	0
b	1	0	1	0	1	0	0	0	0
c	0	1	0	0	0	1	0	0	0
d	1	0	0	0	1	0	1	0	0
e	0	1	0	1	0	1	0	1	0
f	0	0	1	0	1	0	0	0	1
g	0	0	0	1	0	0	0	1	0
h	0	0	0	0	1	0	1	0	1
i	0	0	0	0	0	1	0	1	0

I want to study the squared differences among cells $(x_i-x_j)^2$

Call this the C_{ij}

	a	b	c	d	e	f	g	h	i
a	0	9	36	1	16	49	4	25	64
b	9	0	9	4	1	16	1	4	25
c	36	9	0	25	4	1	16	1	4
d	1	4	25	0	9	36	1	16	49
e	16	1	4	9	0	9	4	1	16
f	49	16	1	36	9	0	25	4	1
g	4	1	16	1	4	25	0	9	36
h	25	4	1	16	1	4	9	0	9
i	64	25	4	49	16	1	36	9	0

Original data

a	b	c
d	e	f
g	h	i

9	6	3
8	5	2
7	4	1

Hamardad Product of Matrices (element by element)

$$(A \circ B)_{i,j} = (A)_{i,j}(B)_{i,j}$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \circ \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} = \begin{pmatrix} a_{11}\,b_{11} & a_{12}\,b_{12} & a_{13}\,b_{13} \\ a_{21}\,b_{21} & a_{22}\,b_{22} & a_{23}\,b_{23} \\ a_{31}\,b_{31} & a_{32}\,b_{32} & a_{33}\,b_{33} \end{pmatrix}$$

Source: Wikipedia

Cross Products (C_{ij} * W_{ij})

C_{ij}

	a	b	c	d	e	f	g	h	i
a	0	9	36	1	16	49	4	25	64
b	9	0	9	4	1	16	1	4	25
c	36	9	0	25	4	1	16	1	4
d	1	4	25	0	9	36	1	16	49
e	16	1	4	9	0	9	4	1	16
f	49	16	1	36	9	0	25	4	1
g	4	1	16	1	4	25	0	9	36
h	25	4	1	16	1	4	9	0	9
i	64	25	4	49	16	1	36	9	0

W_{ij}

	a	b	c	d	e	f	g	h	i
a	0	1	0	1	0	0	0	0	0
b	1	0	1	0	1	0	0	0	0
c	0	1	0	0	0	1	0	0	0
d	1	0	0	0	1	0	1	0	0
e	0	1	0	1	0	1	0	1	0
f	0	0	1	0	1	0	0	0	1
g	0	0	0	1	0	0	0	1	0
h	0	0	0	0	1	0	1	0	1
i	0	0	0	0	0	1	0	1	0

Cross Products ($C_{ij} * W_{ij}$)

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*
a	0	9	0	1	0	0	0	0	0
b	9	0	9	0	1	0	0	0	0
c	0	9	0	0	0	1	0	0	0
d	1	0	0	0	9	0	1	0	0
e	0	1	0	9	0	9	0	1	0
f	0	0	1	0	9	0	0	0	1
g	0	0	0	1	0	0	0	9	0
h	0	0	0	0	1	0	9	0	9
i	0	0	0	0	0	1	0	9	0

Sum of Cross Products ($C_{ij} * W_{ij}$) = 120

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*
a	0	9	0	1	0	0	0	0	0
b	9	0	9	0	1	0	0	0	0
c	0	9	0	0	0	1	0	0	0
d	1	0	0	0	9	0	1	0	0
e	0	1	0	9	0	9	0	1	0
f	0	0	1	0	9	0	0	0	1
g	0	0	0	1	0	0	0	9	0
h	0	0	0	0	1	0	9	0	9
i	0	0	0	0	0	1	0	9	0

General Cross Product Statistic

$$\Gamma = \sum_{i=all\ cells} \sum_{j=all\ cells} W_{ij} C_{ij}$$

This is just one possible set of values

a	b	c
d	e	f
g	h	i

9	6	3
8	5	2
7	4	1

Other possibilities

1	2	3
4	5	6
7	8	9

1	2	3
4	9	6
7	8	5

6	2	3
4	5	1
7	8	9

This is just one possible set of values

a	b	c
d	e	f
g	h	i

9	6	3
8	5	2
7	4	1

If the values in the cells are completely random, each cell could take any value between 1 and 9. Completely means there is systematicity in the way specific values appear in a specific cell.

The possible permutations are 9! = 362880

Each of these permutations lead to a different Γ

This means we can consider the cross product to be one realization from a distribution of random permutations

Then we can repeatedly create these permutations and create a distribution of General cross products, plot them and compare our Γ= 120 to the many possible

If Γ=120 is sort of rare -> we have spatial correlation

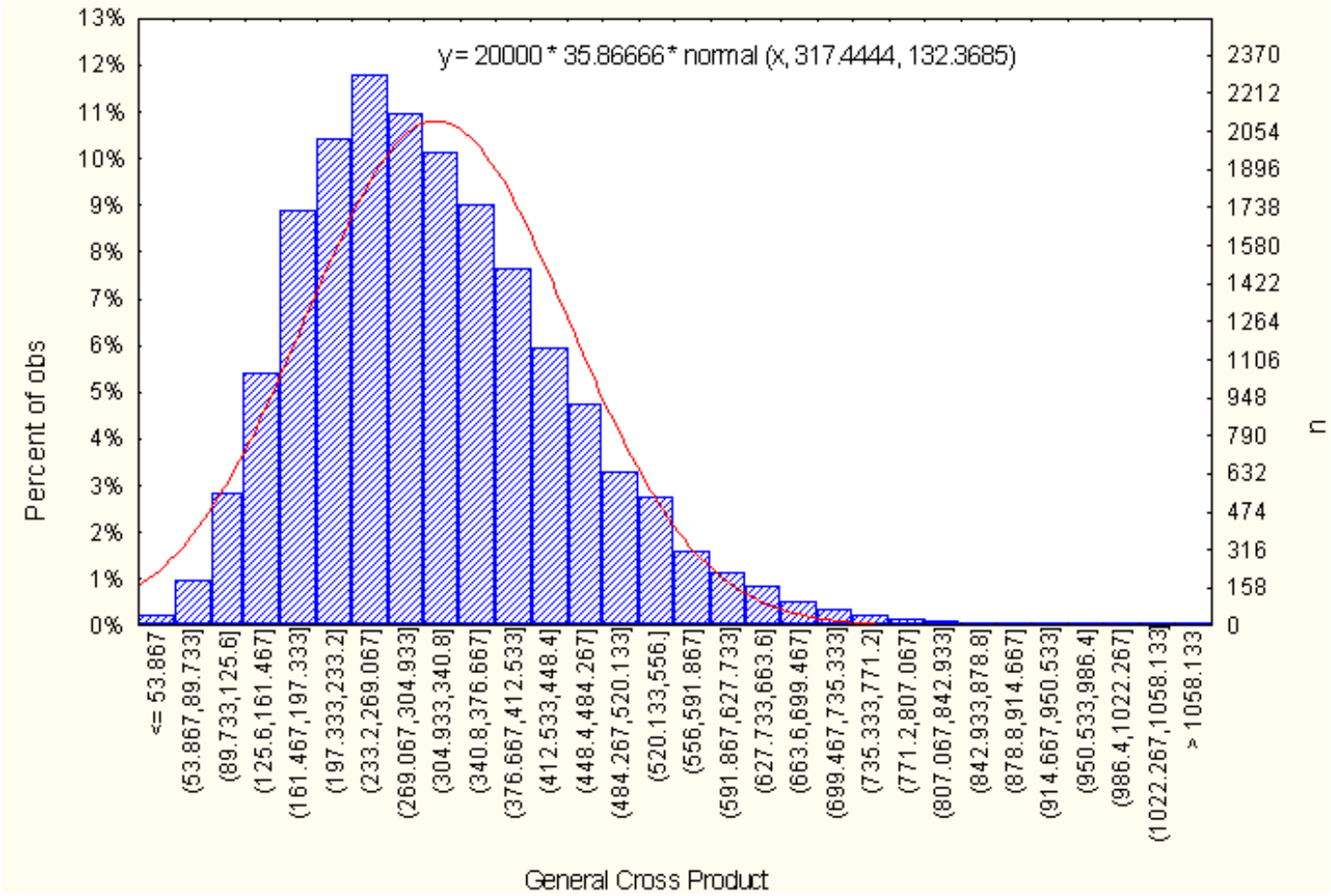
Create 20,000 possible 3 by 3 tables of values. Compute the spatial weights and then compute the cross products

Then check if the value 120 is extremely rare, rare, or just in the center of possibilities

Rare means something is “systematic” in the distribution of x among the cells. G=120 is somewhere in the tail of this distribution and therefore the values we observe are a systematic pattern.

Of course we knew this (large numbers on the left and small numbers on the right)

9	6	3
8	5	2
7	4	1



Moran's I

$$I = \frac{n}{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}} \bullet \frac{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{i=n}\left(x_i - \bar{x}\right)^2}$$

This time we use products of contiguous cell values that are deviations from the mean

When Moran I = +1 indicates strong positive correlation
When Moran I = 0 indicates random pattern of values in space
When Moran I = -1 indicates strong negative correlation

In Sawada's example

n = 9

$\bar{X} = 5$

$C_{ij} = (X_i - \bar{X})(X_j - \bar{X})$

W_{ij} same as above

I = 0.5

Positive spatial correlation among the 9 cell values

C_{ij}

	a	b	c	d	e	f	g	h	i
a	16	4	-8	12	0	-12	8	-4	-16
b	4	1	-2	3	0	-3	2	-1	-4
c	-8	-2	4	-6	0	6	-4	2	8
d	12	3	-6	9	0	-9	6	-3	-12
e	0	0	0	0	0	0	0	0	0
f	-12	-3	6	-9	0	9	-6	3	12
g	8	2	-4	6	0	-6	4	-2	-8
h	-4	-1	2	-3	0	3	-2	1	4
i	-16	-4	8	-12	0	12	-8	4	16

$$I = \frac{n}{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}} \bullet \frac{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}$$

$$I = \frac{9}{24} \bullet \frac{80}{60}$$

When is this zero? Only if the $C_{ij} = (X_i - \bar{X})(X_j - \bar{X})$ are zeros

When is this negative? When many $C_{ij} = (X_i - \bar{X})(X_j - \bar{X})$ are negative
(higher than average Xs right next lower than average Xs)

Geary's C

$$C = \frac{(n-1)\sum_{i=1}^{i=n}\sum_{j=1}^{i=n} w_{ij}(x_i - x_j)^2}{2\left(\sum_{i=1}^{i=n}\sum_{j=1}^{i=n} w_{ij}\right)\sum_{i=1}^{i=n}(x_i - x)^2}$$

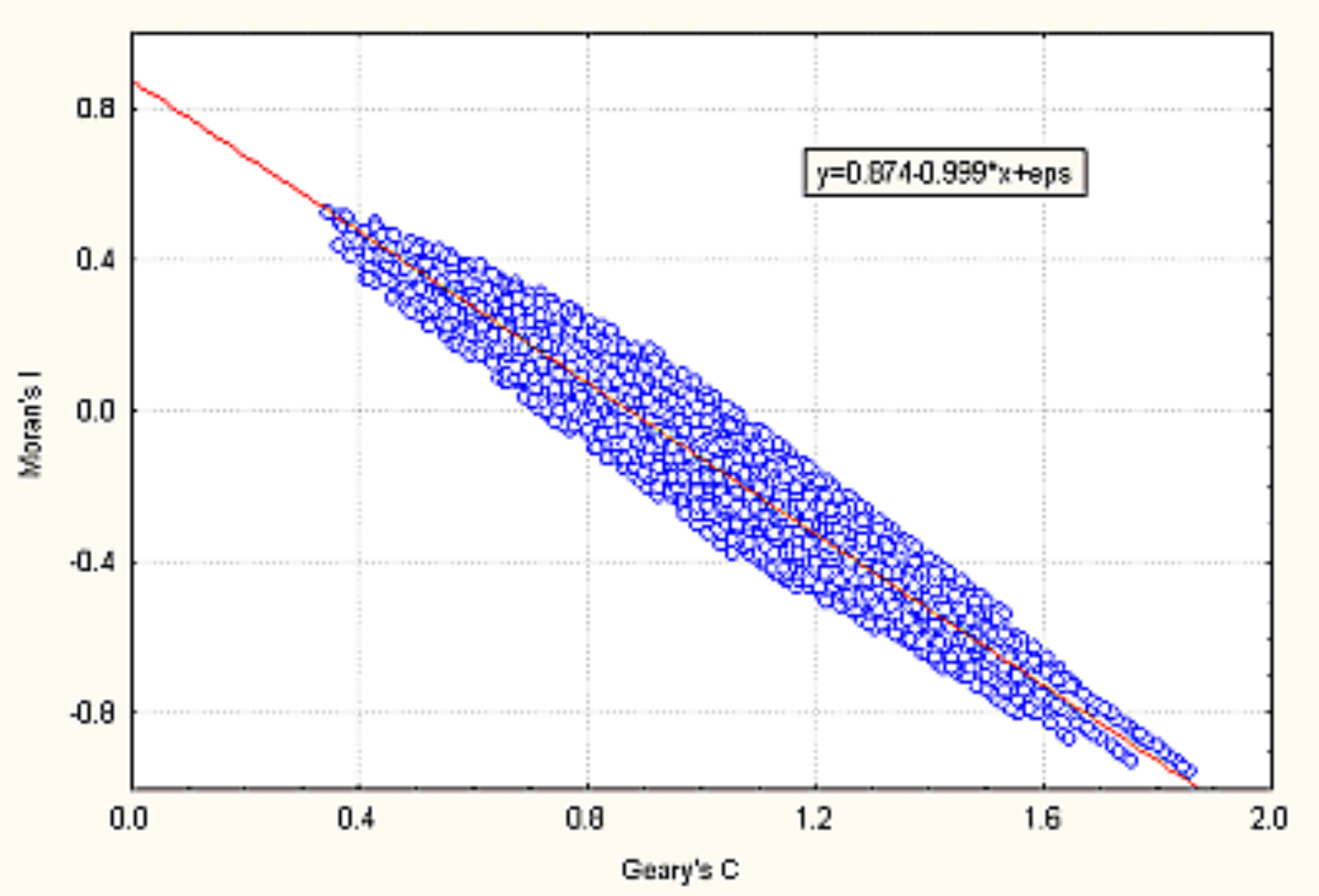
C is between 0 and 1 when we have positive spatial correlation

C = 0 for no correlation

C is between 1 and 2 when we have negative correlation

The Sawada writeup has an error – see above

Relationship between I and C (Sawada)



Global Spatial Correlation
As a function of lag d

Moran's I and Geary's c

Moran's I:

$$I(d) = \frac{\frac{1}{W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - \bar{y})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{for } h \neq i$$

Geary's c:

$$c(d) = \frac{\frac{1}{2W}\sum_{h=1}^{n}\sum_{i=1}^{n} w_{hi}(y_h - y_i)^2}{\frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \text{for } h \neq i$$

n is the number of units indexed by h and i.

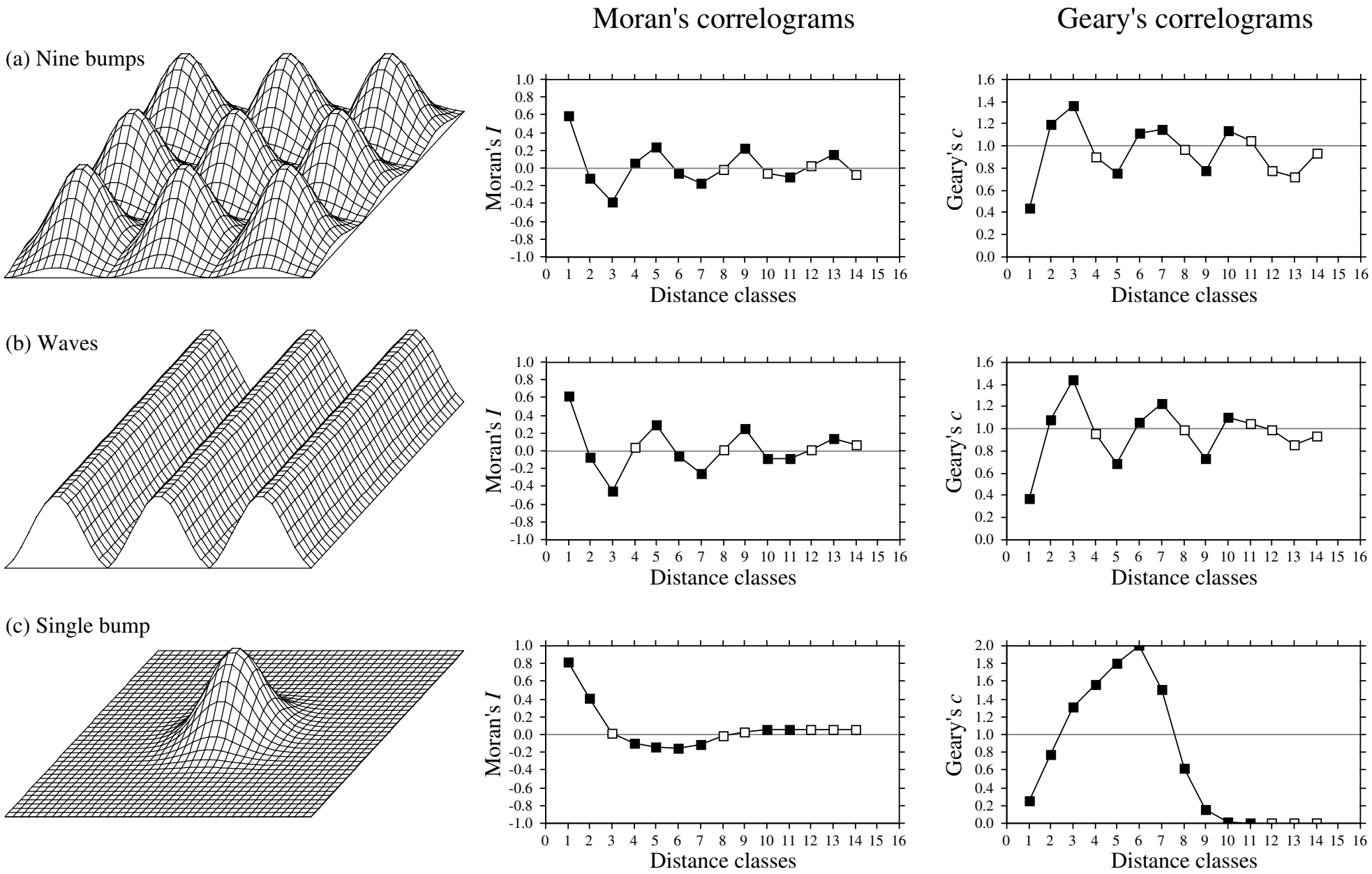
y is the variable we analyze

y bar is its mean

w_{hi} are the elements of the usual matrix (weights) taking values 0 and 1 to represent the neighboring h and i. w_{ii} is usually set to zero (I am a not a neighbor of myself)

W is the sum of all weights

d is the geographic distance between h and I (the lag)



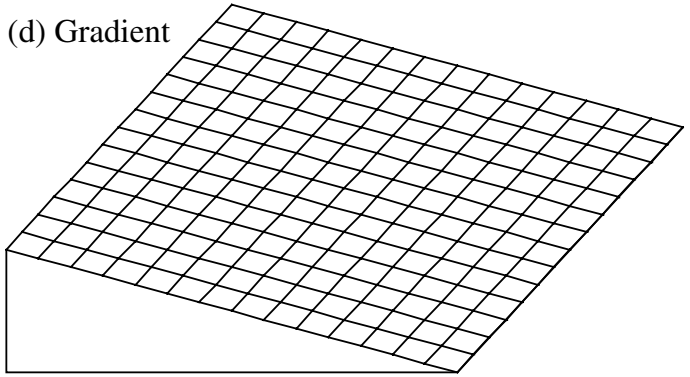
From Legendre & Legendre. Numerical Ecology, 1998, Elsevier

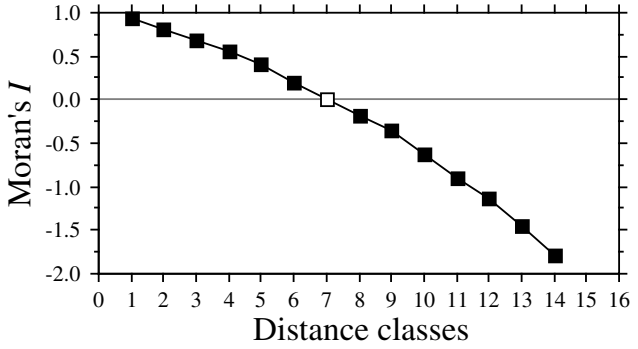
Mynote: page 723 (in my research paper directory under “spatial analysis”)

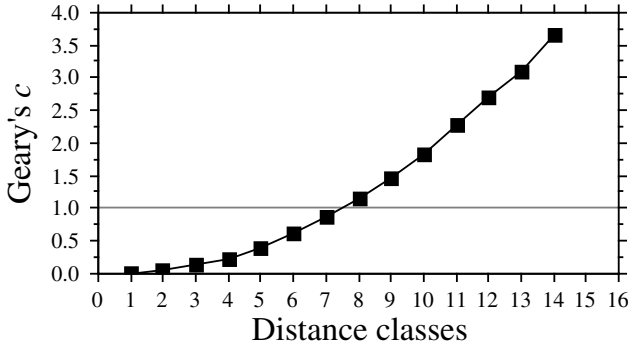
Moran's correlograms

Geary's correlograms

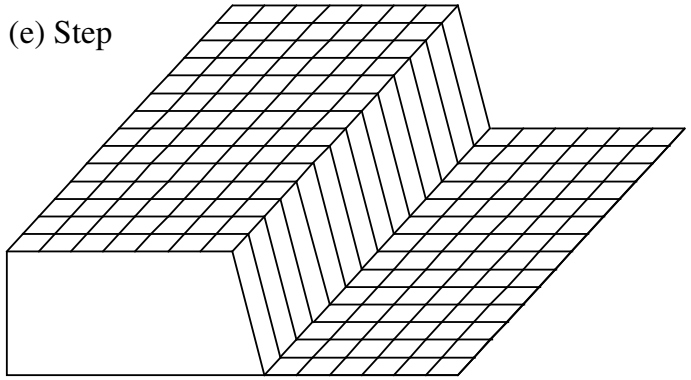
(d) Gradient

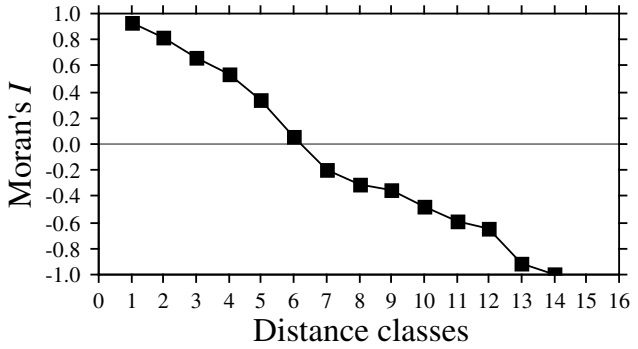


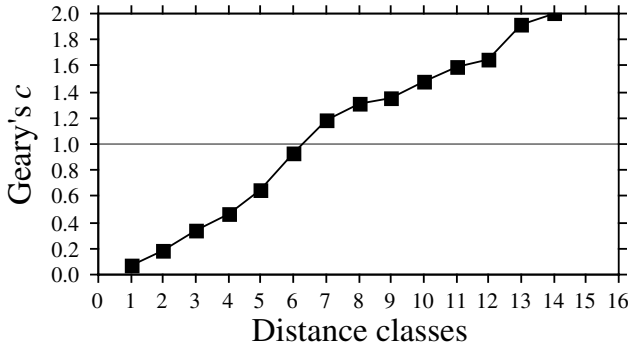




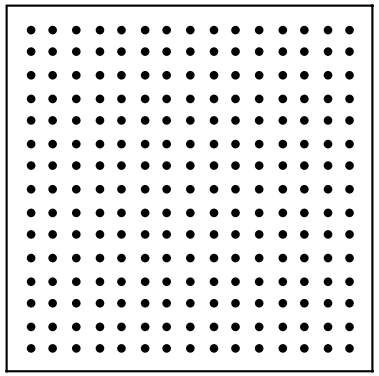
(e) Step

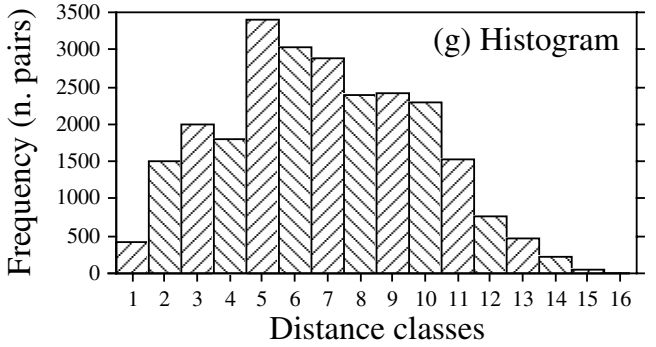


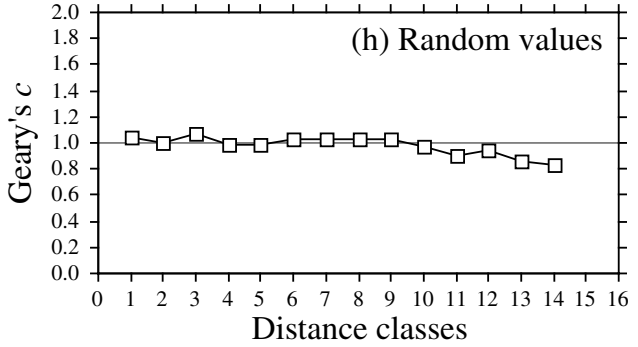




(f) Sampling grid (15×15)







Moran's I

$$I = \frac{n}{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}} \bullet \frac{\sum_{i=1}^{i=n}\sum_{j=1}^{j=n} W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{i=n}(x_i - \bar{x})^2}$$

- This is a statistic
- Different permutations of the Xs tables create many different values of I
- If we assume the Xs are from a Normal Distribution with no spatial correlation we can get E(I) and Var (I)
- If we assume the Xs are from a random permutation of the 3 by 3 table values with no spatial correlation we also get E(I) and Var(I)
- Then we can use z-scores to judge if "significant" spatial correlation is present in our data

$$Var_N(I) = \left(\frac{1}{S_0^2 (n2-1)} (n^2 S_1 - nS_2 + 3S_0^2) \right) - E_N(I)^2$$

whereas, the variance of Moran's I under Randomization - $Var_R(I)$ - is given as:

$$Var_R(I) = \frac{\left(\left\{ n\left[(n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2 \right] \right\} - \left\{ k\left[(n^2 - n)S_1 - 2nS_2 + 6S_0^2 \right] \right\} \right)}{(n-1)(n-2)(n-3)S_0^2} - ER(I)2$$

The following variables in the variance equations are defined as:

n = Number of Observations

$$E_N(I) = \frac{-1}{(n-1)}$$

$$E_R(I) = E_N(I)$$

$S_0 = \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} W_{ij}$ The sum of the Spatial Weight Matrix

$S_1 = \frac{\sum_{i=1}^{i=n} \sum_{j=i}^{j=n} (W_{ij} + W_{ji})^2}{2}$ If weight matrix symmetric then $S_1 = 2 \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} W_{ij}$

$S_2 = \sum_{i=1}^{i=n} (W_{i\bullet} + W_{\bullet i})^2$ The sum of the (i^{th}column + i^{th}Row)2 of weight matrix. If symmetrix $S_2 = 4 \sum_{i=1}^{i=n} W_{i\bullet}^2$

$k = \frac{\left[\sum_{i=1}^{i=n} (x_i - \bar{x})^4 / n \right]}{\left[\sum_{i=1}^{i=n} (x_i - \bar{x})^2 / n \right]^2}$ Involves the sum of each value in the data matrix minus the mean

The Standard deviation and standard z-scores of I are given by:

$$SD_{NorR}(I) = \sqrt{Var_{NorR}(I)}$$

and the z - score is given as

$$z = \frac{(I - E_{NorR}(I))}{\sqrt{Var_{NorR}(I)}}$$

The expectation of no spatial correlation on the 9 cell values is – 1/(9-1) = -1/8 = -0.125

The variance depends on the assumption about the underlying no spatial correlation random process

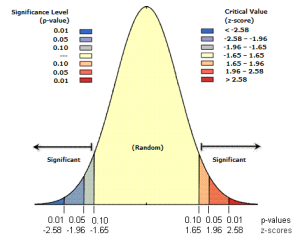
Normal and Random are the two labels used by most references (Sagawa, Goodchild, Cliff & Ord, Getis & Ord)

Comparison

	E(I)	Sqrt(Var(I)	Z-score
Normal Assumption	-0.125	0.23049	2.7116
Random Permutations	-0.125	0.24431	2.5582
20,000 Random Samples	-0.128	0.24200	2.5950
Observed	0.5		

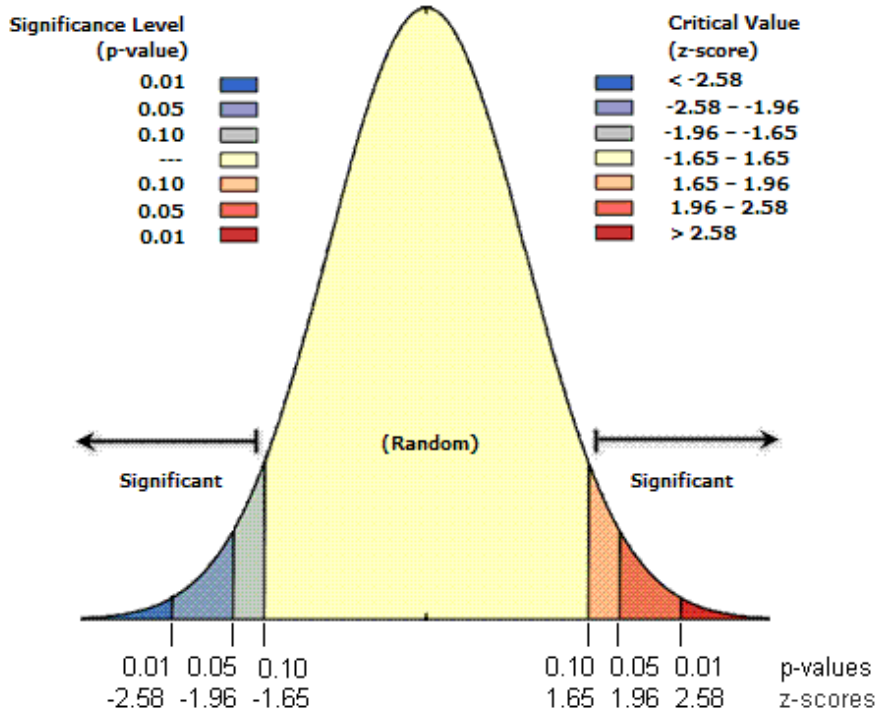
With confidence more than 95% we can say there is significant positive correlation among the cell values in the 9 cells we analyzed.

Z-scores are standard deviations. If, for example, a tool returns a z-score of +2.5, you would say that the result is 2.5 standard deviations. Both z-scores and p-values are associated with the standard normal distribution as shown below.



Nice explanation in http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/what-is-a-z-score-what-is-a-p-value.htm

Z-scores are standard deviations. If, for example, a tool returns a z-score of +2.5, you would say that the result is 2.5 standard deviations. Both z-scores and p-values are associated with the standard normal distribution as shown below.



The mechanics of diagnosing spatial autocorrelation

- Choose an indicator that depicts spatial autocorrelation (Moran, Geary, or whatever else people developed)
- Choose a matrix that represents relationships among spatial units (W_{ij}) . This can be Rook, Queen, some kind of distance.
- Find theoretical support for the Expectation and Variance of the indicator under the null hypothesis of no spatial correlation
- Compute a z-score and its p-value and then judge if your data display spatial correlation

Task 1: Space Support (space representation)

- **Points** - pairs of coordinates (x,y), representing events, observation posts, individuals, cities or any other discrete object in space.

- **Polygons & Lines** - sequences of connected points: polygon= the first point is the same as the last; line = open polygon with the sequence of points does not result in a closed shape (R calles these SpatialPolygons and SpatialLines).

- **Raster Grid** - divides the study region into a set of identical, regularly-spaced, discrete elements (pixels), each of which records the value or presence/absence of a quantity of interest (used in image processing, remote sensing data, weather forecasts, disease maps, elevation models).

Example in California

- Key packages/libraries are:
  - spdep
  - maptools
  - leaflet

  - We set the working directory:\

  - setwd("~/Documents/COURSES UCSB/Course Winter 2018/California")
  - ## We read in R a shapefile in an object called CA.poly . #

  - CA.poly <- readShapePoly('LPA_Pop_Char_bg.shp')

SpatialPolygonsDataFrame object in R

- A **SpatialPolygonsDataFrame** object brings together the spatial representations of the polygons with data.
- A Spatial Polygon Data Frame has four components (in the R jargon these are called **slots**)
- **Component 1**: Data contains the variables that are used in the analysis in each unit of analysis that is a spatial object (in this example a Census block group) with variables like number of households with zero cars that live in each block group, # vehicle miles of travel produced in each block group.

Component 1: Data contains the variables that are used in the analysis such as number of households with zero cars,# vehicle miles of travel. 23198 obs. (the US Census blockgroups) each containing values of 105 variables (McBride's MA thesis)

```
> str(slot(CA.poly, "data"))
'data.frame':   23198 obs. of  105 variables:
 $ OBJECTID  : int  1 2 3 4 5 6 7 8 9 10 ...
 $ STATEFP   : Factor w/ 1 level "06": 1 1 1 1 1 1 1 1 1 1 ...
 $ COUNTYFP  : Factor w/ 58 levels "001","003","005",..: 42 37 30 12 19 30 30 19 19 33
...
 $ TRACTCE   : Factor w/ 6521 levels "000100","000101",..: 319 1482 2412 1083 5937 249
2 1898 3170 3757 2302 ...
 $ BLKGRPCE  : Factor w/ 10 levels "0","1","2","3",..: 2 3 3 2 2 3 6 2 2 2 ...
 $ AFFGEOID  : Factor w/ 23198 levels "1500000US060014001001",..: 19731 17394 11471 27
65 8865 11726 11313 4165 5247 13779 ...
 $ NAME      : Factor w/ 10 levels "0","1","2","3",..: 2 3 3 2 2 3 6 2 2 2 ...
 $ LSAD      : Factor w/ 1 level "BG": 1 1 1 1 1 1 1 1 1 1 ...
 $ ALAND     : num  6.70e+06 3.80e+06 2.89e+05 8.03e+08 2.61e+05 ...
 $ AWATER    : num  502651 2039168 0 1577551 2184908 ...
 $ GEO_ID2   : num  6.08e+10 6.07e+10 6.06e+10 6.02e+10 6.04e+10 ...
 $ GEOID_1   : num  6.08e+10 6.07e+10 6.06e+10 6.02e+10 6.04e+10 ...
 $ HHAGE1    : num  3 14 0 21 50 5 7 4 6 11 ...
 $ HHAGE2    : num  12 87 0 81 206 82 105 20 31 45 ...
 $ HHAGE3    : num  34 180 0 67 117 73 193 63 51 57 ...
 $ HHAGE4    : num  103 225 6 113 80 96 75 75 49 82 ...
 $ HHAGE5    : num  67 96 20 68 27 59 22 52 30 50 ...
 $ HHAGE6    : num  57 97 18 79 22 43 27 40 11 85 ...
 $ HHAGE7    : num  99 119 107 59 16 45 15 39 18 256 ...
 $ HHAGE8    : num  60 67 221 48 7 25 8 16 7 187 ...
```

Component 2: This is the polygon slot and contains the “shape” information.

Plot(CA.poly)

Component 3: this is the bobox (bounding box of coordinates that is drawn around the boundaries of CA)
Component 4: Is the proj4string that contains the projections.

- The $ is used to access a variable
- The @ is used to access a specific slot of the spatial data frame

The following is an example:

- summary (CA.poly@data$VMT)
- This produces:

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
       0   21738   31288   37816   45815 1148168
```

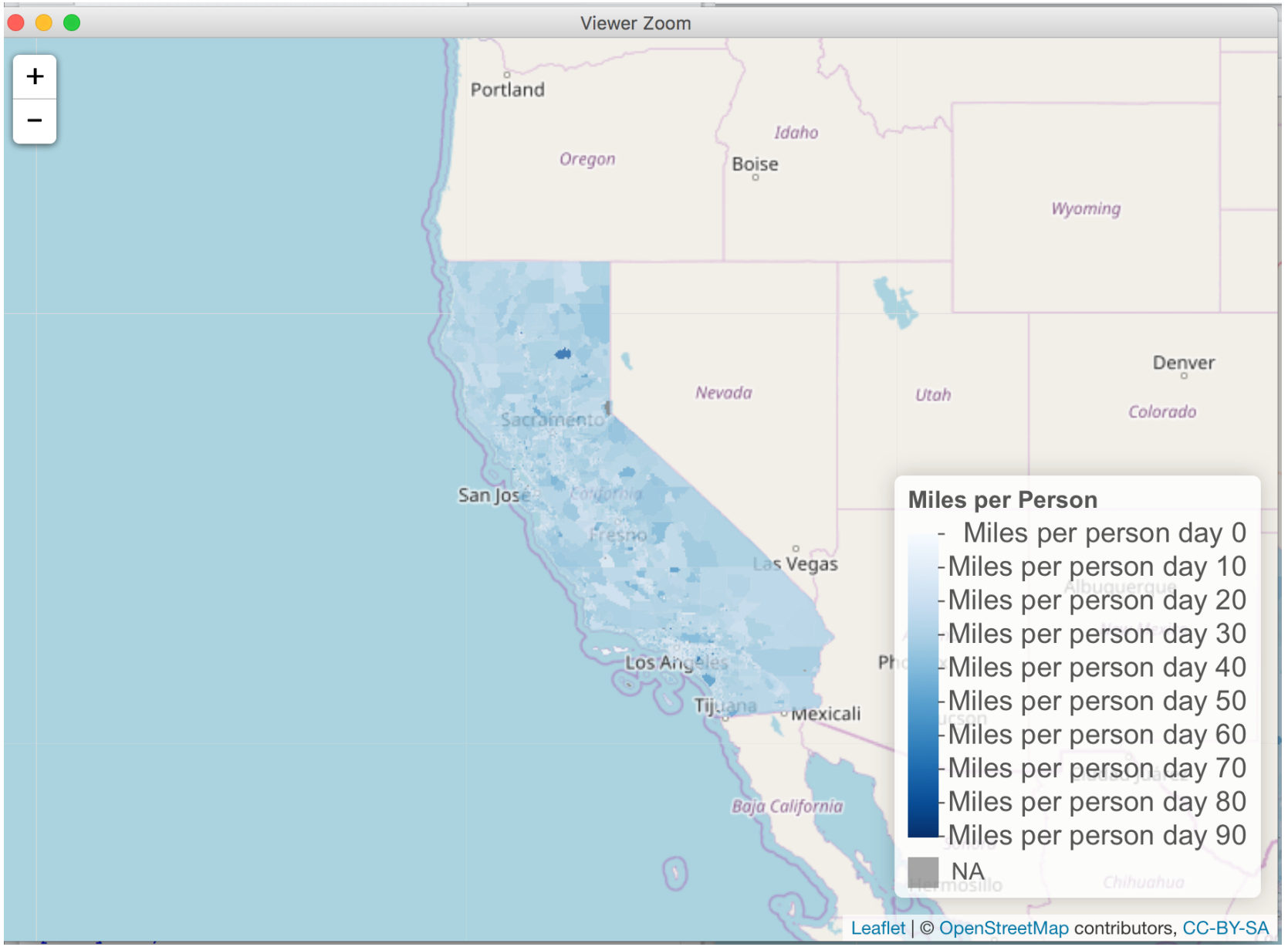
- *We can use all kinds of operations like in a usual dataframe*

- CA.poly@data$VMTpr = CA.poly@data$VMT/CA.poly@data$n_pr
- *The above creates a new variable*
- summary (CA.poly@data$VMTpr)
- *This produces the output*

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
0.00   18.49   22.46   24.06   28.37   90.00      63
```

Using package leaflet we can display this

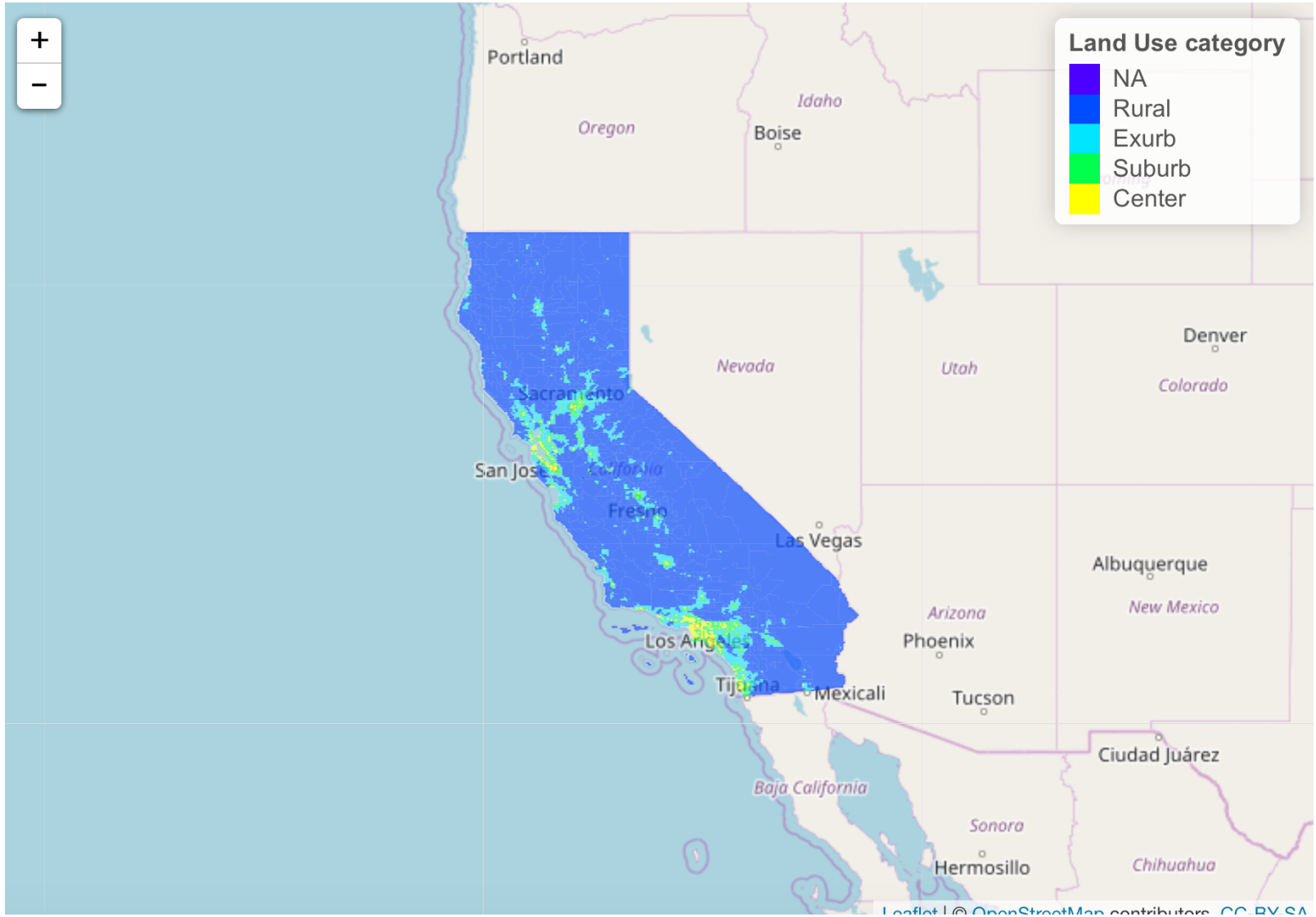
This is for a continuous variable



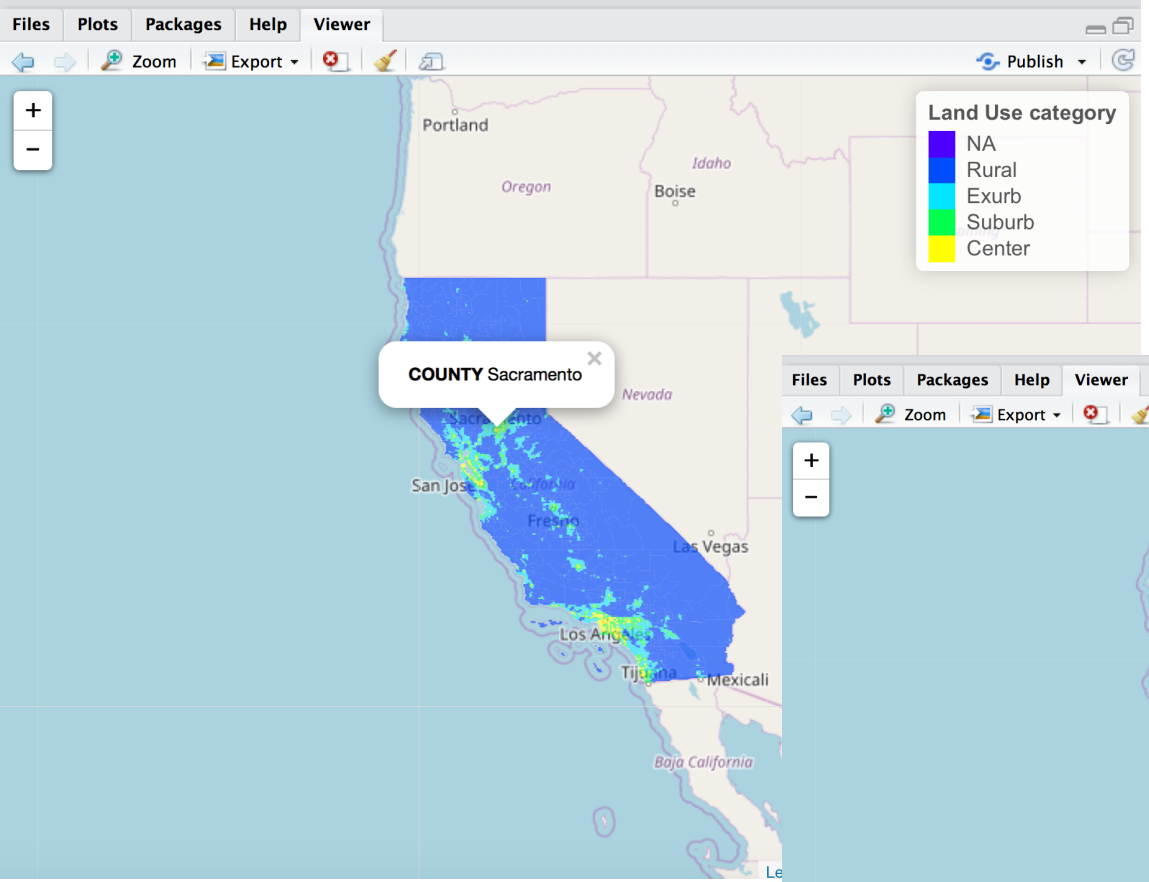
Defining a factor (categorical variable) we can create this map

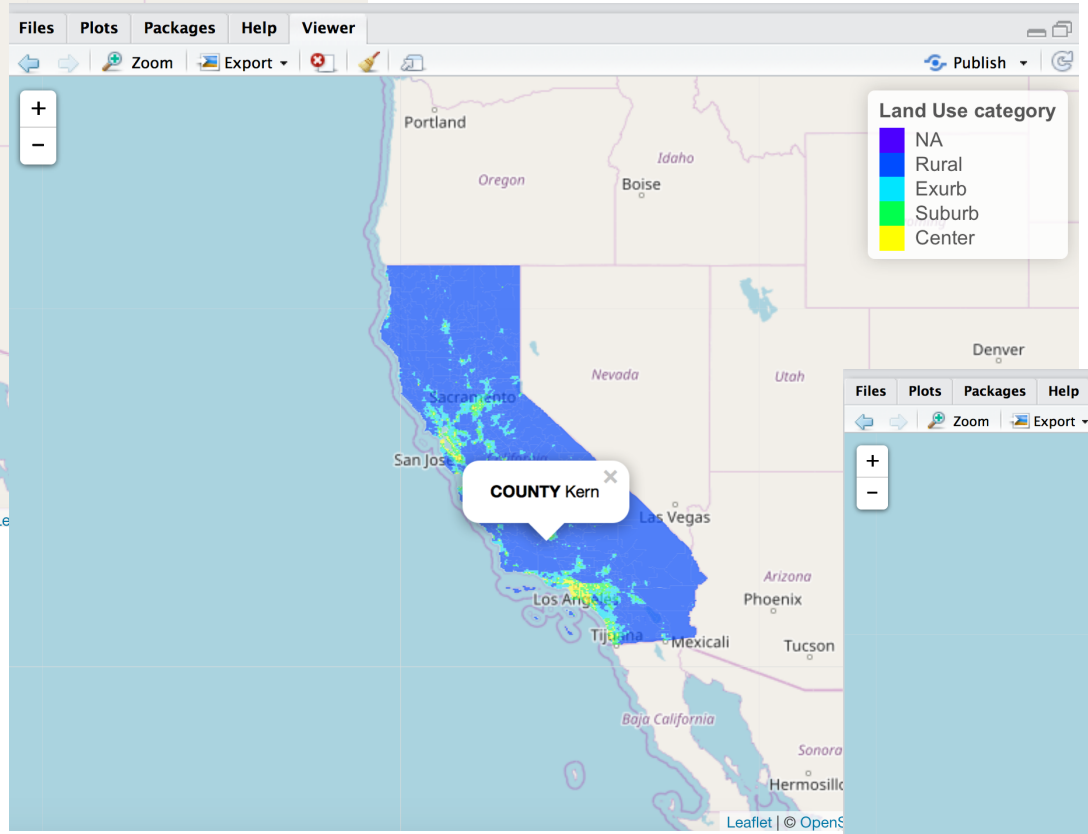
This classification was done by Elizabeth McBride in her MA thesis

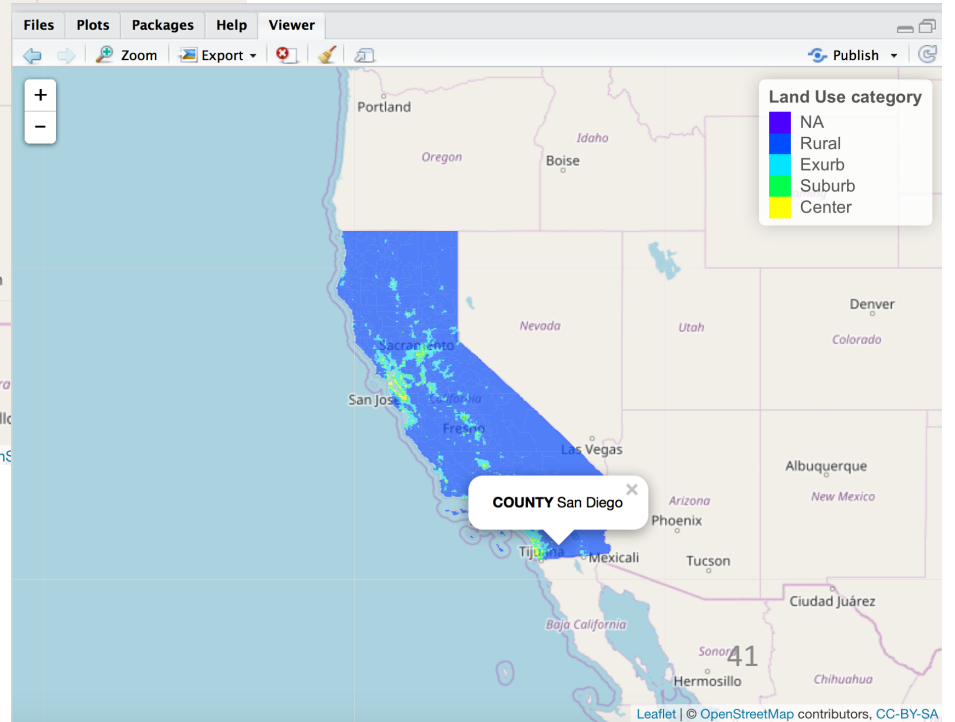
In essence uses data on business establishments to classify each block group in the four categories here



...and fancy popups







Linear Regression of VMT per person:
VMTprOLS<-lm(VMTpr~suburb+exurb+rural+HHVEH0+HHVEH1+HHVEH2+HHVEH3+HHVEH4+HHVEH5+HHVEH6+HHAGE7, data=CA.poly@data)

```
> summary(VMTprOLS)

Call:
lm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 + HHVEH1 +
    HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7, data = CA.poly@data)

Residuals:
    Min      1Q  Median      3Q     Max
-30.705  -3.446  -0.754   2.503  63.486

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.4747095  0.1044064 186.528  < 2e-16 ***
suburbTRUE   5.5996042  0.1005532  55.688  < 2e-16 ***
exurbTRUE    5.3916213  0.1555270  34.667  < 2e-16 ***
ruralTRUE   11.6717350  0.2011197  58.034  < 2e-16 ***
HHVEH0      -0.0199781  0.0017925 -11.145  < 2e-16 ***
HHVEH1      -0.0092536  0.0007676 -12.055  < 2e-16 ***
HHVEH2       0.0108465  0.0008535  12.708  < 2e-16 ***
HHVEH3       0.0109570  0.0019188   5.710 1.14e-08 ***
HHVEH4       0.0466177  0.0040481  11.516  < 2e-16 ***
HHVEH5      -0.1474161  0.0087433 -16.861  < 2e-16 ***
HHVEH6      -0.0645967  0.0099456  -6.495 8.47e-11 ***
HHAGE7       0.0025441  0.0011925   2.133   0.0329 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.687 on 23123 degrees of freedom
  (63 observations deleted due to missingness)
Multiple R-squared:  0.4257,    Adjusted R-squared:  0.4254
F-statistic:  1558 on 11 and 23123 DF,  p-value: < 2.2e-16
```

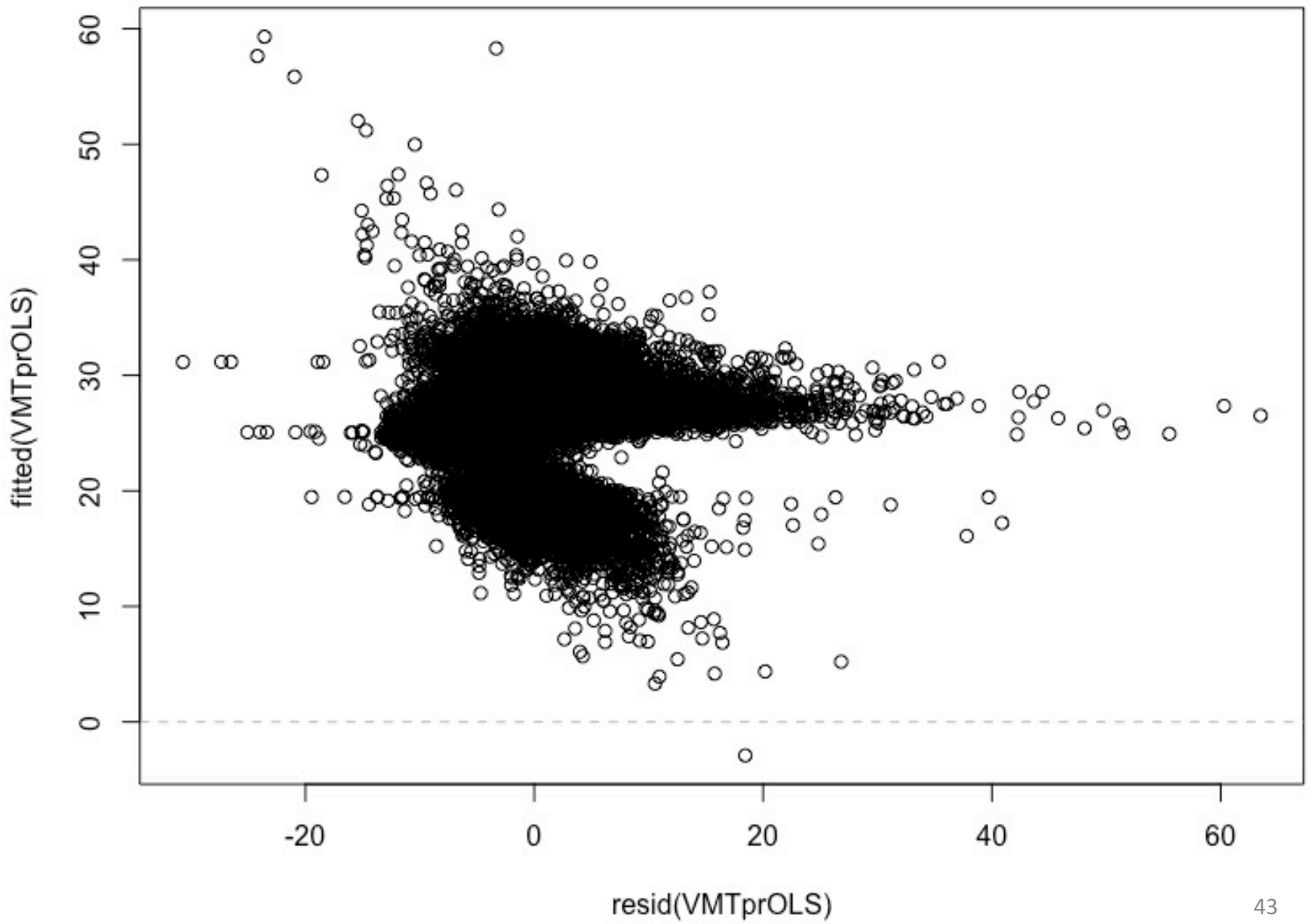
Compare center, suburbs, exurbs, and rural areas

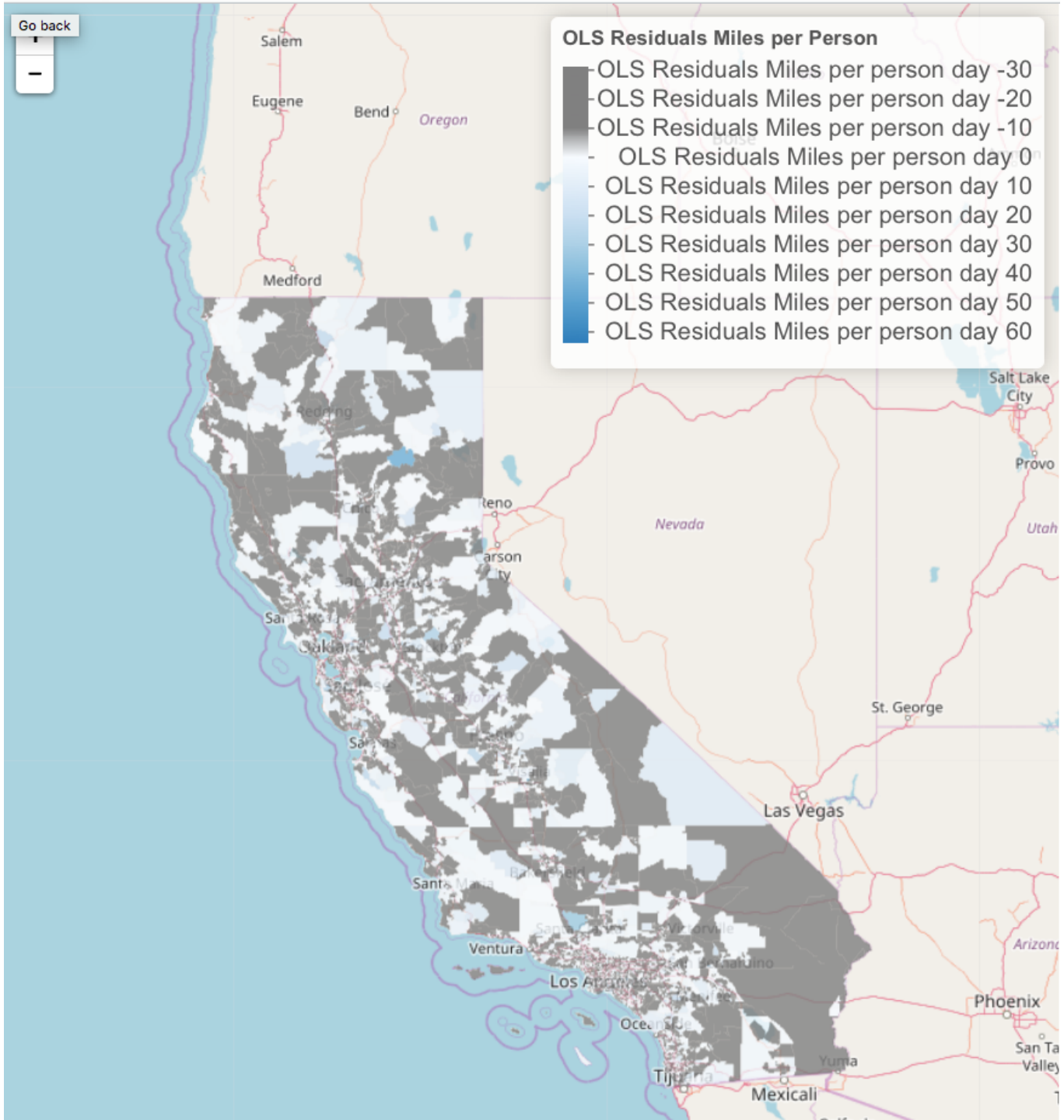
In Assignment 1 and 2 our observation units are households

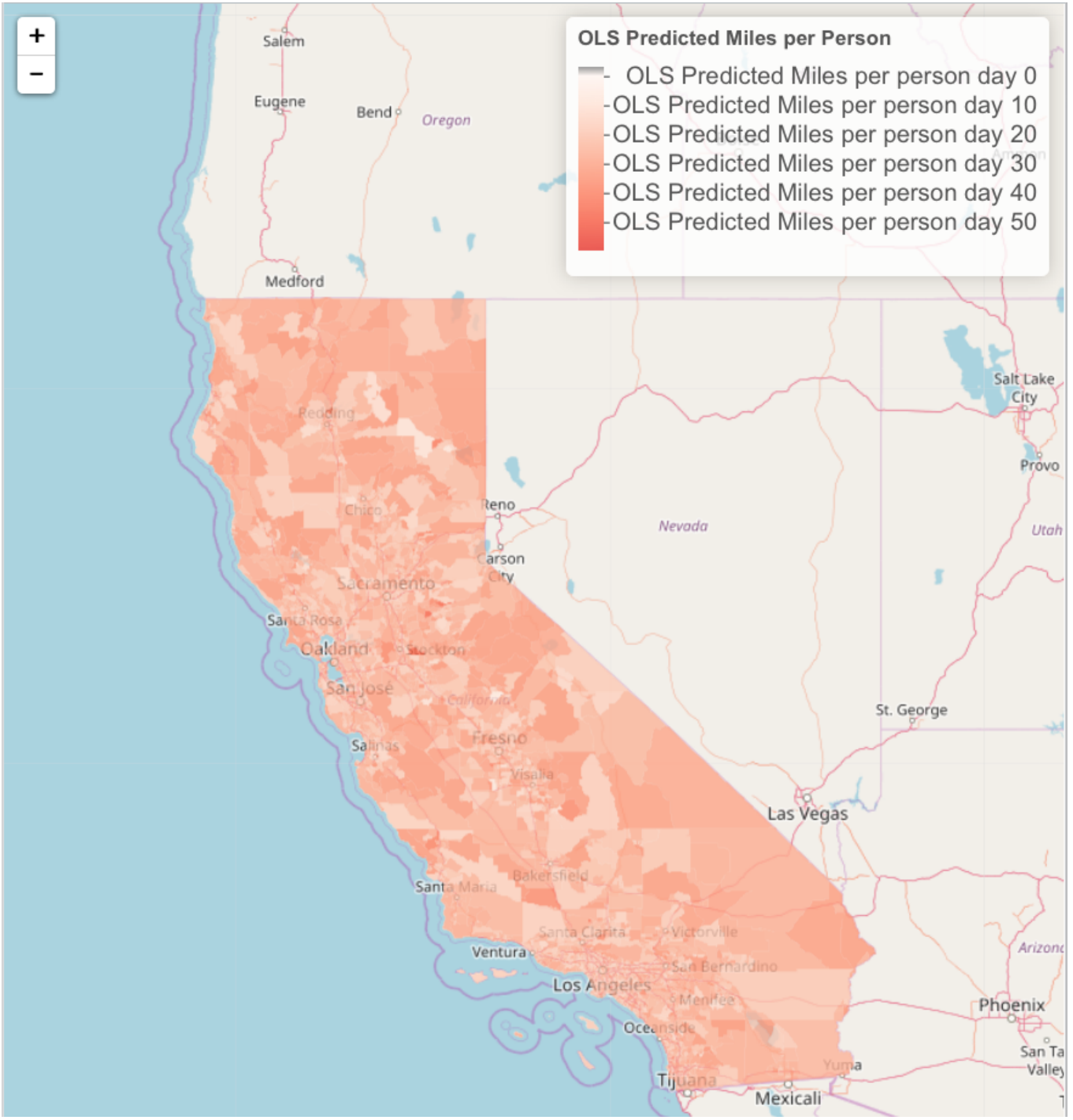
The observation units now are the Census blockgroup

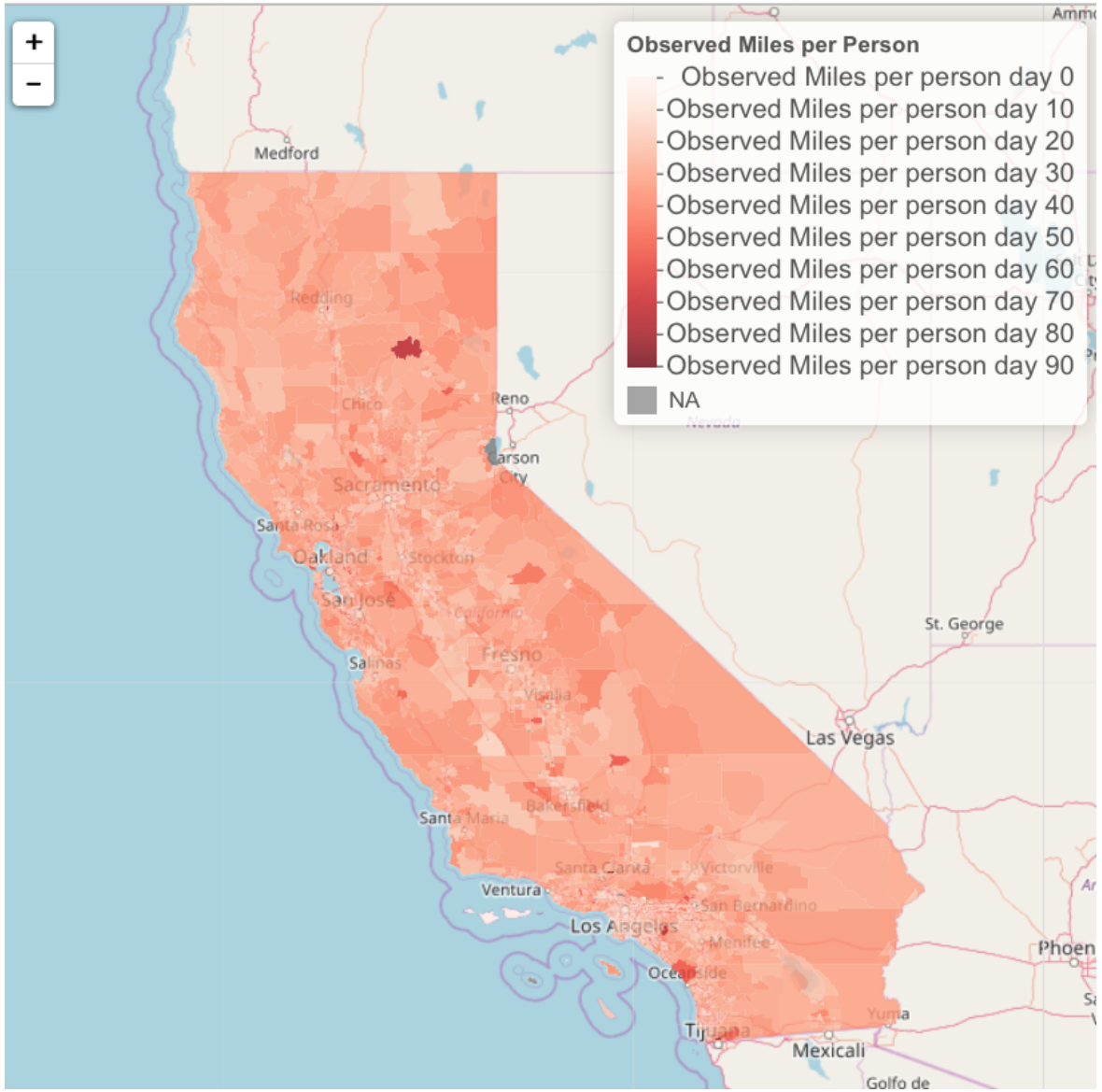
R-square? How does it compare with assignment 1?

Residual vs Fitted values




Go back
OLS Residuals Miles per Person
OLS Residuals Miles per person day -30
OLS Residuals Miles per person day -20
OLS Residuals Miles per person day -10
OLS Residuals Miles per person day 0
OLS Residuals Miles per person day 10
OLS Residuals Miles per person day 20
OLS Residuals Miles per person day 30
OLS Residuals Miles per person day 40
OLS Residuals Miles per person day 50
OLS Residuals Miles per person day 60
Salem
Eugene
Bend
Oregon
Medford
Salt Lake City
Provo
Nevada
Utah
St. George
Las Vegas
Ventura
Phoenix
Yuma
Mexicali



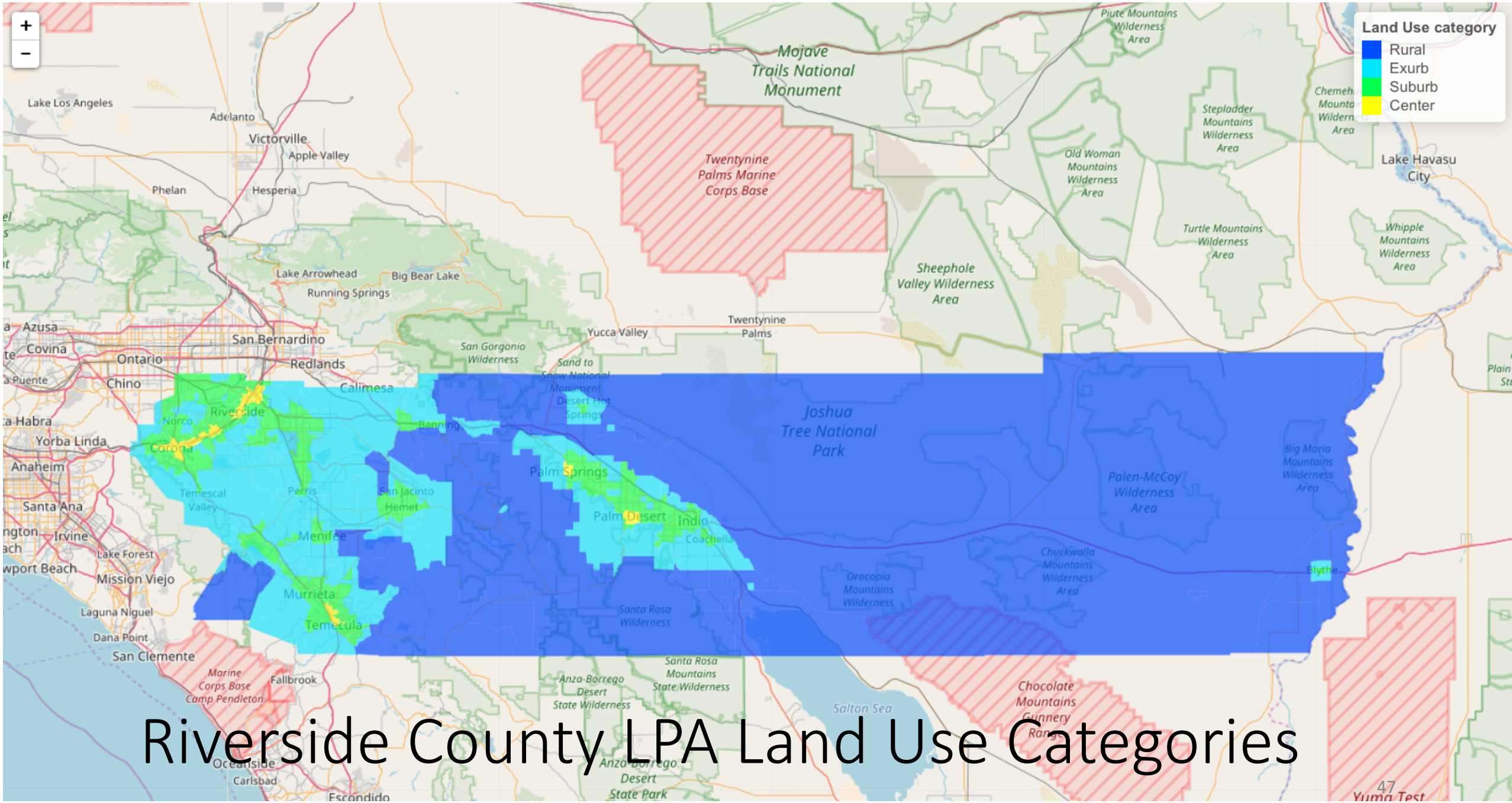


Observed and Predicted have major differences – look at span of values and spatial distribution of values. We have a systematic way to explore spatial structure (see later analysis using Riverside)

Select one county for analysis

- YCOUNTY <- CA.poly[CA.poly@data$countyname== c("Riverside"),]

- This selects all the polygons that belong to the county with name Riverside



Riverside County LPA Land Use Categories

```
Call:
lm(formula = VMTpr ~ suburb + exurb + rural + HHVEH0 + HHVEH1 +
    HHVEH2 + HHVEH3 + HHVEH4 + HHVEH5 + HHVEH6 + HHAGE7, data = YCOUNTY@data)

Residuals:
    Min      1Q  Median      3Q     Max
-17.319  -3.779  -0.688   3.021  49.862

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 21.687987   0.734600  29.524  < 2e-16 ***
suburbTRUE   5.004393   0.688883   7.265 7.44e-13 ***
exurbTRUE    2.831428   0.862265   3.284  0.00106 **
ruralTRUE    8.050345   1.153020   6.982 5.25e-12 ***
HHVEH0      -0.039912   0.009968  -4.004 6.68e-05 ***
HHVEH1      -0.018997   0.003620  -5.248 1.87e-07 ***
HHVEH2       0.001059   0.003655   0.290  0.77216
HHVEH3       0.020531   0.007350   2.793  0.00532 **
HHVEH4       0.112590   0.015355   7.332 4.61e-13 ***
HHVEH5      -0.313633   0.038005  -8.253 4.77e-16 ***
HHVEH6       0.107290   0.056601   1.896  0.05830 .
HHAGE7       0.017379   0.003758   4.625 4.23e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.685 on 1017 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.4126,    Adjusted R-squared:  0.4062
F-statistic: 64.93 on 11 and 1017 DF,  p-value: < 2.2e-16
```

What is the story?

Exurbs vs Center?

Not the same as the entire State of California

What is happening? (especially in terms of spatial dependency – remember these are Census blockgroups)

- Maybe Riverside is different than the rest of California
- Maybe our original definition of center, suburb, exurb, rural has a problem (measurement error = big problem because the x is stochastic and most likely correlated with random error terms)
- Maybe the closeness of exurbs to center play a “spatial” role (exclusion of this influence = misspecification)
- Maybe we should look for other influences (exclusion of important variables = misspecification)

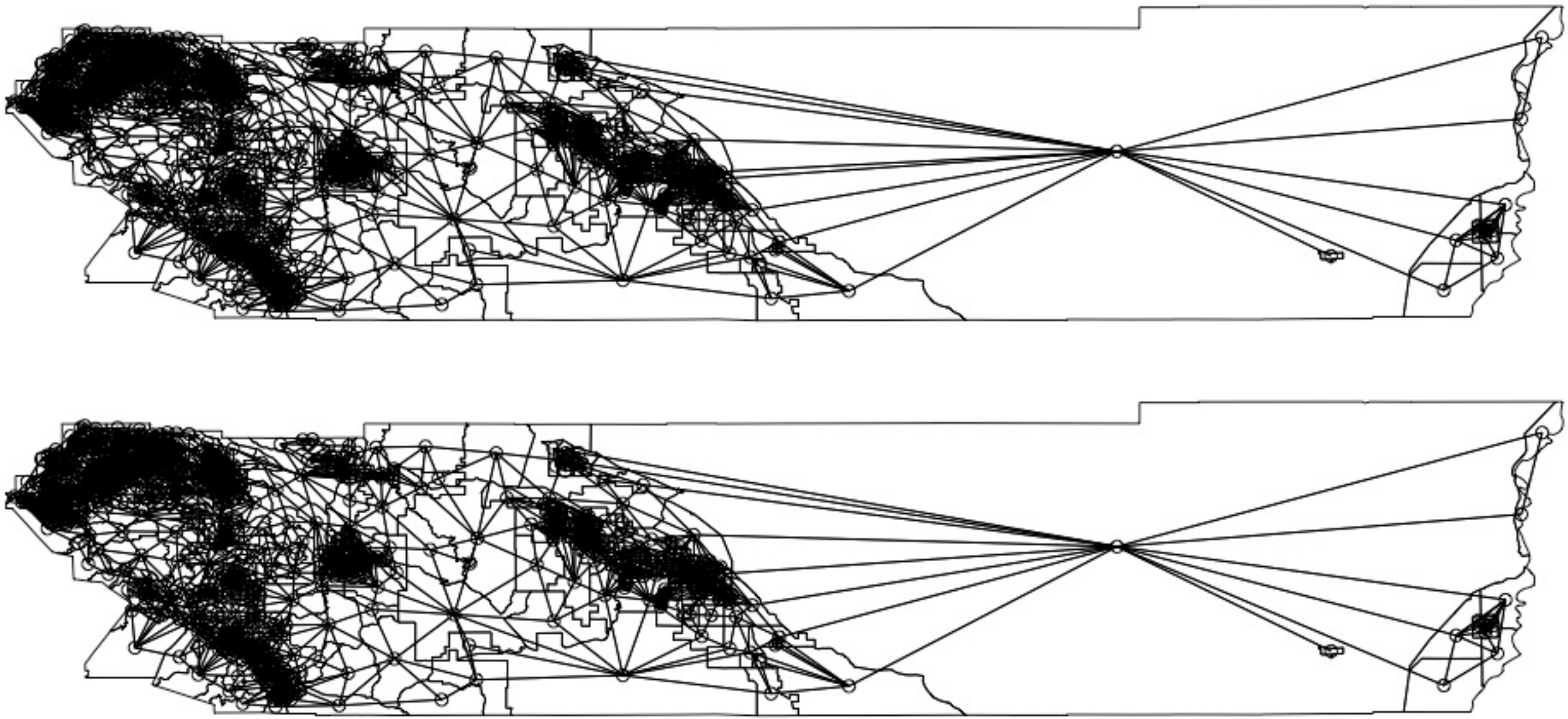
- Most likely all of the above

Neighborhoods

- To test for spatial correlation we first need to define neighborhoods for each unit of analysis (in our case the blockgroup polygons in Riverside county are the units)
- There are many ways to define neighborhoods
  - Contiguity
  - Distance
  - Other special relationships

Riverside County Polygons (US Census Block Groups)

Note the “automated” creation of centroids for each polygon!



Very small difference between queen and rook contiguity in this case

Nearest neighbor (k=10 vs k=1)

Spatial Weights

This in reality is a connectivity matrix that can be modified to reflect how we want neighbors to influence each other when we compute statistics

Weight Matrices from Queen and Rook for item I in space

			1	1	1		
			1	i	1		
			1	1	1		

				1			
			1	i	1		
				1			

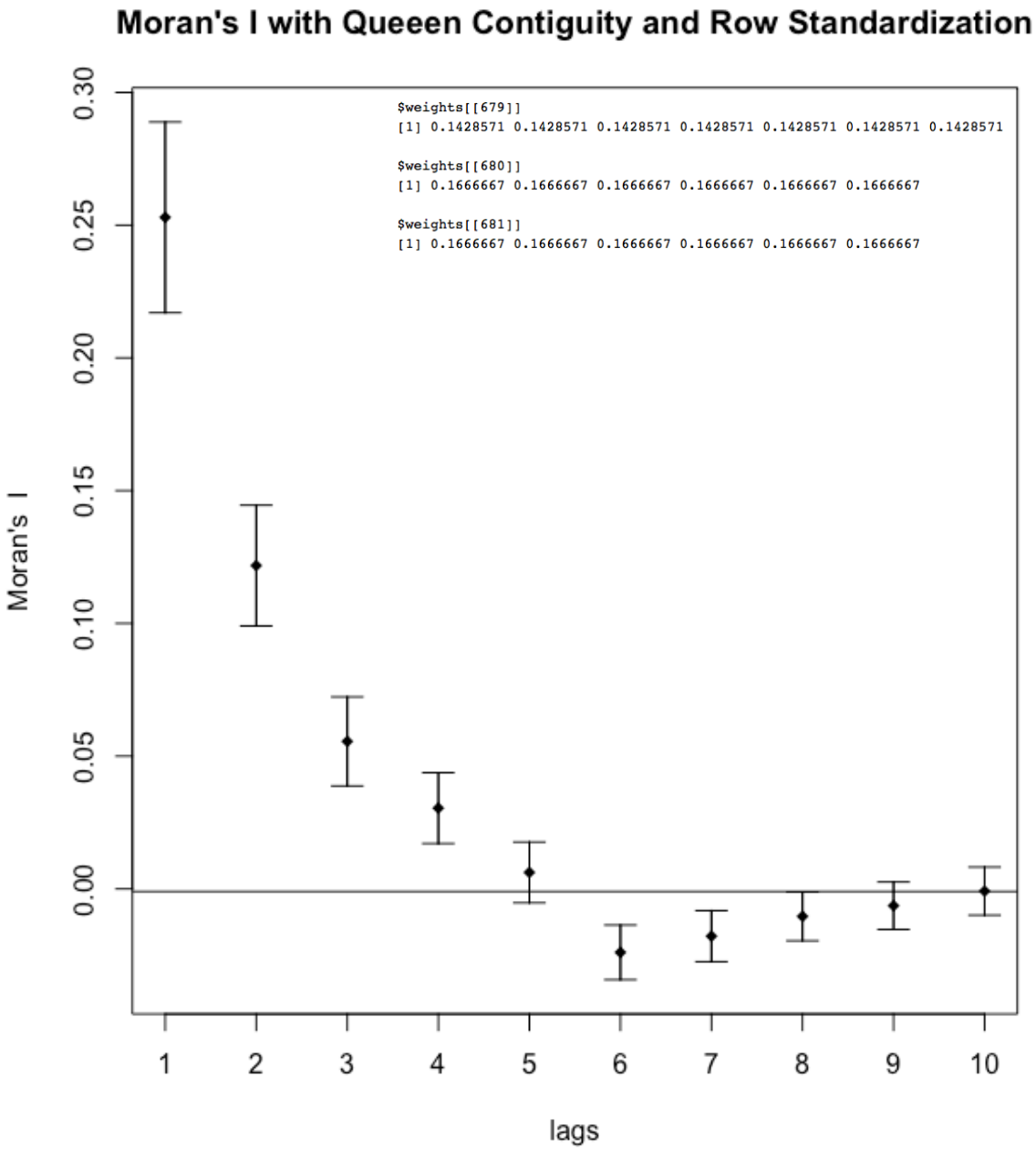
Weight Matrices with row weight standardization from Queen and Rook for item i

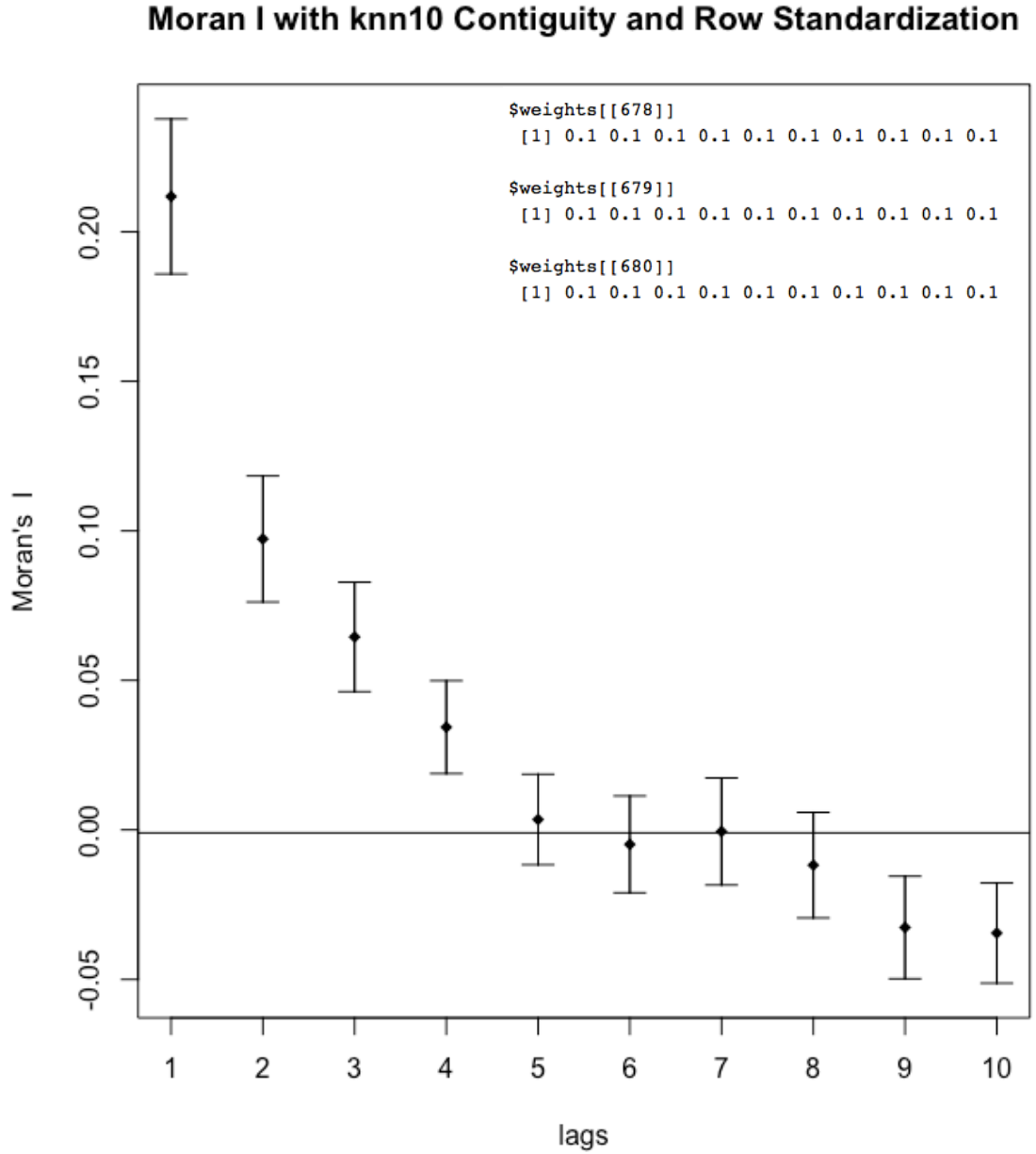
			1/8	1/8	1/8		
			1/8	i	1/8		
			1/8	1/8	1/8		

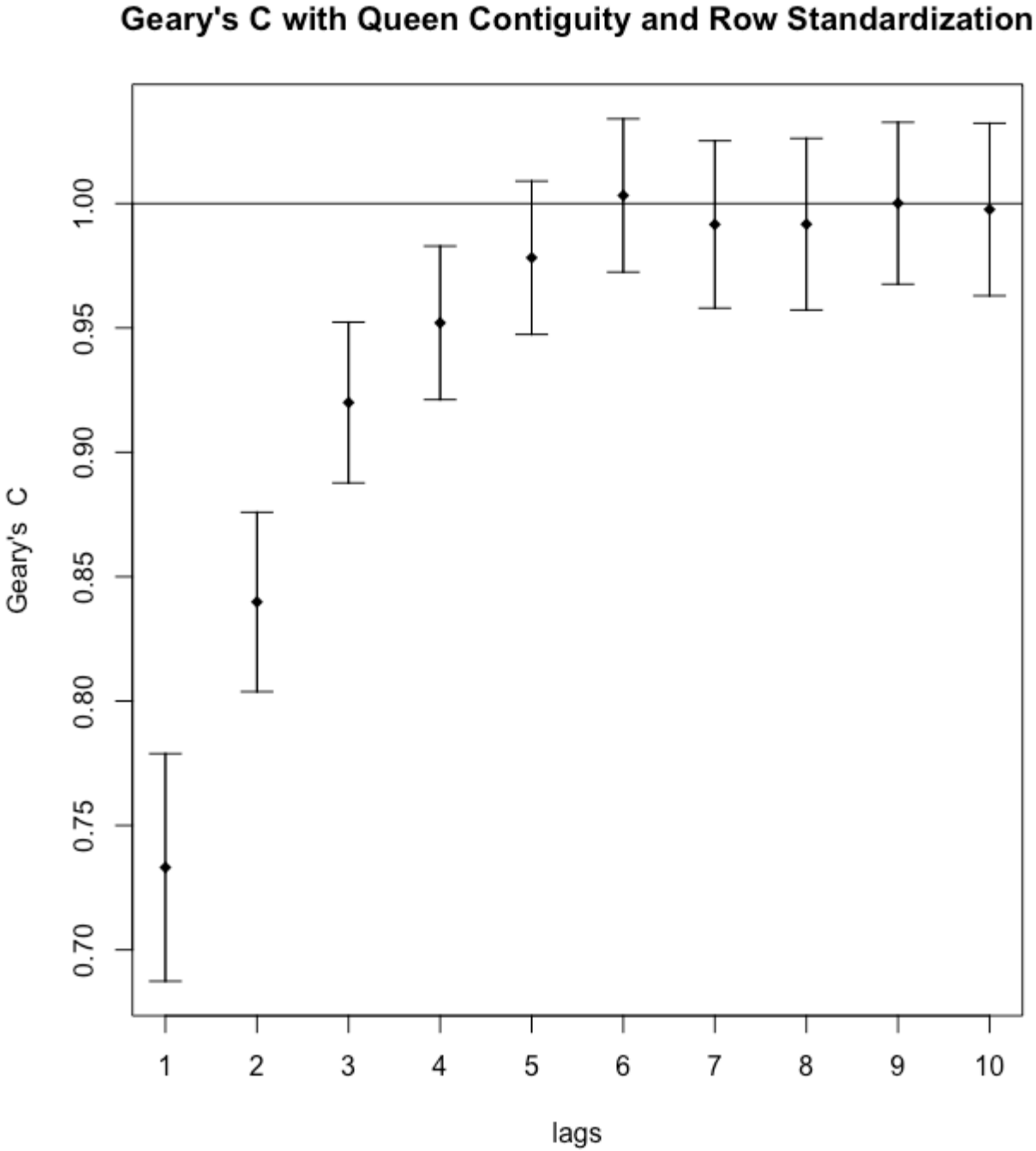
If I make the value of Y at i a linear sum of the values in each neighbor, I will multiply each neighbor's y by 1/8

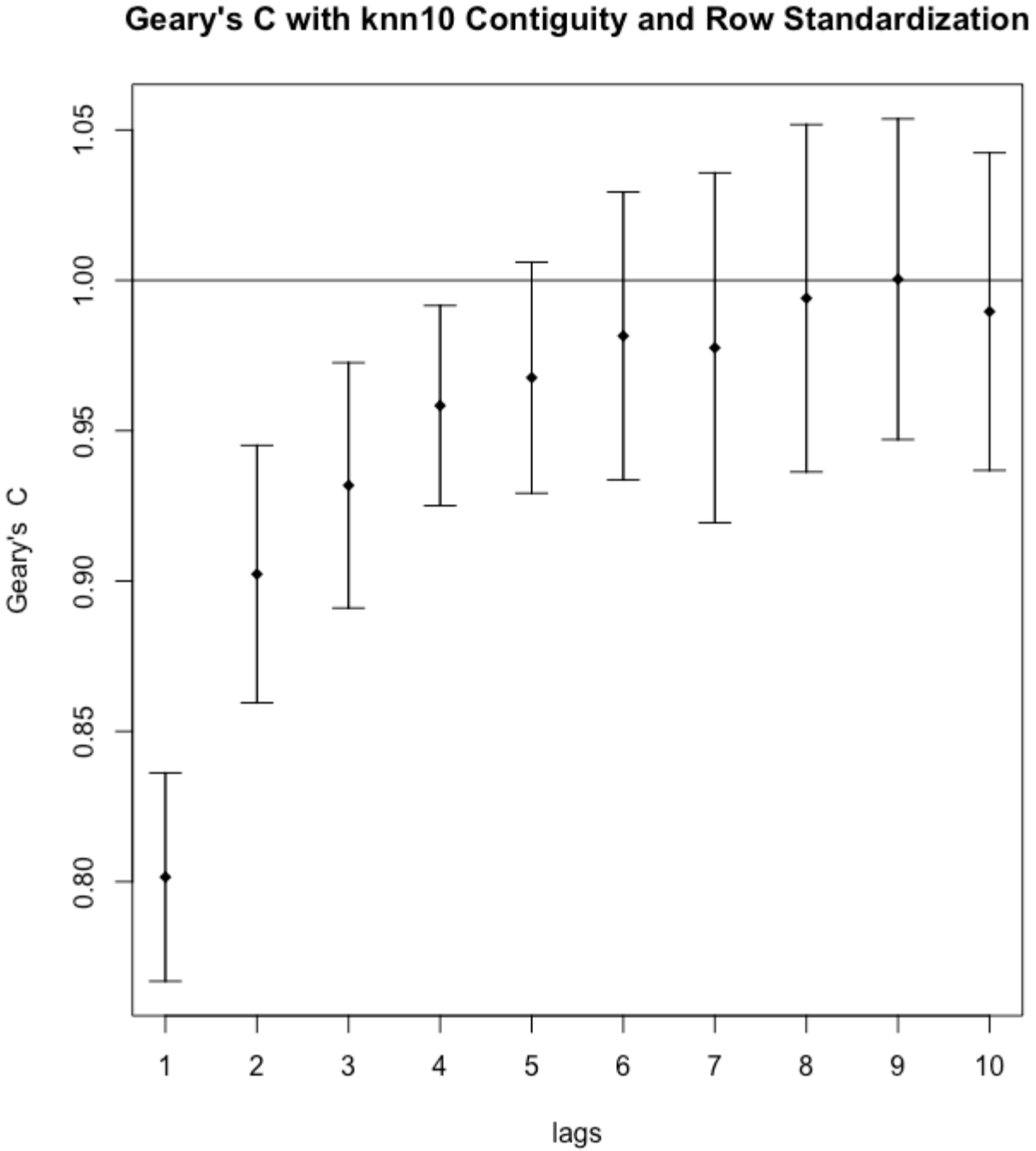
				1/4			
			1/4	i	1/4		
				1/4			

If I make the value of Y at i a linear sum of the values in each neighbor, I will multiply each neighbor's y by 1/4


Moran's I with Queeen Contiguity and Row Standardization
$weights[[679]]
[1] 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571 0.1428571
$weights[[680]]
[1] 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667
$weights[[681]]
[1] 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667
Moran's I
0.00
0.05
0.10
0.15
0.20
0.25
0.30
1
2
3
4
5
6
7
8
9
10
lags


Moran I with knn10 Contiguity and Row Standardization
$weights[[678]]
[1] 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
$weights[[679]]
[1] 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
$weights[[680]]
[1] 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1
Moran's I
-0.05
0.00
0.05
0.10
0.15
0.20
1
2
3
4
5
6
7
8
9
10
lags


Geary's C with Queen Contiguity and Row Standardization
Geary's C
0.70
0.75
0.80
0.85
0.90
0.95
1.00
1
2
3
4
5
6
7
8
9
10
lags


Geary's C with knn10 Contiguity and Row Standardization
Geary's C
0.80
0.85
0.90
0.95
1.00
1.05
1
2
3
4
5
6
7
8
9
10
lags

From slides 27-29 we can compute a z-score

```{r}
moran.test(YCOUNTY@data$VMTpr, listw=nb2listw(list.queenY))
```

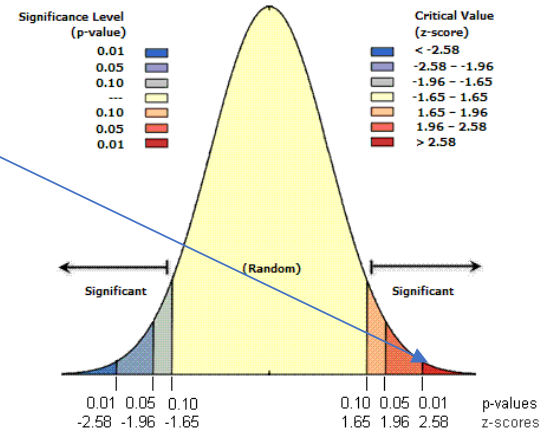
```
        Moran I test under randomisation

data:  YCOUNTY@data$VMTpr
weights: nb2listw(list.queenY)

Moran I statistic standard deviate = 14.137, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.2529411655     -0.0009718173      0.0003225841
```

In the Table above

```{r}
standarddeviate=((0.2529411655-(-0.0009718173))/sqrt(0.0003225841))
standarddeviate
```

$$z = \frac{(I - E_{NorR}(I))}{\sqrt{Var_{NorR}(I)}}$$

Z-scores are standard deviations. If, for example, a tool returns a z-score of +2.5, you would say that the result is 2.5 standard deviations. Both z-scores and p-values are associated with the standard normal distribution as shown below.



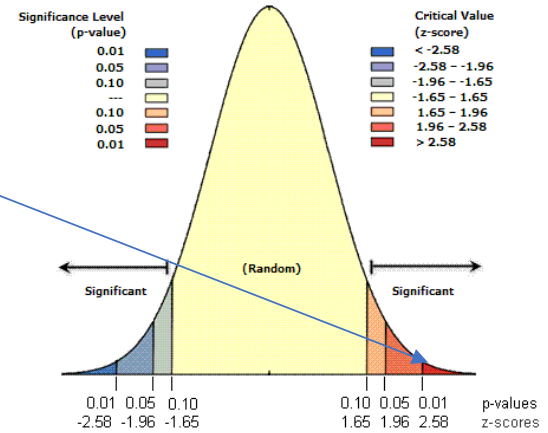
From slides 27-29 we can compute a z-score (this time under normality)

```{r}

moran.test(YCOUNTY@data$VMTpr, listw=nb2listw(list.queenY), randomisation=FALSE)

```

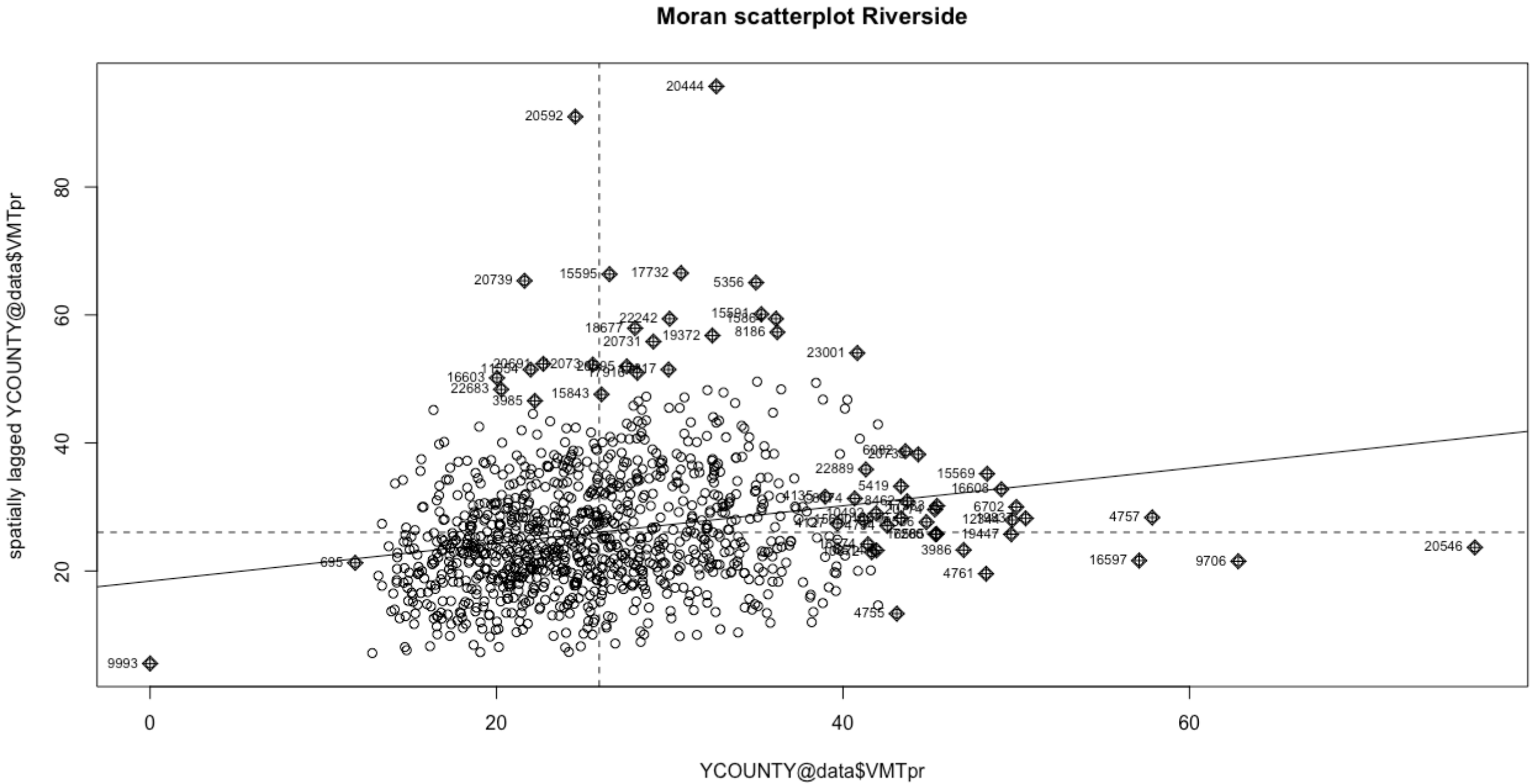
$$z = \frac{(I - E_{NorR}(I))}{\sqrt{Var_{NorR}(I)}}$$

```
        Moran I test under normality

data:  YCOUNTY@data$VMTpr
weights: nb2listw(list.queenY)

Moran I statistic standard deviate = 14.117, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.2529411655     -0.0009718173      0.0003235222
```

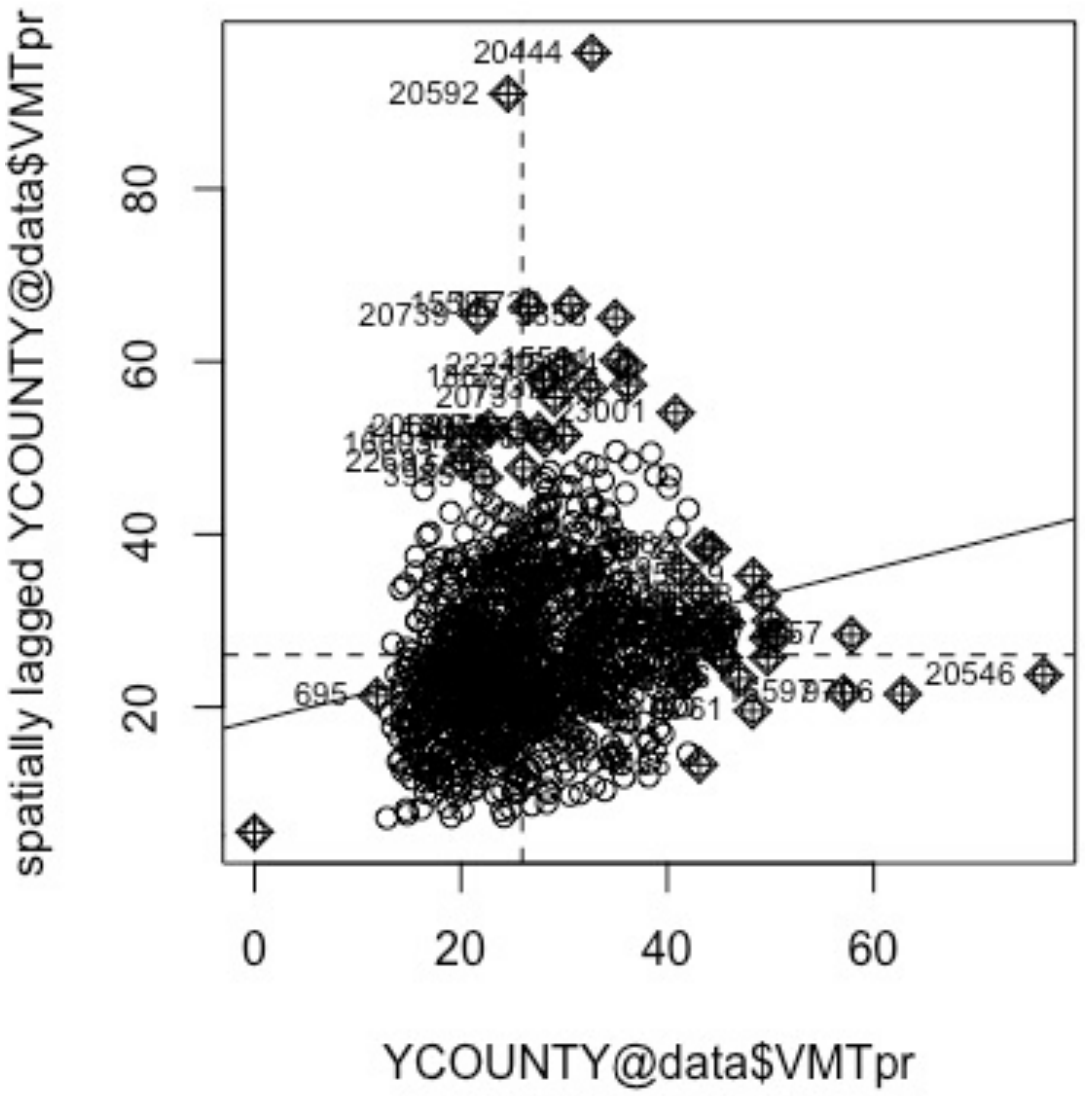
Z-scores are standard deviations. If, for example, a tool returns a z-score of +2.5, you would say that the result is 2.5 standard deviations. Both z-scores and p-values are associated with the standard normal distribution as shown below.



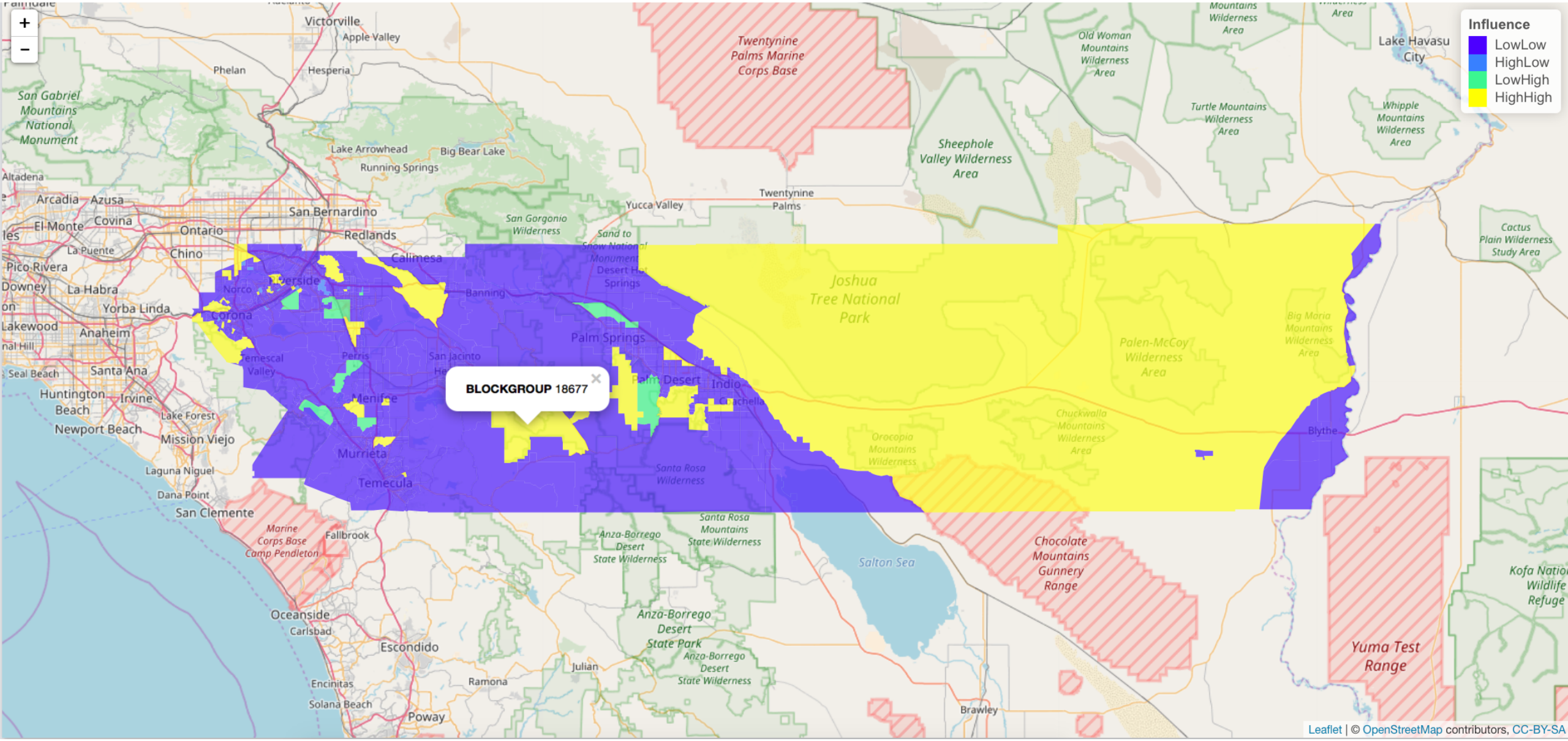
The standard deviate is not very different from the previous result.
This usually happens when we have many units (the polygons)


Moran scatterplot Riverside
spatially lagged YCOUNTY@data$VMTpr
YCOUNTY@data$VMTpr
0
20
40
60
80
20444
20592
20739
15595
17732
5356
22242
18677
19372
8186
20731
23001
16603
22683
3985
15843
22889
15569
5419
16608
6702
4757
19447
3986
16597
9706
20546
695
4761
4755
9993

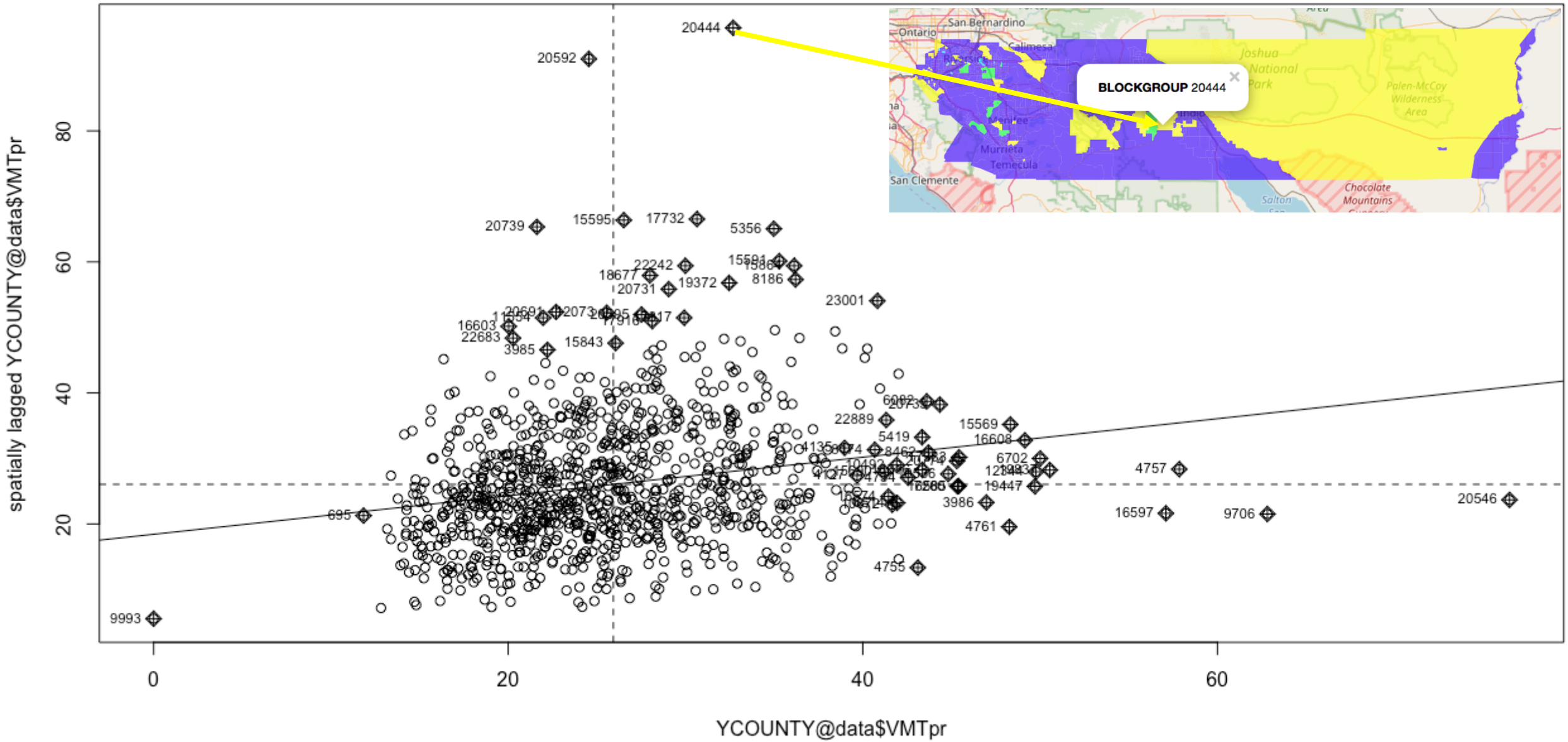
Moran scatterplot Riverside



Block groups with influence

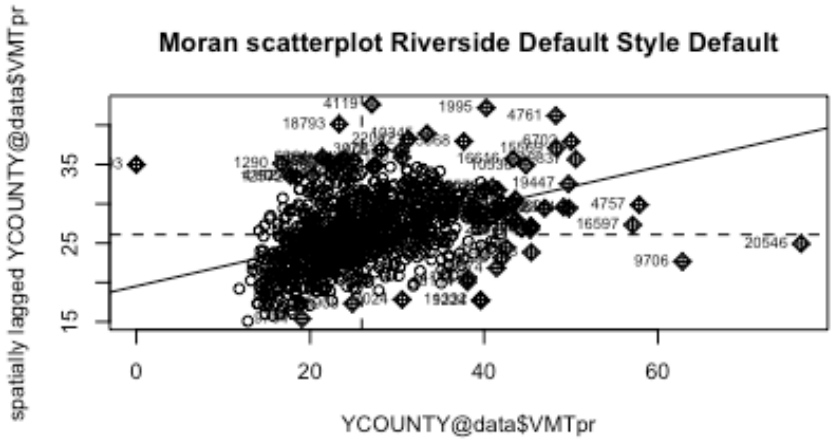

Influence
LowLow
HighLow
LowHigh
HighHigh
BLOCKGROUP 18677
Victorville
Apple Valley
Phelan
Hesperia
Twentynine Palms Marine Corps Base
Old Woman Mountains Wilderness Area
Lake Havasu City
San Gabriel Mountains National Monument
Lake Arrowhead
Big Bear Lake
Running Springs
Sheephole Valley Wilderness Area
Turtle Mountains Wilderness Area
Whipple Mountains Wilderness Area
Arcadia
Azusa
El Monte
Covina
Ontario
San Bernardino
Redlands
Yucca Valley
Twentynine Palms
San Gorgonio Wilderness
Chino
Calimesa
Cactus Plain Wilderness Study Area
Pico Rivera
La Habra
Yorba Linda
Anaheim
Joshua Tree National Park
Palm Springs
Big Maria Mountains Wilderness Area
Palen-McCoy Wilderness Area
Santa Ana
Irvine
Huntington Beach
Newport Beach
Lake Forest
Mission Viejo
Murrieta
Temecula
Indio
Blythe
Chuckwalla Mountains Wilderness Area
Orocopia Mountains Wilderness
Santa Rosa Wilderness
Laguna Niguel
Dana Point
San Clemente
Marine Corps Base Camp Pendleton
Fallbrook
Santa Rosa Mountains State Wilderness
Anza-Borrego Desert State Wilderness
Salton Sea
Chocolate Mountains Gunnery Range
Oceanside
Carlsbad
Escondido
Anza-Borrego Desert State Park
Julian
Ramona
Encinitas
Solana Beach
Poway
Brawley
Yuma Test Range
Leaflet | © OpenStreetMap contributors, CC-BY-SA

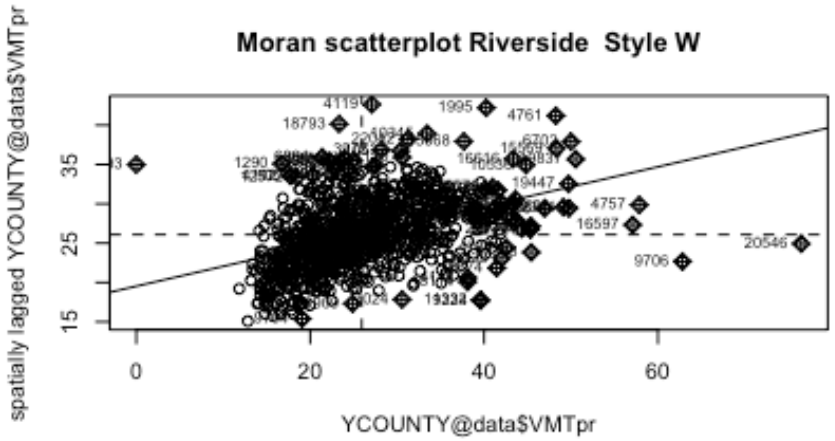
Moran scatterplot Riverside

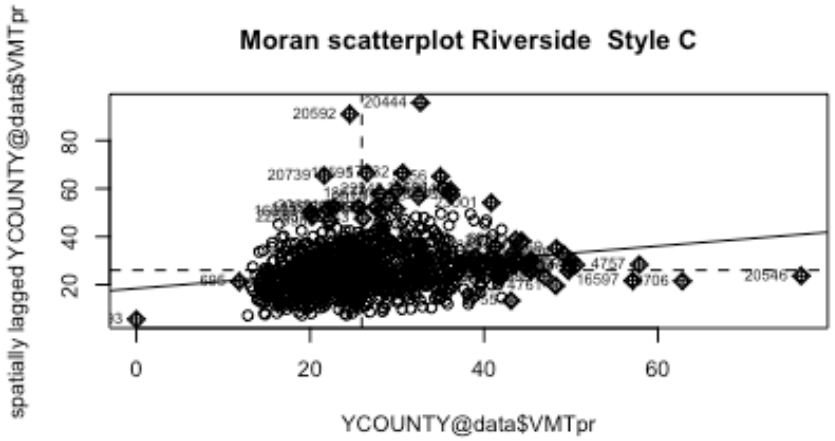


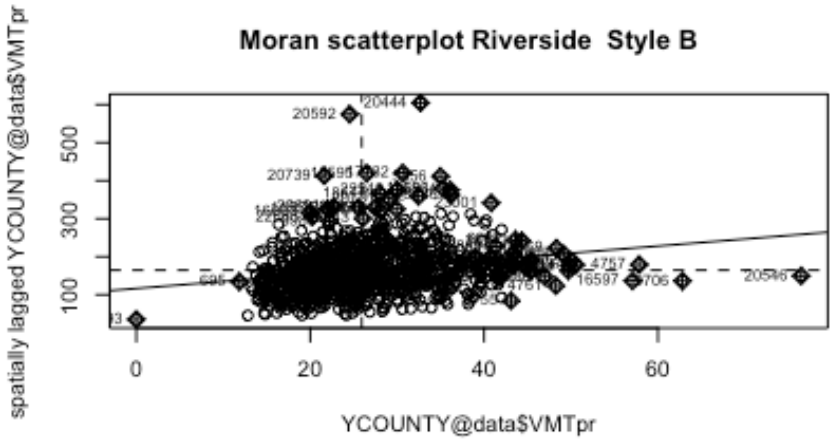
Different weighting schemes lead to different results

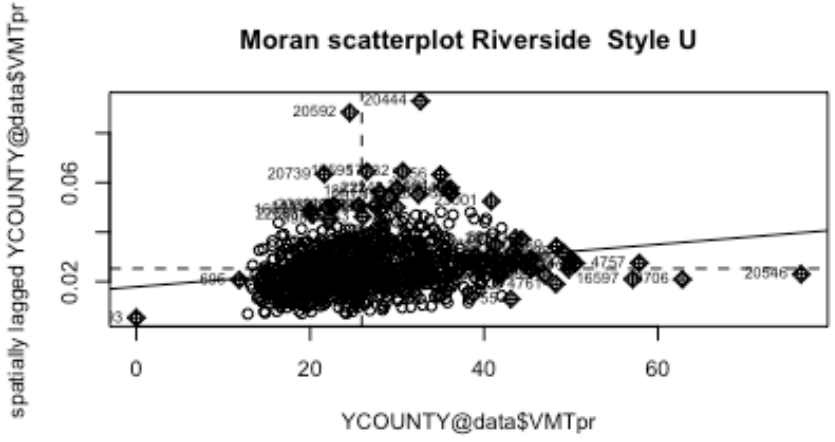
Neighboring values influence each observation with higher or lower impact and contribution to Moran’s I, Geary C or any other cross-product indicator

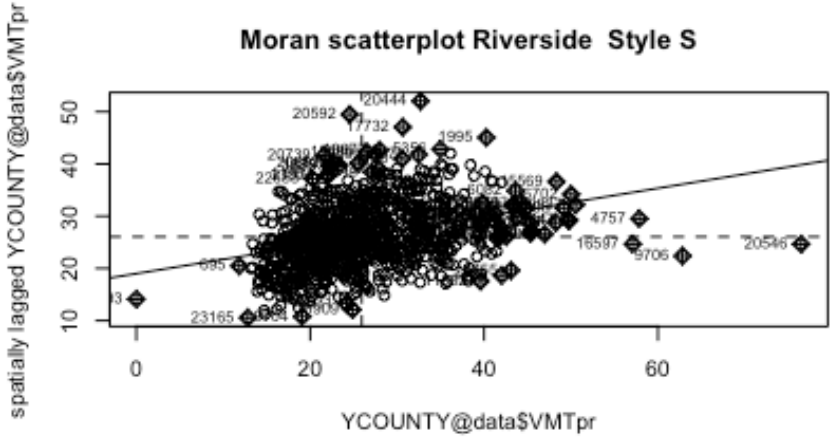

Moran scatterplot Riverside Default Style Default
spatially lagged YCOUNTY@data$VMTpr
YCOUNTY@data$VMTpr


Moran scatterplot Riverside Style W
spatially lagged YCOUNTY@data$VMTpr
YCOUNTY@data$VMTpr


Moran scatterplot Riverside Style C
spatially lagged YCOUNTY@data$VMTpr
YCOUNTY@data$VMTpr

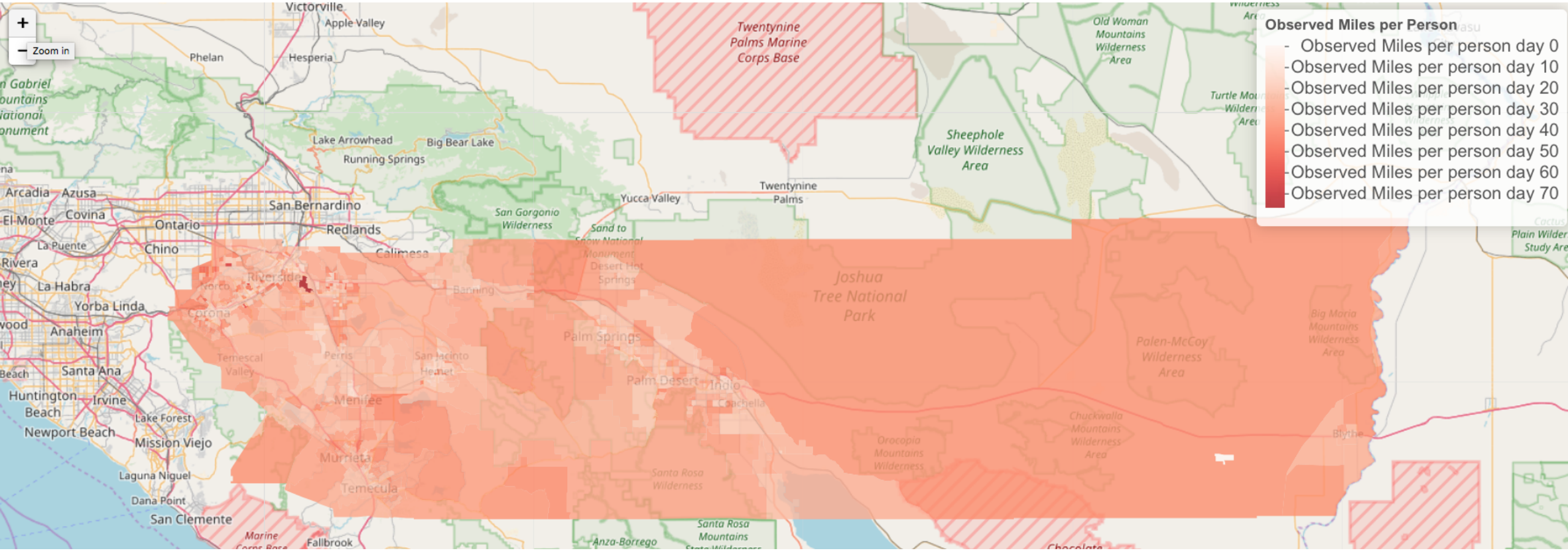

Moran scatterplot Riverside Style B
spatially lagged YCOUNTY@data$VMTpr
YCOUNTY@data$VMTpr

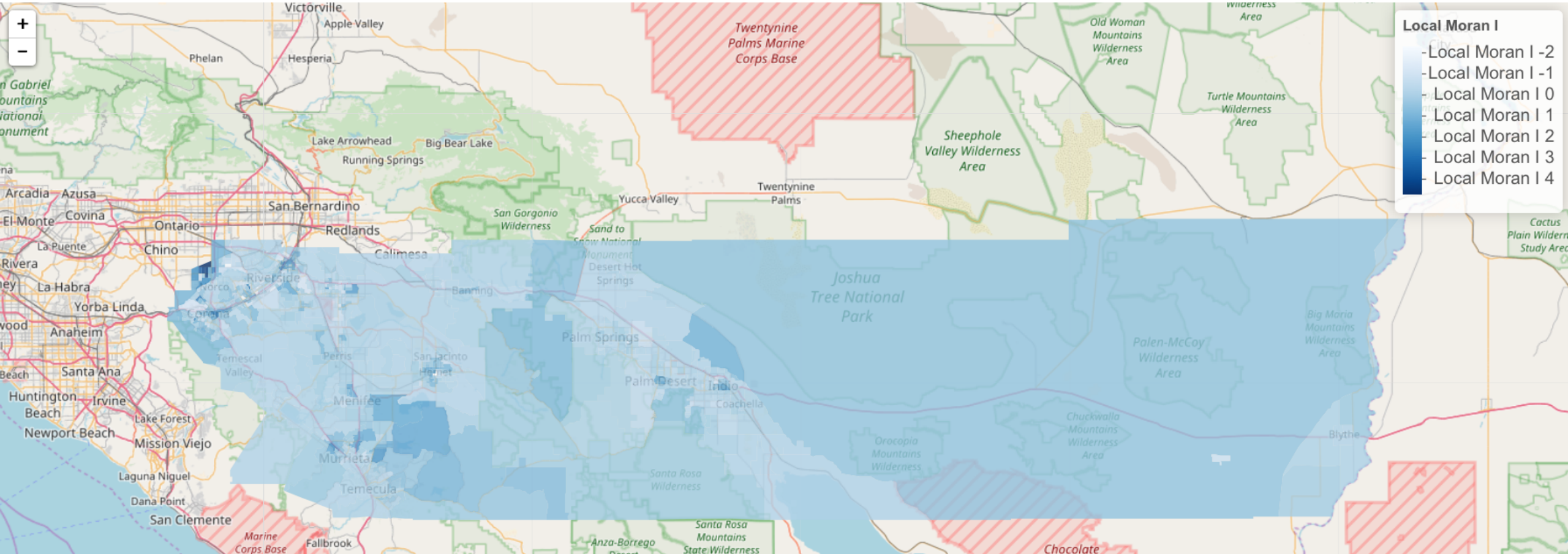

Moran scatterplot Riverside Style U
spatially lagged YCOUNTY@data$VMTpr
YCOUNTY@data$VMTpr

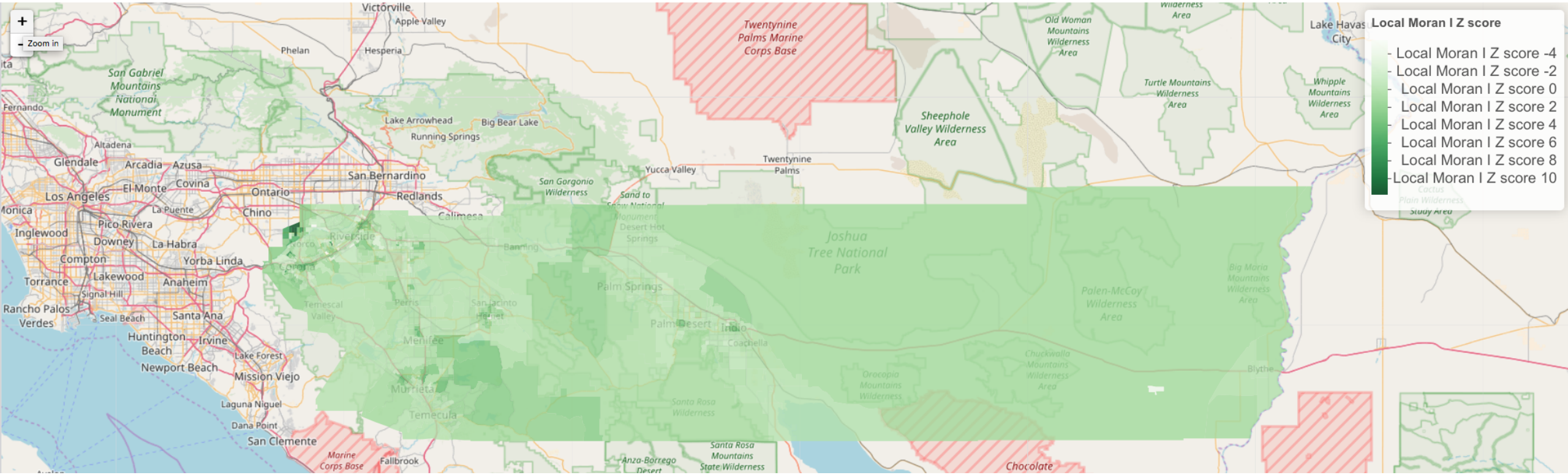

Moran scatterplot Riverside Style S
spatially lagged YCOUNTY@data$VMTpr
YCOUNTY@data$VMTpr

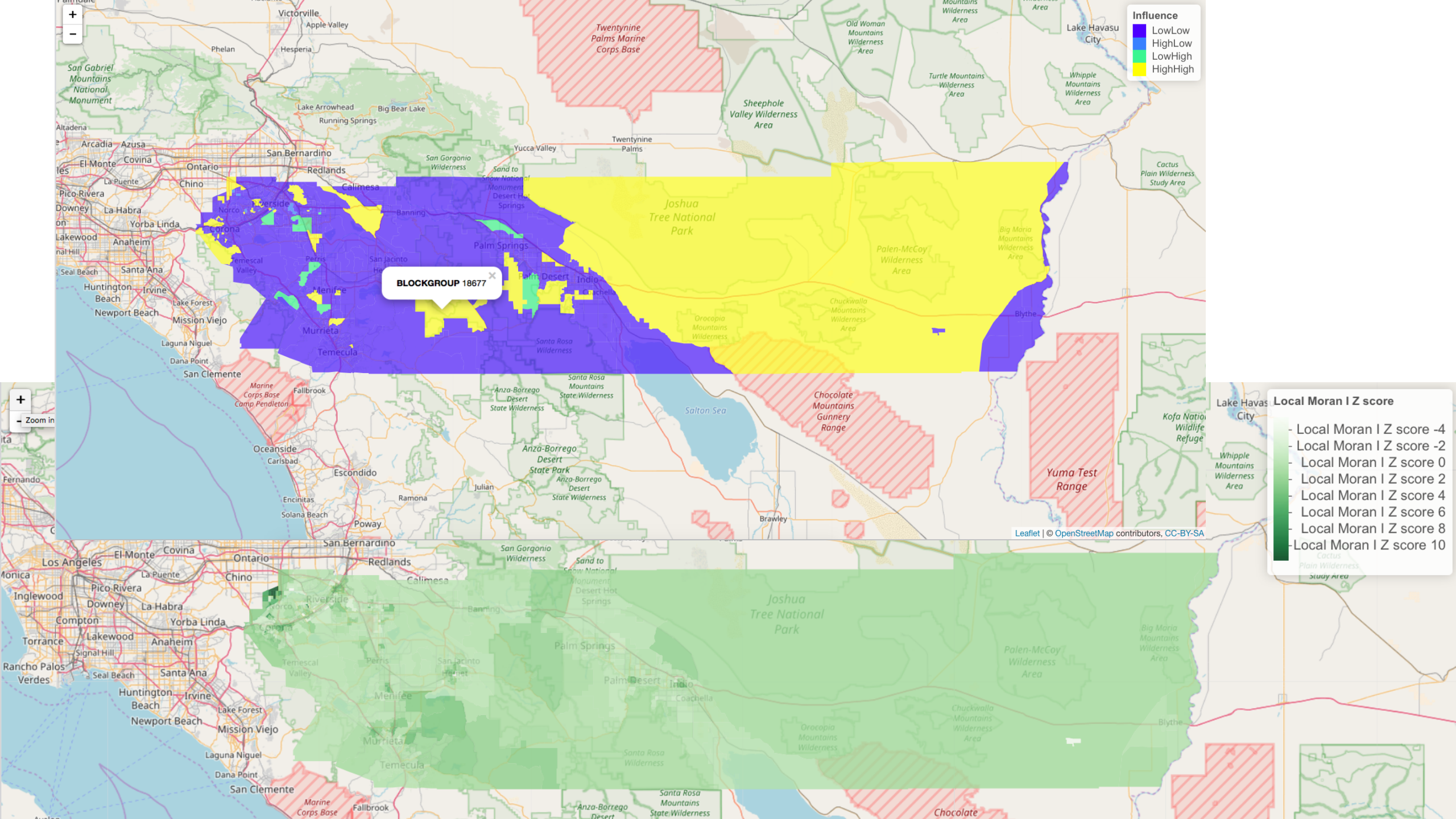
Local Moran I

$$I_i = \frac{(x_i - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2/(n-1)} \sum_{j=1}^{n} w_{ij}(x_j - \bar{x})$$


Observed Miles per Person
Observed Miles per person day 0
Observed Miles per person day 10
Observed Miles per person day 20
Observed Miles per person day 30
Observed Miles per person day 40
Observed Miles per person day 50
Observed Miles per person day 60
Observed Miles per person day 70
Zoom in
Victorville
Apple Valley
Phelan
Hesperia
Twentynine Palms Marine Corps Base
Old Woman Mountains Wilderness Area
Sheephole Valley Wilderness Area
Lake Arrowhead
Big Bear Lake
Running Springs
Twentynine Palms
Yucca Valley
San Gorgonio Wilderness
San Bernardino
Redlands
Ontario
Chino
Calimesa
Riverside
Banning
Corona
Joshua Tree National Park
Palm Springs
Palm Desert
Indio
Coachella
Palen-McCoy Wilderness Area
Big Maria Mountains Wilderness Area
Arcadia
Azusa
Covina
El Monte
La Puente
Rivera
La Habra
Yorba Linda
Anaheim
Santa Ana
Huntington Beach
Irvine
Newport Beach
Lake Forest
Mission Viejo
Laguna Niguel
Dana Point
San Clemente
Temecula
Murrieta
Menifee
Perris
Temescal Valley
San Jacinto
Hemet
Chuckwalla Mountains Wilderness Area
Orocopia Mountains Wilderness
Santa Rosa Wilderness
Santa Rosa Mountains
Anza-Borrego
Fallbrook
Blythe


Local Moran I
-Local Moran I -2
-Local Moran I -1
- Local Moran I 0
- Local Moran I 1
- Local Moran I 2
- Local Moran I 3
- Local Moran I 4
Victorville
Apple Valley
Phelan
Hesperia
Twentynine Palms Marine Corps Base
Old Woman Mountains Wilderness Area
Turtle Mountains Wilderness Area
Sheephole Valley Wilderness Area
Lake Arrowhead
Big Bear Lake
Running Springs
Twentynine Palms
Yucca Valley
San Gorgonio Wilderness
San Bernardino
Redlands
Ontario
Chino
Calimesa
Riverside
Banning
Joshua Tree National Park
Palm Springs
Palm Desert
Indio
Coachella
Palen-McCoy Wilderness Area
Big Maria Mountains Wilderness Area
Arcadia
Azusa
Covina
El Monte
La Puente
Rivera
La Habra
Yorba Linda
Anaheim
Santa Ana
Huntington Beach
Irvine
Lake Forest
Newport Beach
Mission Viejo
Laguna Niguel
Dana Point
San Clemente
Temescal Valley
Perris
San Jacinto
Menifee
Murrieta
Temecula
Santa Rosa Wilderness
Orocopia Mountains Wilderness
Chuckwalla Mountains Wilderness Area
Blythe
Marine Corps Base
Fallbrook
Anza-Borrego
Santa Rosa Mountains State Wilderness
Chocolate


Zoom in
Local Moran I Z score
Local Moran I Z score -4
Local Moran I Z score -2
Local Moran I Z score 0
Local Moran I Z score 2
Local Moran I Z score 4
Local Moran I Z score 6
Local Moran I Z score 8
Local Moran I Z score 10
Victorville
Apple Valley
Phelan
Hesperia
San Gabriel Mountains National Monument
Twentynine Palms Marine Corps Base
Old Woman Mountains Wilderness Area
Lake Havasu City
Turtle Mountains Wilderness Area
Whipple Mountains Wilderness Area
Sheephole Valley Wilderness Area
Lake Arrowhead
Big Bear Lake
Running Springs
Twentynine Palms
Yucca Valley
San Gorgonio Wilderness
Glendale
Arcadia
Azusa
El Monte
Covina
Los Angeles
Ontario
San Bernardino
Redlands
La Puente
Chino
Calimesa
Pico Rivera
Inglewood
Downey
La Habra
Riverside
Norco
Corona
Banning
Joshua Tree National Park
Compton
Yorba Linda
Lakewood
Anaheim
Torrance
Signal Hill
Palm Springs
Rancho Palos Verdes
Seal Beach
Santa Ana
Perris
San Jacinto
Temescal Valley
Palm Desert
Indio
Coachella
Palen-McCoy Wilderness Area
Big Maria Mountains Wilderness Area
Huntington Beach
Irvine
Lake Forest
Menifee
Newport Beach
Mission Viejo
Murrieta
Chuckwalla Mountains Wilderness Area
Blythe
Orocopia Mountains Wilderness
Laguna Niguel
Santa Rosa Wilderness
Temecula
Dana Point
San Clemente
Marine Corps Base
Fallbrook
Anza-Borrego Desert
Santa Rosa Mountains State Wilderness
Chocolate


Influence
LowLow
HighLow
LowHigh
HighHigh
BLOCKGROUP 18677
Local Moran I Z score
Local Moran I Z score -4
Local Moran I Z score -2
Local Moran I Z score 0
Local Moran I Z score 2
Local Moran I Z score 4
Local Moran I Z score 6
Local Moran I Z score 8
Local Moran I Z score 10
Zoom in
Leaflet | © OpenStreetMap contributors, CC-BY-SA

Intro to spatial regression

Recall that in the Linear Regression Model

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

$\varepsilon_i \sim N(0, \sigma^2)$

$\text{Cov}(\varepsilon_i \varepsilon_j) = 0$ residuals are independent

This is a convenient simplification and allows the use of “simple” estimation methods for the regression coefficients

What happens if we have spatial influence?

- Do we know the data generating process (DGP) ?
- What type of spatial units do we have?
- What type of variables are we analyzing?
- Observed (dependent variable and/or independent variables) spatial dependency?
- Unobserved (error term) spatial dependency?

- Answers to these questions guide the way we create models

Data Generating Process (linear regression)

$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$

$\varepsilon_i \sim N(0, \sigma^2)$

$Cov(\varepsilon_i \varepsilon_j) = 0$ residuals are independent

This says that we have i=1,…,N observations that are independent from each other. The values of y from observation i are determined by xs that are of unit i alone. It also says the coefficients betas are as many as the xs and they are the same for all observations.

It also says that the error terms are all distributed according to the Normal distribution with the same mean and standard deviation sigma and the random error terms between unit i and unit j for i different than j are uncorrelated

Data Generating Process (a neighbor influences another neighbor)

$y_i = \alpha_i + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i + \gamma w_{ij} y_j$

$\varepsilon_i \sim N(0, \sigma^2)$

$Cov(\varepsilon_i \varepsilon_j) = 0$ residuals are independent

This says that we have i=1,…,N observations that are independent from each other except for the w_{ij} relationship. The values of y from observation i are determined by xs that are of unit i and what happens in the values of y to a neighboring unit j (w is the amount of influence from j to i). It also says the coefficients betas are as many as the xs and they are the same for all observations.

It also says that the error terms are all distributed according to the Normal distribution with the same mean and standard deviation sigma and the random error terms between unit i and unit j for i different than j are uncorrelated

Many different Data Generating Processes

- We can have Y's depending on the Ys of neighbors
- We can have X's depending on the Xs of neighbors
- We can have random error terms depending on the error terms of neighbors
- We can have spatial dependency on all (Ys, Xs, and es)
- We can have temporal dependency on all (Ys, Xs, and es)
- We can also have spatial and temporal dependencies (spatial panel data)

- The more dependencies we have the more complex the Data Generating Process gets and the harder etiology becomes.