Introduction of geospatial data visualization and geographically weighted regression (GWR)

Frank Fan

Vanderbilt University

August 16, 2012

Study Background

(1) Information of breast cancer or population characteristic such as mortality rate, median household income, and ethnicity ...etc around Nashville area or US.

Study Background

(1) Information of breast cancer or population characteristic such as mortality rate, median household income, and ethnicity ...etc around Nashville area or US.

(2) Data are summarized in different levels.

- State
- County
- Zip Code

Study Background

(1) Information of breast cancer or population characteristic such as mortality rate, median household income, and ethnicity ...etc around Nashville area or US.

(2) Data are summarized in different levels.
   - State
   - County
   - Zip Code

(3) The goal is generating several geographic maps with color representing information of interests.

Tennessee

Zip Code	Zip City	County	Female Mortality Rate per 100K
37013	Antioch	Davidson	20.05
37072	Goodlettsville	Davidson	28.13
37076	Hermitage	Davidson	25.34
37080	Joelton	Davidson	29.39
37115	Madison	Davidson	31.52
37138	Old Hickory	Davidson	28.71
37189	Whites Creek	Davidson	34.24
37201	Nashville	Davidson	27.40
.	.	.	.
.	.	.	.
.	.	.	.

Tasks

- Generate a color scale

  How to represent, for example mortality rate, in a map with color ?

Tasks

- Generate a color scale

  How to represent, for example mortality rate, in a map with color ?
- Draw a map

  Where to get information or R functions to plot the map ?

Tasks

- Generate a color scale

  How to represent, for example mortality rate, in a map with color ?
- Draw a map

  Where to get information or R functions to plot the map ?
- Fill in the color

  Who is who ?

Get Colors

- The idea is get a single color scale showing severity of mortality from light (less server) to dark (more server).

Get Colors

- The idea is get a single color scale showing severity of mortality from light (less server) to dark (more server).
- *colorRampPalette* {*grDevices*}
  Return as a function

Get Colors

- The idea is get a single color scale showing severity of mortality from light (less server) to dark (more server).
- *colorRampPalette* {*grDevices*}
  Return as a function
- Usage :
  color.scales <- colorRampPalette(c("orange", "red"))
  col.lists <- color.scales(8)

Get Colors

- The idea is get a single color scale showing severity of mortality from light (less server) to dark (more server).
- *colorRampPalette* {*grDevices*}
  Return as a function
- Usage :
  color.scales <- colorRampPalette(c("orange", "red"))
  col.lists <- color.scales(8)
- *col.lists*
  "#FFA500" "#FF8D00" "#FF7500" "#FF5E00"
  "#FF4600" "#FF2F00" "#FF1700" "#FF0000"

Get Colors

- The idea is get a single color scale showing severity of mortality from light (less server) to dark (more server).
- *colorRampPalette* {*grDevices*}
  Return as a function
- Usage :
  color.scales <- colorRampPalette(c("orange", "red"))
  col.lists <- color.scales(8)
- *col.lists*
  "#FFA500" "#FF8D00" "#FF7500" "#FF5E00"
  "#FF4600" "#FF2F00" "#FF1700" "#FF0000"
- It's called hexadecimal color codes.
  There are also RGB color function in R available.
  rgb(228,26,28,max=255)

8 Levels Between Orange and Red

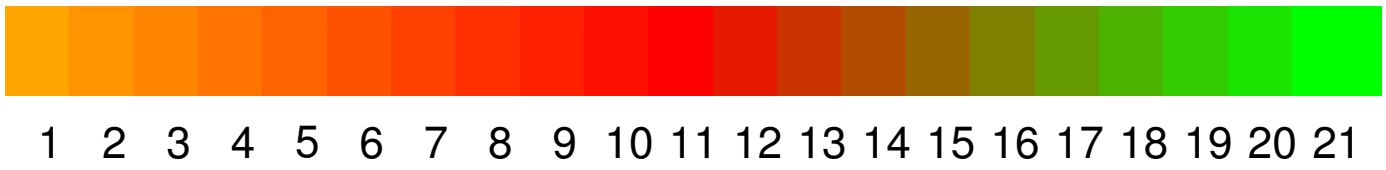
1 2 3 4 5 6 7 8

More Colors ?!?!

- Can we get more than two colors and have the scales between all of them ?

More Colors ?!?!

- Can we get more than two colors and have the scales between all of them ?
- Usage :
  color.scales <- colorRampPalette(c("orange", "red", "green"))
  col.lists <- color.scales(21)



Zip Code Map

- Maps package doesn't have either zip code's boundary or maps of Alaska & Hawaii, so I can't use map function to get those plotted.

Zip Code Map

- Maps package doesn't have either zip code's boundary or maps of Alaska & Hawaii, so I can't use map function to get those plotted.
- More tasks need to be done.

Zip Code Map

- Maps package doesn't have either zip code's boundary or maps of Alaska & Hawaii, so I can't use map function to get those plotted.
- More tasks need to be done.
  - Zip code boundary data (or Alaska & Hawaii boundary data)

Zip Code Map

- Maps package doesn't have either zip code's boundary or maps of Alaska & Hawaii, so I can't use map function to get those plotted.
- More tasks need to be done.
  - Zip code boundary data (or Alaska & Hawaii boundary data)
  - R function to draw the frame of each zip code area.

Zip Code Map

- Maps package doesn't have either zip code's boundary or maps of Alaska & Hawaii, so I can't use map function to get those plotted.
- More tasks need to be done.
  - Zip code boundary data (or Alaska & Hawaii boundary data)
  - R function to draw the frame of each zip code area.
  - Same or different function to fill in the color.

Zip Code Boundaries

- Boundary Data.

Zip Code Boundaries

- Boundary Data.
- US Census

Zip Code Boundaries

- Boundary Data.
- US Census
- It's back !! called "TIGER/Line Shapefiles"

Zip Code Boundaries

- Boundary Data.
- US Census
- It's back !! called "TIGER/Line Shapefiles"
- http://www.census.gov/geo/www/tiger/shp.html

Zip Code Boundaries

- Boundary Data.
- US Census
- It's back !! called "TIGER/Line Shapefiles"
- http://www.census.gov/geo/www/tiger/shp.html
- http://www.census.gov/cgi-bin/geo/shapefiles2010/main

Maptools Package

- Need a way to read .shp file into R.

Maptools Package

- Need a way to read .shp file into R.
- *readShapeSpatial*{*maptools*}

Maptools Package

- Need a way to read .shp file into R.
- *readShapeSpatial*{*maptools*}
- *TN.zip.map <- readShapeSpatial("zt47_d00.shp")*

Maptools Package

- Need a way to read .shp file into R.
- *readShapeSpatial{maptools}*
- *TN.zip.map <- readShapeSpatial("zt47_d00.shp")*
- *str(TN.zip.map)*

```
Formal class 'SpatialPolygonsDataFrame' [package ''sp"] with
5 slots
..@ data :'data.frame':  964 obs.  of 8 variables:
```

Maptools Package

- Need a way to read .shp file into R.
- *readShapeSpatial{maptools}*
- *TN.zip.map <- readShapeSpatial("zt47_d00.shp")*
- *str(TN.zip.map)*

```
Formal class 'SpatialPolygonsDataFrame' [package ''sp"] with
5 slots
..@ data :'data.frame':  964 obs.  of 8 variables:
```

- and another 1277 more lines !!!

Stracture of .shp File

```
Formal class 'SpatialPolygonsDataFrame' [package ``sp"] with 5 slots
 ..@ data :'data.frame':  964 obs.  of 8 variables:
 ..  ..$ AREA : num [1:964] 2.92e-03 2.34e-02 4.18e-03 2.02e-05 2.95e-05 ...
 ..  ..$ PERIMETER : num [1:964] 0.8641 1.3754 2.9457 0.0226 0.0326 ...
 ..  ..$ ZT47_D00_ :  int [1:964] 2 3 4 5 6 7 8 9 10 11 ...
 ..  ..$ ZT47_D00_I: int [1:964] 1 2 3 4 5 6 7 8 9 10 ...
 ..  ..$ ZCTA : Factor w/ 634 levels ``37010",``37012",..:  65 66 65 15 15 15 88 108 81 98 ...
 ..  ..$ NAME : Factor w/ 634 levels ``37010",``37012",..:  65 66 65 15 15 15 88 108 81 98 ...
 ..  ..$ LSAD : Factor w/ 1 level ``Z5":  1 1 1 1 1 1 1 1 1 1 ...
 ..  ..$ LSAD_TRANS: Factor w/ 1 level ``5-Digit ZCTA":  1 1 1 1 1 1 1 1 1 1 ...
 ..  ..- attr(*, ``data_types")= chr [1:8] ``F" ``F" ``N" ``N" ...
 ..@ polygons :List of 964
 ..  ..$ :Formal class 'Polygons' [package ``sp"] with 5 slots
 ..  ..  ..  ..@ Polygons :List of 1
 ..  ..  ..  ..  ..$ :Formal class 'Polygon' [package ``sp"] with 5 slots
 ..  ..  ..  ..  ..  ..  ..@ labpt :  num [1:2] -88 36.6
 ..  ..  ..  ..  ..  ..  ..@ area :  num 0.00292
 ..  ..  ..  ..  ..  ..  ..@ hole :  logi FALSE
 ..  ..  ..  ..  ..  ..  ..@ ringDir:  int 1
 ..  ..  ..  ..  ..  ..  ..@ coords :  num [1:169, 1:2] -88 -88 -88 -88 -88 ...
 ..  ..  ..  ..@ plotOrder:  int 1
 ..  ..  ..  ..@ labpt :  num [1:2] -88 36.6
 ..  ..  ..  ..@ ID : chr "0"
 ..  ..  ..  ..@ area :  num 0.00292
```

Decompose the Structure

- Take Hawaii as an example.

Decompose the Structure

- Take Hawaii as an example.
- Think it as a box with lots boxes inside.

Decompose the Structure

- Take Hawaii as an example.
- Think it as a box with lots boxes inside.
- One huge box : object itself.

Decompose the Structure

- Take Hawaii as an example.
- Think it as a box with lots boxes inside.
- One huge box : object itself.
- 5 middle size box :

Decompose the Structure

- Take Hawaii as an example.
- Think it as a box with lots boxes inside.
- One huge box : object itself.
- 5 middle size box :
  - General information.

Decompose the Structure

- Take Hawaii as an example.
- Think it as a box with lots boxes inside.
- One huge box : object itself.
- 5 middle size box :
  - General information.
  - Main parts with numbers of small boxes inside.

Decompose the Structure

- Take Hawaii as an example.
- Think it as a box with lots boxes inside.
- One huge box : object itself.
- 5 middle size box :
  - General information.
  - Main parts with numbers of small boxes inside.
    - Coordinates.

Decompose the Structure

- Take Hawaii as an example.
- Think it as a box with lots boxes inside.
- One huge box : object itself.
- 5 middle size box :
  - General information.
  - Main parts with numbers of small boxes inside.
    - Coordinates.
    - Center Coordinate.

Decompose the Structure

- Take Hawaii as an example.
- Think it as a box with lots boxes inside.
- One huge box : object itself.
- 5 middle size box :
  - General information.
  - Main parts with numbers of small boxes inside.
    - Coordinates.
    - Center Coordinate.
    - Area, ID, ...etc.

Decompose the Structure

- Take Hawaii as an example.
- Think it as a box with lots boxes inside.
- One huge box : object itself.
- 5 middle size box :
  - General information.
  - Main parts with numbers of small boxes inside.
    - Coordinates.
    - Center Coordinate.
    - Area, ID, ...etc.
    - No zip code.

Decompose the Structure

- Take Hawaii as an example.
- Think it as a box with lots boxes inside.
- One huge box : object itself.
- 5 middle size box :
  - General information.
  - Main parts with numbers of small boxes inside.
    - Coordinates.
    - Center Coordinate.
    - Area, ID, ...etc.
    - No zip code.
  - The other 3 median boxes which is not so important.

Decompose the Structure

- Take Hawaii as an example.
- Think it as a box with lots boxes inside.
- One huge box : object itself.
- 5 middle size box :
  - General information.
  - Main parts with numbers of small boxes inside.
    - Coordinates.
    - Center Coordinate.
    - Area, ID, ...etc.
    - No zip code.
  - The other 3 median boxes which is not so important.
- Assumption : the zip code is saved with the order.

Plot the Shape & Colored Only by Coordinates

- Need a function to plot the shape by only giving a set of coordinates.

Plot the Shape & Colored Only by Coordinates

- Need a function to plot the shape by only giving a set of coordinates.
- If can handle the color as welll will be better.

Plot the Shape & Colored Only by Coordinates

- Need a function to plot the shape by only giving a set of coordinates.
- If can handle the color as welll will be better.
- *polygon*{*graphics*}

Plot the Shape & Colored Only by Coordinates

- Need a function to plot the shape by only giving a set of coordinates.
- If can handle the color as welll will be better.
- *polygon{graphics}*
- Useage :

$$polygon(x,\ y = NULL,\ density = NULL,\ angle = 45,\ border = NULL,\ col = NA,\ lty = par(\text{"lty"}),\ldots)$$

Representing Information with Colors

Zip Code	Female Mortality Rate per 100K	20% Quantile
37013	20.05	1
37072	28.13	4
37076	25.34	3
37115	22.52	2
37138	28.71	5
37189	34.24	5
37201	27.40	4
.	.	.
.	.	.
.	.	.

Representing Information with Colors

Zip Code	Female Mortality Rate per 100K	20% Quantile	Corresponding Color
37013	20.05	1	#9999FF
37072	28.13	4	#4C4CFF
37076	25.34	3	#6666FF
37115	22.52	2	#7F7FFF
37138	28.71	5	#3333FF
37189	34.24	5	#3333FF
37201	27.40	4	#4C4CFF
.	.	.	.
.	.	.	.
.	.	.	.

Two Parts of Data

- Zip code with corresponding color.

Two Parts of Data

- Zip code with corresponding color.
- Boundary coordinates for each zip code.

Two Parts of Data

- Zip code with corresponding color.
- Boundary coordinates for each zip code.
- Now we are ready to plot !!!!

GWR

- PI's question : Does breast cancer mortality related to education, household income...etc.

GWR

- PI's question : Does breast cancer mortality related to education, household income...etc.
- My first idea : Spearman's correlation.

GWR

- PI's question : Does breast cancer mortality related to education, household income...etc.
- My first idea : Spearman's correlation.
- The problem are :
  - What if we have more or fewer data points ?
  - Simple correlation can't deal with multiple covariates at the same time.

GWR

- PI's question : Does breast cancer mortality related to education, household income...etc.
- My first idea : Spearman's correlation.
- The problem are :
  - What if we have more or fewer data points ?
  - Simple correlation can't deal with multiple covariates at the same time.
- Brunsdon (1996), "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity", Geographical analysis, Vol. 28, No. 4, P.281.

GWR

- Wherever a house is located, for example, the marginal price increase associated with an additional bedroom is fixed.

GWR

- Wherever a house is located, for example, the marginal price increase associated with an additional bedroom is fixed.
- It might be more reasonable to assume that rates of change are determined by local culture or local knowledge, rather than a global utility assumed for each commodity.

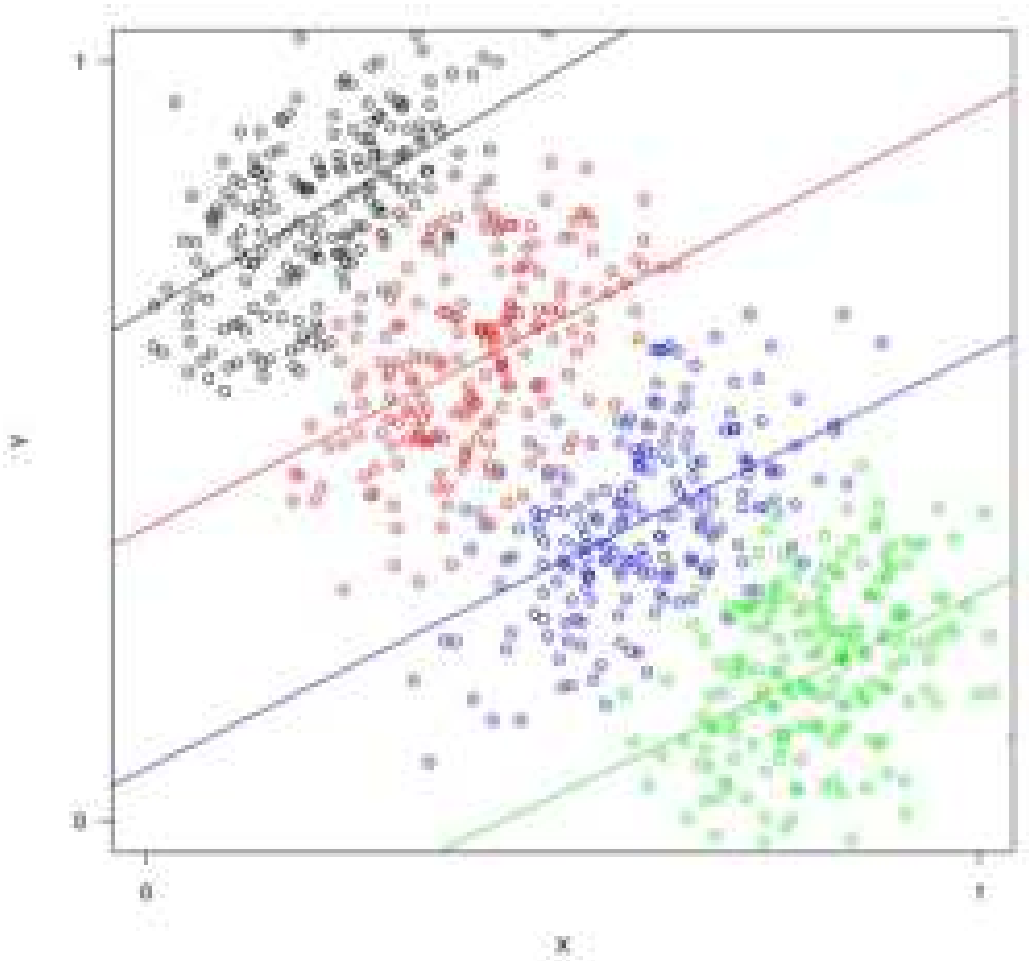
GWR

- Wherever a house is located, for example, the marginal price increase associated with an additional bedroom is fixed.
- It might be more reasonable to assume that rates of change are determined by local culture or local knowledge, rather than a global utility assumed for each commodity.
- Variations in relationships over space, such as those described above, are referred to as spatial nonstationarity.

GWR

- Wherever a house is located, for example, the marginal price increase associated with an additional bedroom is fixed.
- It might be more reasonable to assume that rates of change are determined by local culture or local knowledge, rather than a global utility assumed for each commodity.
- Variations in relationships over space, such as those described above, are referred to as spatial nonstationarity.
- Simple linear model is not not appropriate, but widely used in early 90s for the data with geographic information.

Simpson's Paradox



GWR

- $y = \alpha + \beta X + \epsilon$

GWR

- $y = \alpha + \beta X + \epsilon$
- $y_i = \alpha_i + \beta_i X + \epsilon_i$
- Where i indicates that there is a set of coefficients estimated for every observation in the data set.

GWR

- $y = \alpha + \beta X + \epsilon$
- $y_i = \alpha_i + \beta_i X + \epsilon_i$
- Where i indicates that there is a set of coefficients estimated for every observation in the data set.
- Key difference : We estimate a set of regression coefficients for each observation (i).

GWR

- $y = \alpha + \beta X + \epsilon$
- $y_i = \alpha_i + \beta_i X + \epsilon_i$
- Where i indicates that there is a set of coefficients estimated for every observation in the data set.
- Key difference : We estimate a set of regression coefficients for each observation (i).
- To do so we weight near observations more heavily than more distant ones.

GWR

- $y = \alpha + \beta X + \epsilon$
- $y_i = \alpha_i + \beta_i X + \epsilon_i$
- Where i indicates that there is a set of coefficients estimated for every observation in the data set.
- Key difference : We estimate a set of regression coefficients for each observation (i).
- To do so we weight near observations more heavily than more distant ones.
- First law of geography: everything is related with everything else, but closer things are more related.

GWR

- Decide how you are going to weight your nearby locations.
  - Fixed bandwidth
  - Variable bandwidth
  - User-defined bandwidth

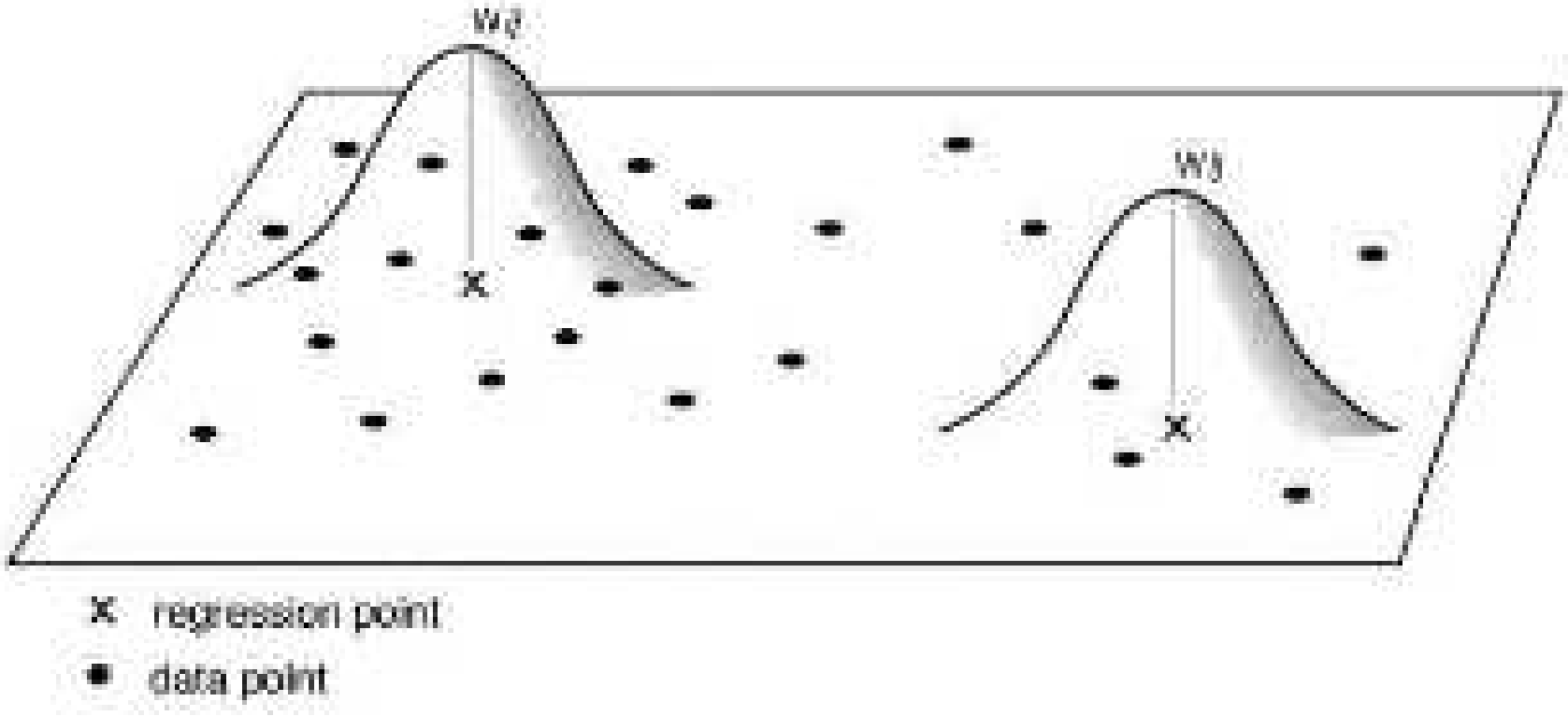
GWR

- Decide how you are going to weight your nearby locations.
  - Fixed bandwidth
  - Variable bandwidth
  - User-defined bandwidth
- Bandwidth specifies shape of weights curve.

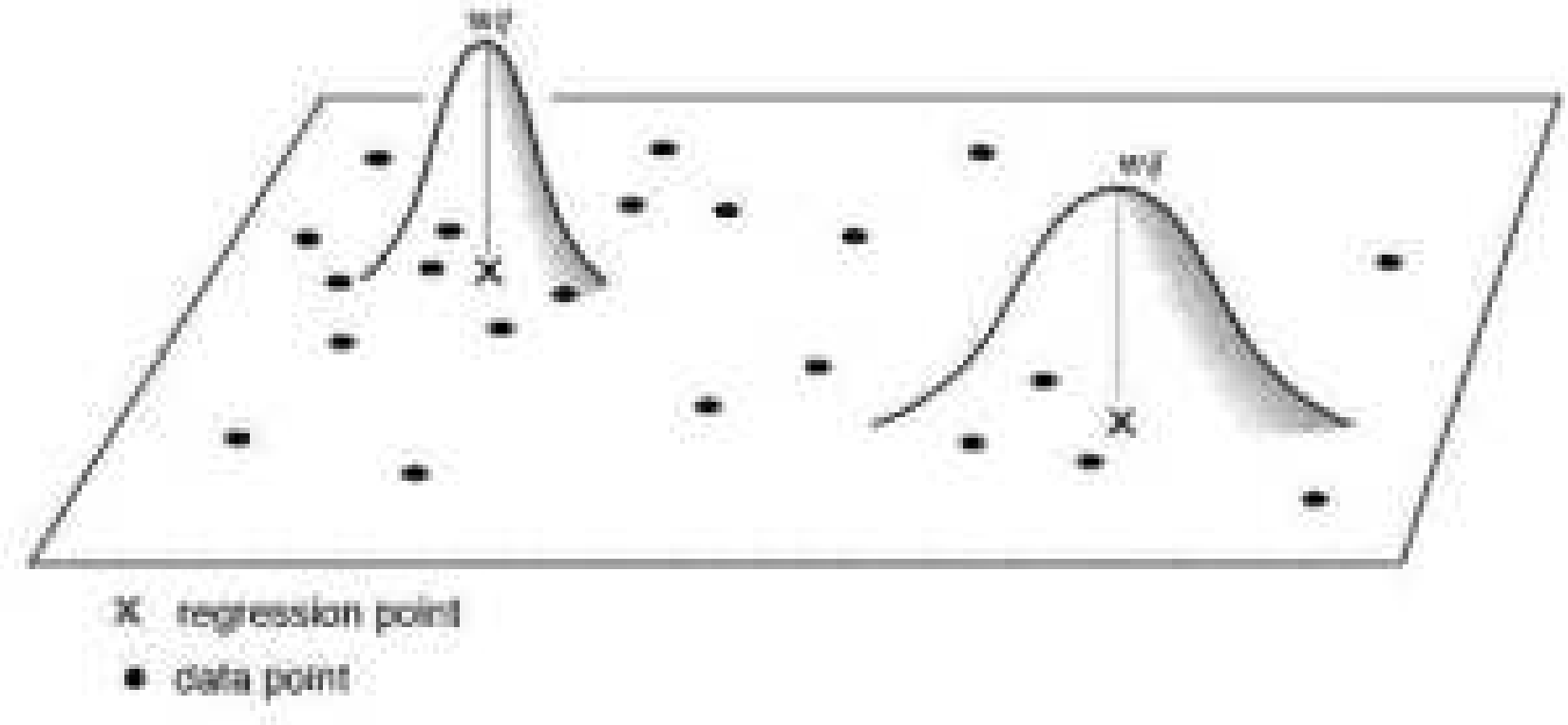
GWR

- Decide how you are going to weight your nearby locations.
  - Fixed bandwidth
  - Variable bandwidth
  - User-defined bandwidth
- Bandwidth specifies shape of weights curve.
- Whether we will define our bandwidth based on distance (fixed) or number of neighbors (adaptive).

Fixed Bandwidth



Adaptive Bandwidth



GWR

- R package : spgwr (dependent packages: sp, maptools)

GWR

- R package : spgwr (dependent packages: sp, maptools)
- data(columbus)

GWR

- R package : spgwr (dependent packages: sp, maptools)
- data(columbus)
- 5 variables
  - crime - recorded crime per inhabitant
  - income - average income values
  - housing - average housing costs
  - x - latitude
  - y - longitude

GWR

- R package : spgwr (dependent packages: sp, maptools)
- data(columbus)
- 5 variables
  - crime - recorded crime per inhabitant
  - income - average income values
  - housing - average housing costs
  - x - latitude
  - y - longitude
- Relationship between crime and income & housing.

GWR

- Need to find a bandwidth for a given geographically weighted regression by optimizing a selected function.

GWR

- Need to find a bandwidth for a given geographically weighted regression by optimizing a selected function.
- gwr.sel(formula, data, coords, adapt=FALSE, ...)

GWR

- Need to find a bandwidth for a given geographically weighted regression by optimizing a selected function.
- gwr.sel(formula, data, coords, adapt=FALSE, ...)
  - formula, data, coords.

GWR

- Need to find a bandwidth for a given geographically weighted regression by optimizing a selected function.
- gwr.sel(formula, data, coords, adapt=FALSE, ...)
  - formula, data, coords.
  - adapt - either TRUE: find the proportion between 0 and 1 of observations to include in weighting scheme (k-nearest neighbors), or FALSE find global bandwidth.

GWR

- Need to find a bandwidth for a given geographically weighted regression by optimizing a selected function.
- gwr.sel(formula, data, coords, adapt=FALSE, ...)
  - formula, data, coords.
  - adapt - either TRUE: find the proportion between 0 and 1 of observations to include in weighting scheme (k-nearest neighbors), or FALSE find global bandwidth.
- Returns the cross-validation bandwidth.

GWR

- Fit GWR model.

GWR

- Fit GWR model.
- gwr(formula, data, coords, bandwidth, ...)

GWR

- Fit GWR model.
- gwr(formula, data, coords, bandwidth, ...)
  - formula, data, coords.

GWR

- Fit GWR model.
- gwr(formula, data, coords, bandwidth, ...)
  - formula, data, coords.
  - bandwidth - bandwidth used in the weighting function.

GWR

- Fit GWR model.
- gwr(formula, data, coords, bandwidth, ...)
  - formula, data, coords.
  - bandwidth - bandwidth used in the weighting function.
  - An example.

QR & GWQR

- Quantile regression is also a method that could be applied to spatial data.

QR & GWQR

- Quantile regression is also a method that could be applied to spatial data.
- Another regression method called Geographically Weighted Quantile Regression (GWQR), as well as the corresponding R package are under development.

QR & GWQR

- Quantile regression is also a method that could be applied to spatial data.
- Another regression method called Geographically Weighted Quantile Regression (GWQR), as well as the corresponding R package are under development.
- By integrating the features of QR into GWR, the GWQR permits researchers to investigate the spatial non-stationary across both space and the whole distribution of a dependent variable.

Discussion

- We can find lots of boundaries on line. If it's a .shp file, we can easily read it into R and use it now.

Discussion

- We can find lots of boundaries on line. If it's a .shp file, we can easily read it into R and use it now.
- However, the biggest assumption (but it should always followed) is, those blocks are following certain order.

Discussion

- We can find lots of boundaries on line. If it's a .shp file, we can easily read it into R and use it now.
- However, the biggest assumption (but it should always followed) is, those blocks are following certain order.
- There are also lots of color platter available on line.

Discussion

- We can find lots of boundaries on line. If it's a .shp file, we can easily read it into R and use it now.
- However, the biggest assumption (but it should always followed) is, those blocks are following certain order.
- There are also lots of color platter available on line.
- GWR, QR, or GWQR should be applied for spatial data.

Discussion

- We can find lots of boundaries on line. If it's a .shp file, we can easily read it into R and use it now.
- However, the biggest assumption (but it should always followed) is, those blocks are following certain order.
- There are also lots of color platter available on line.
- GWR, QR, or GWQR should be applied for spatial data.
- Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. WILEY (2002). $120 !!!

Discussion

- We can find lots of boundaries on line. If it's a .shp file, we can easily read it into R and use it now.
- However, the biggest assumption (but it should always followed) is, those blocks are following certain order.
- There are also lots of color platter available on line.
- GWR, QR, or GWQR should be applied for spatial data.
- Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. WILEY (2002). $120 !!!
- Thank you !! & Questions ??