Local statistics: Geographically-weighted regression

GIS 5923 Spatial Statistics

COLLEGE OF ATMOSPHERIC AND GEOGRAPHIC SCIENCES
DEPARTMENT OF GEOGRAPHY
AND ENVIRONMENTAL SUSTAINABILITY
The UNIVERSITY of OKLAHOMA

Local Statistics

- A local statistic is any descriptive statistic associated with a spatial data set whose value varies from place to place.
- In the broadest sense, any spatial data set is a collection of local statistics, since the recorded attribute values are different at each location.
- But, a "local statistic" usually is one that is derived by considering a subset of the data local to or nearby the spatial location where it is being calculated. Example: the localized mean

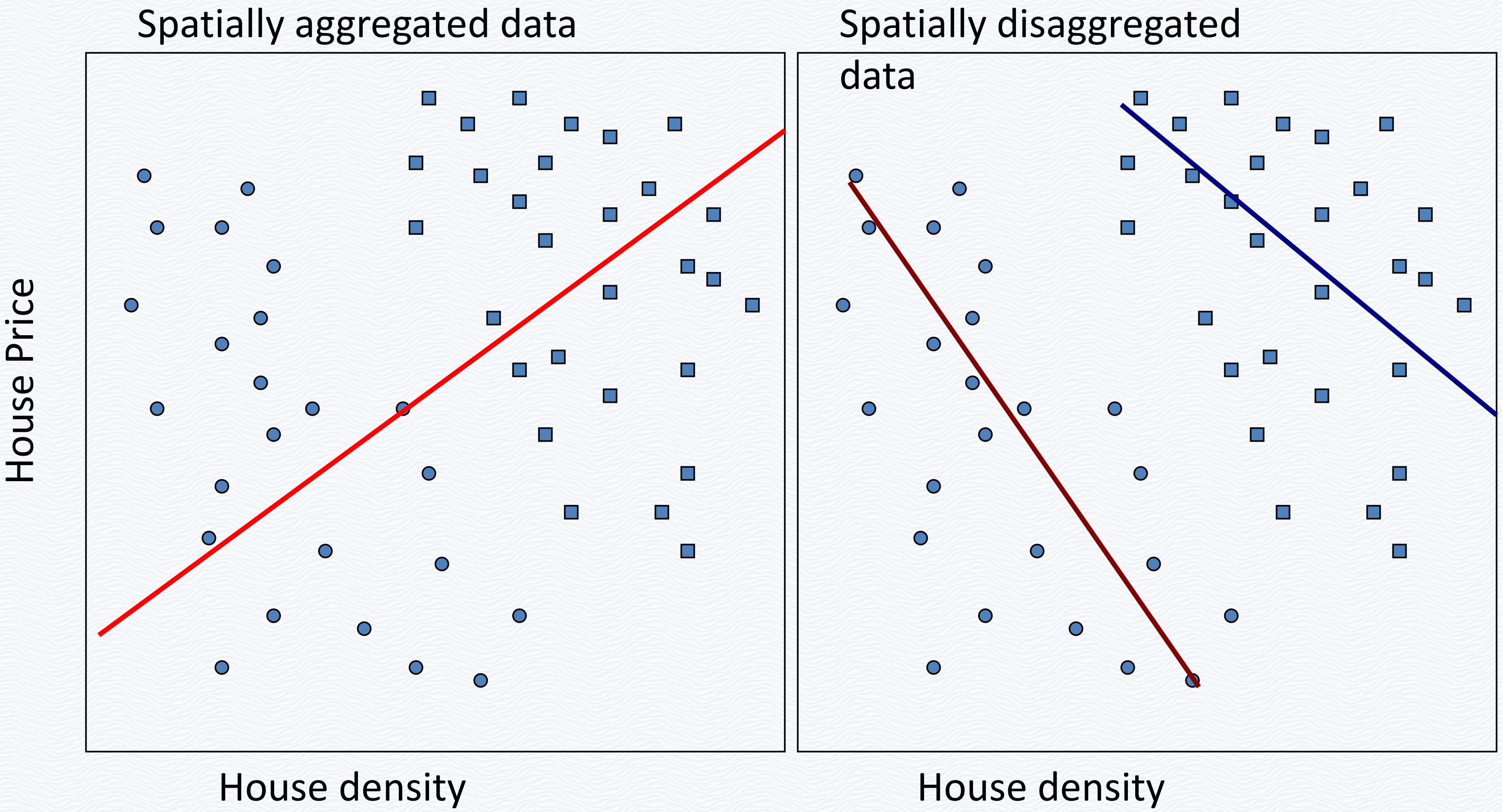
Two topics in Local Statistics

- We will cover two general topics:

- In this set of slides, we will focus on geographically-weighted regression: regression models in which we allow the coefficients to vary spatially

- In the next set of slides, we will cover Local Indicators of Spatial Association (LISAs): used to identify hotspots and outliers

Simpson’s paradox



COLLEGE OF ATMOSPHERIC AND GEOGRAPHIC SCIENCES
DEPARTMENT OF GEOGRAPHY
AND ENVIRONMENTAL SUSTAINABILITY
The UNIVERSITY of OKLAHOMA

Source: K. de Beurs

Some definitions

- Spatial nonstationarity exists when the relationship between two covariates changes across the study region.
  - Example: the relationship between housing price and square footage may differ between urban and rural regions

- Global models (i.e., fitting a single regression model to the entire study region) assume that the relationships between covariates and response are stationary.

- Local models (e.g., Geographically-weighted regression) account for spatial nonstationarity by allowing model coefficients to vary spatially

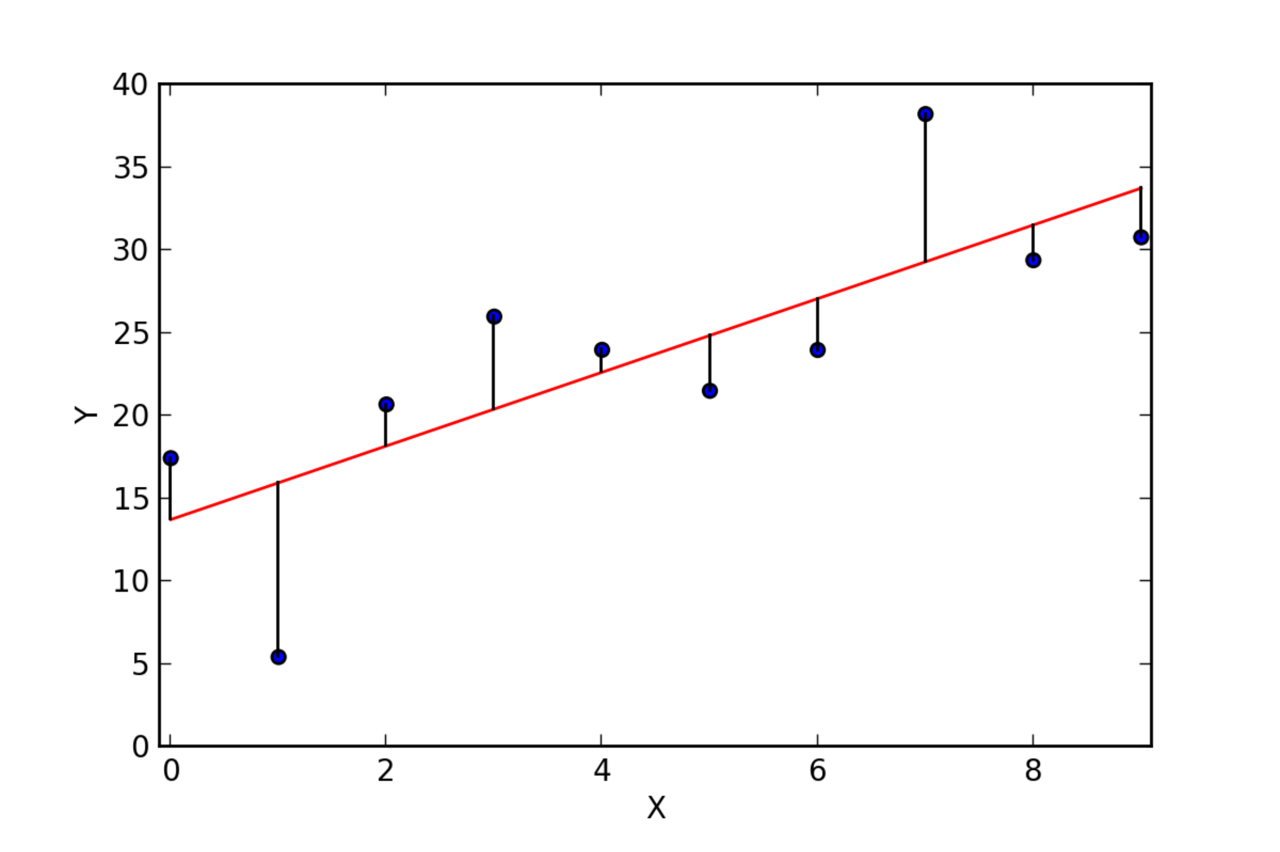
When to consider a GWR model

- Before fitting a GWR model or some other spatial regression model, the first step is to fit an ordinary least squares model and assess the fit

- An OLS model is said to be mis-specified if there is strong spatial autocorrelation in the residuals… This is good justification for fitting a GWR model. If the residuals are spatially random, there is not good justification for fitting a GWR model.

- Let's look at some examples…

Regression model residuals

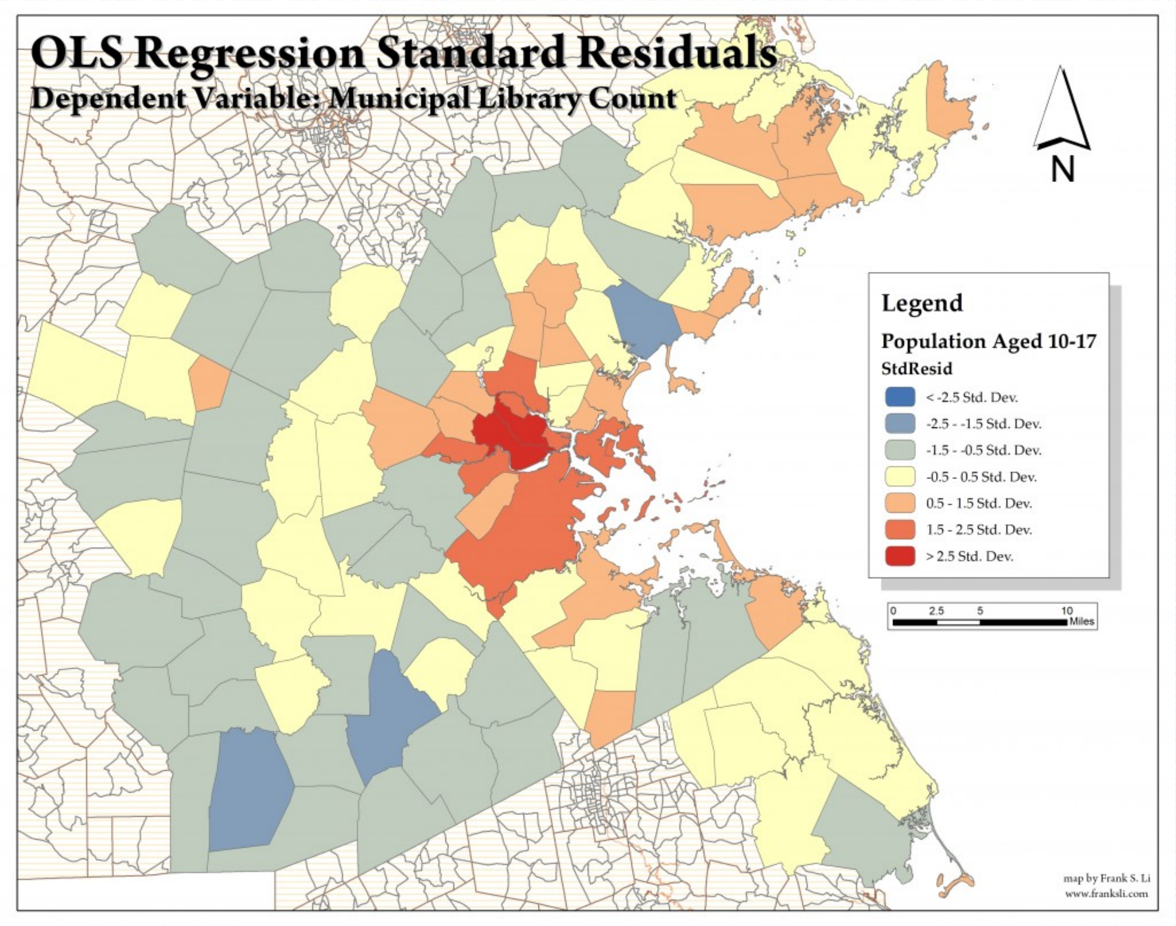


Residuals are the difference between observed and estimated values

Positive residuals occur when the point lies above the regression line; i.e., the model **underpredicts** Y

Negative residuals occur when the point lies below the regression line; i.e., the model **overpredicts** Y

Is there stationarity in the relationship between # of kids and library density in Boston?

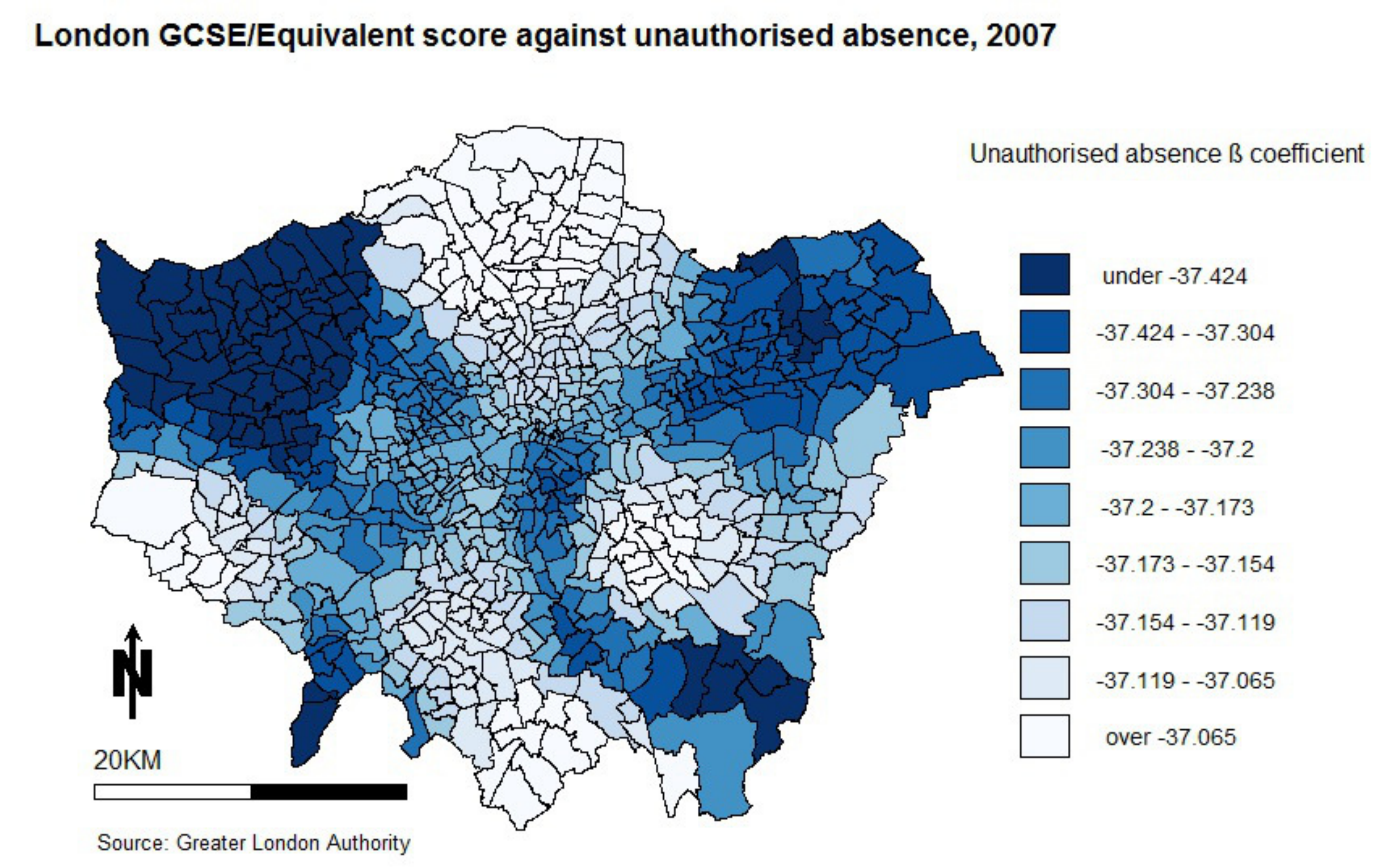


COLLEGE OF ATMOSPHERIC AND GEOGRAPHIC SCIENCES
DEPARTMENT OF GEOGRAPHY
AND ENVIRONMENTAL SUSTAINABILITY
The UNIVERSITY of OKLAHOMA

Geographically Weighted Regression

- In of the preceding example, we see spatial autocorrelation in the residuals from the ordinary least squares model. This is evidence that the OLS model is mis-specified, and good justification for fitting a GWR model.

- In Geographically Weighted Regression, we allow the regression coefficients to vary spatially:



COLLEGE OF ATMOSPHERIC AND GEOGRAPHIC SCIENCES
DEPARTMENT OF GEOGRAPHY
AND ENVIRONMENTAL SUSTAINABILITY
The UNIVERSITY of OKLAHOMA

A review of the OLS model

- In OLS regression, we have a data set $\{y_i, x_{i1}, \ldots, x_{ip}\}$ with $i = 1,\ldots, n$ spatial units and p regressors or explanatory variables. The OLS model is:

$$y_i = \beta_o + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p} + \varepsilon_i$$

The OLS model in vector form

- We can write the same model in vector form as:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \ldots \\ y_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & \ldots & x_{1,p} \\ & \ldots & & \\ & & \ldots & \\ 1 & x_{n,1} & \ldots & x_{n,p} \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \ldots \\ \beta_p \end{pmatrix}$$

Estimating the coefficients in OLS

- In the OLS model, we estimate the unknown parameters by minimizing the sum of the squared residuals. The estimated value of the unknown parameters is:

$$\beta = (X^T X)^{-1} X^T y$$

Estimating the coefficients in GWR

- In GWR, we estimate n distinct models (one for each of the n spatial units). To estimate the regression coefficients for each spatial unit, we modify the OLS estimator by introducing a set of weights

$$\beta = (X^T G X)^{-1} G X^T y$$

where G is an $n \times n$ diagonal matrix for each spatial unit in the data set. The diagonal elements give the weight we wish to associate with each observation.

How do we choose G?

- The weighting scheme in the matrix G is an essential feature of the GWR model and depends on two choices:
  - 1) What is the kernel that describes the shape of the relationship between two spatial units?
  - 2) What is the bandwidth of the kernel, i.e., how quickly does influence decay with distance?

- The choice of the appropriate bandwidth is an essential element of fitting a GWR model

Some very simple types of G

- The two simplest rules for specifying G for a spatial unit *j* are to choose binary weights based on some distance threshold:

$$g_{i,i} = \begin{cases} 1 & \text{if } d_{ij} < d^* \\ 0 & \text{else} \end{cases}$$

- Or to choose binary weights based on a k-nearest neighbors approach:

$$g_{i,i} = \begin{cases} 1 & \text{if j is one of the k-nearest neighbors of i} \\ 0 & \text{else} \end{cases}$$

COLLEGE OF ATMOSPHERIC AND GEOGRAPHIC SCIENCES
DEPARTMENT OF GEOGRAPHY
AND ENVIRONMENTAL SUSTAINABILITY
The UNIVERSITY of OKLAHOMA

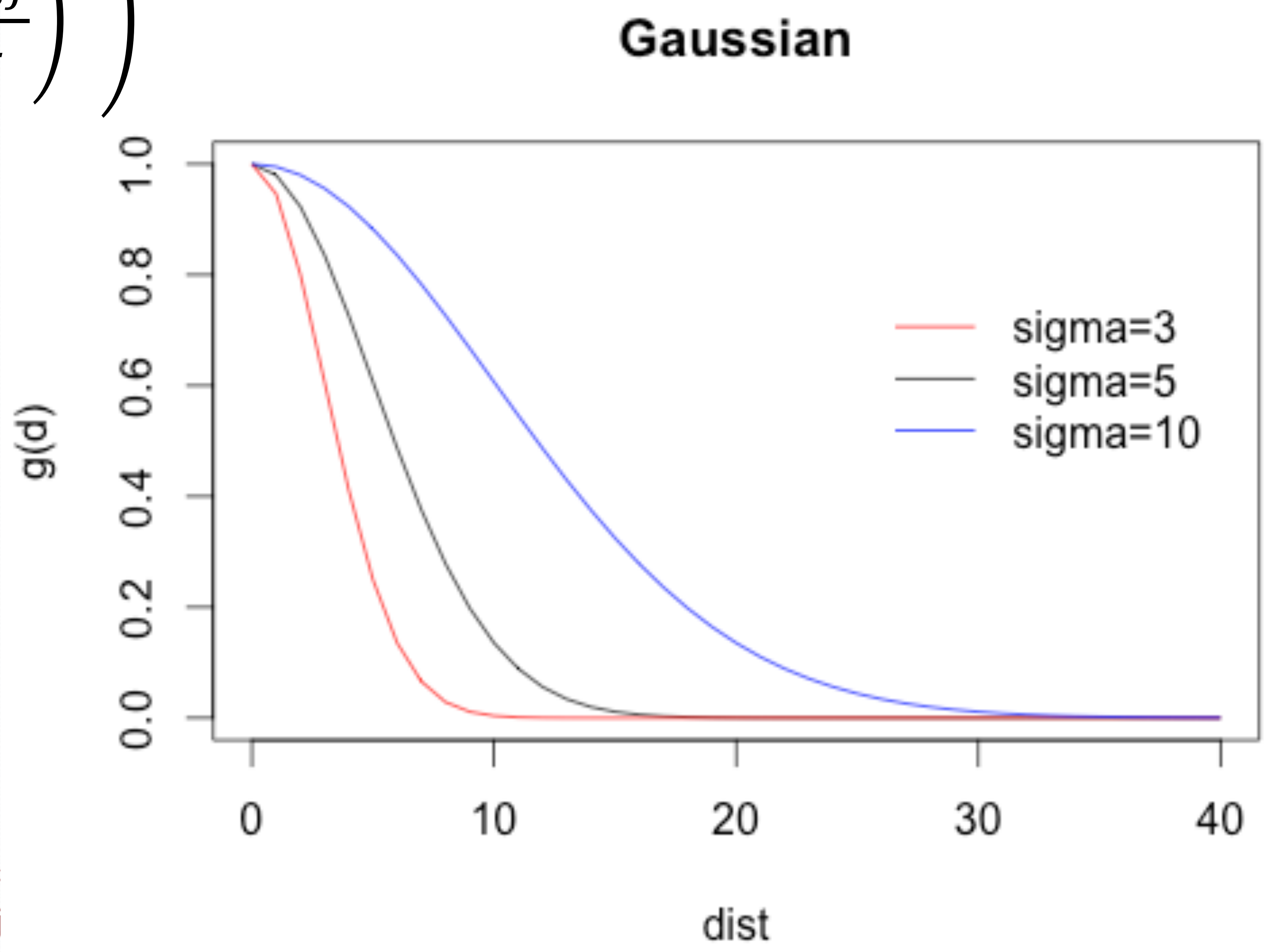
More complex (but more common) G

- A more common approach is to specify a kernel function that weights nearby locations more strongly than distant locations. The most common is the Gaussian kernel. For each spatial unit *j*

$$g_{i,i} = exp\left(-\frac{1}{2}\left(\frac{d_{i,j}}{\sigma}\right)^2\right)$$

- Other common kernels are the bi-square and tri-cube; they just specify different relationships between distance and weight.
- All of these kernels have a parameter σ called the bandwidth through which we can control the range of observations in the samples.

$$g_{i,i} = exp\left(-\frac{1}{2}\left(\frac{d_{i,j}}{\sigma}\right)^2\right)$$



COLLEGE OF ATMOSPHERIC AND GEOGRAPHIC
DEPARTMENT OF GEOGR
AND ENVIRONMENTAL SU
The UNIVERSITY of OKLAHOMA

How to choose the bandwidth

- There are two approaches the specifying the bandwidth: a priori and via the data in a process called calibration.

- If the bandwidth is determined a priori, the research chooses the bandwidth based on knowledge of the data set or research problem

- If the bandwidth is determined via calibration, an iterative search process identifies the bandwidth that minimizes the error of the prediction model.