Department of Geography

GEOG0114: PRINCIPLES OF SPATIAL ANALYSIS

WEEK 8: SPATIAL MODELS (PART 2)

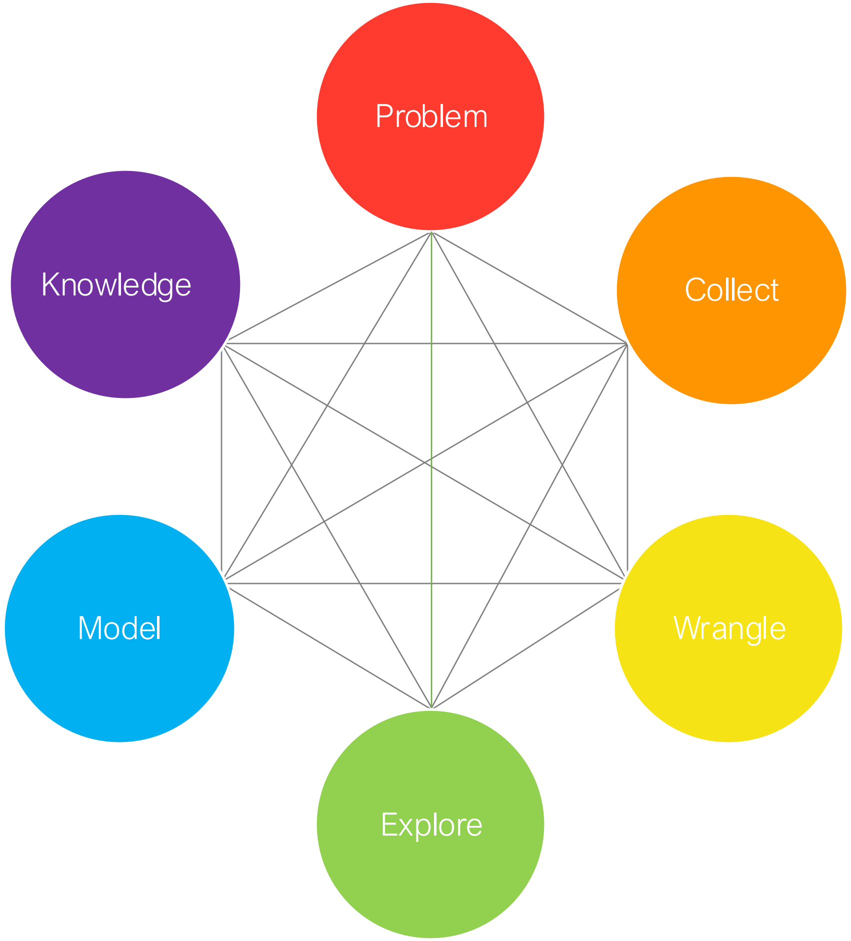
Dr Anwar Musah (a.musah@ucl.ac.uk)
Lecturer in Social and Geographic Data Science
UCL Geography

Contents

1. Introduction to Geographically Weighted Regression Modelling

2. Extending the standard linear regression to GWR
   - Using GWRs to estimate the local (not global) relationships between a dependent and independent variable
   - Determination of whether local relationships are statistically significant or not
   - Model performance through Local R-squared

3. Methodology for statistical analysis and interpretation



Recap

Spatial Models

Week 7

Spatial Lag and Error Models

Week 8

Geographically Weighted Regression

Week 9

Spatial Risk Models

In Week 7, we learnt special types of spatial model which accounts for spatial configuration of areal data

Recap

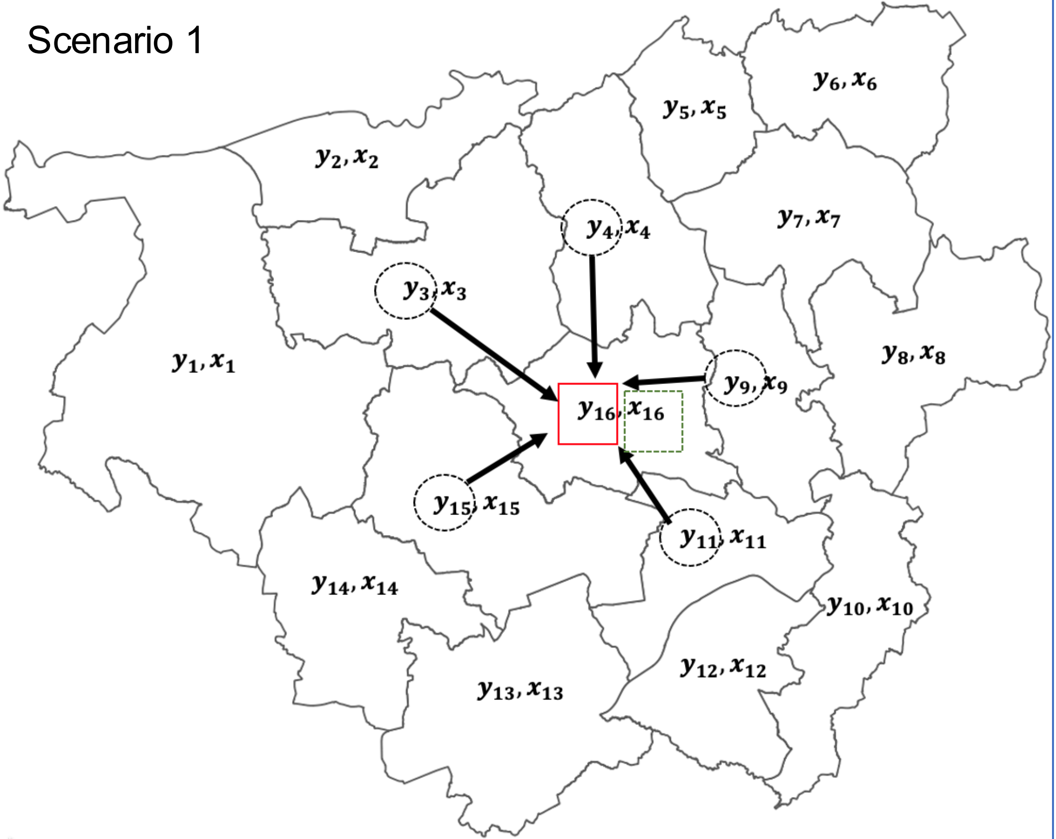
Multivariable Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

1. Spatial Lag Model (lagged on the dependent variable)

$$y = \rho W Y + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Scenario 1

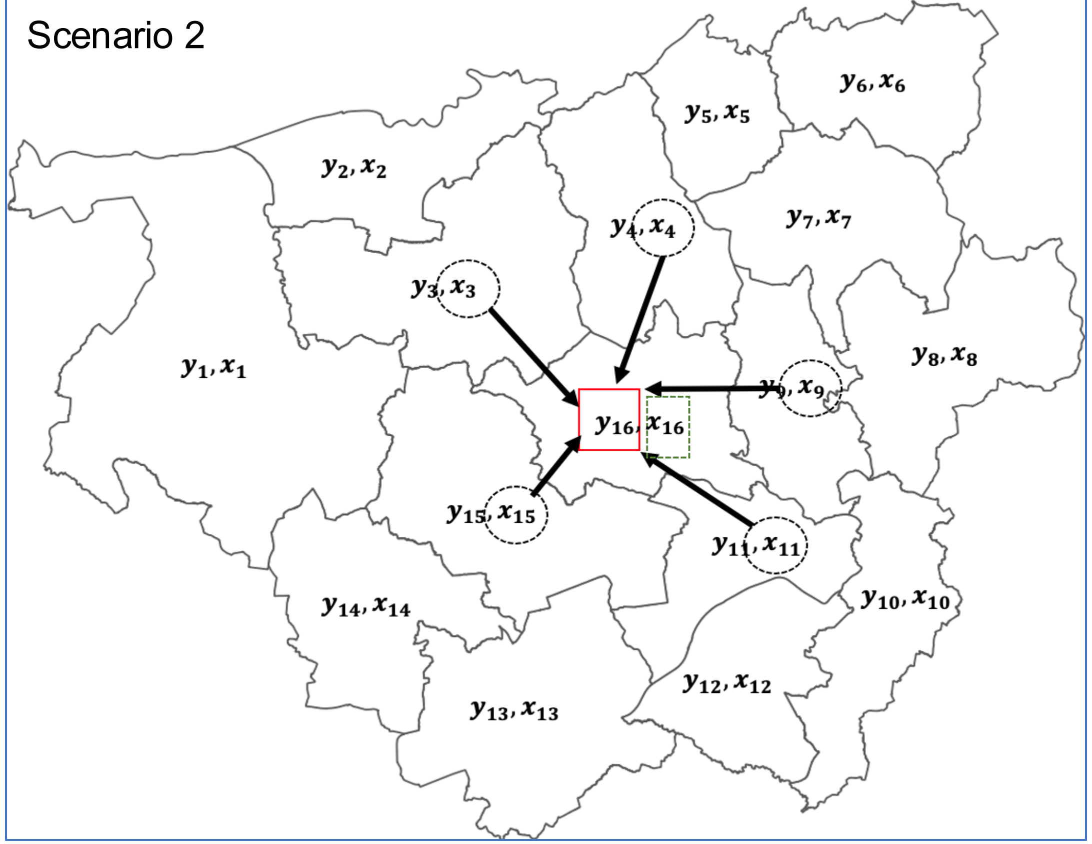


- W is a spatial weights matrix (contiguity – based)
- Y in the right hand of the equation represents the observed outcome from other areas neighbouring that influences what we're trying to predict
- ρ "Rho" is the degree of how our predicted outcome are influenced by its neighbouring Y measures.

2. Spatial Lag Model (lagged on the independent variable)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \theta W X + \varepsilon$$

Scenario 2



- W is a spatial weights matrix (contiguity – based)
- X in the right hand of the equation represents the observed values from the independent variable in other areas neighbouring that influences what we're trying to predict
- θ "theta" is the degree of how our predicted outcome are influenced by its neighbouring X measures.

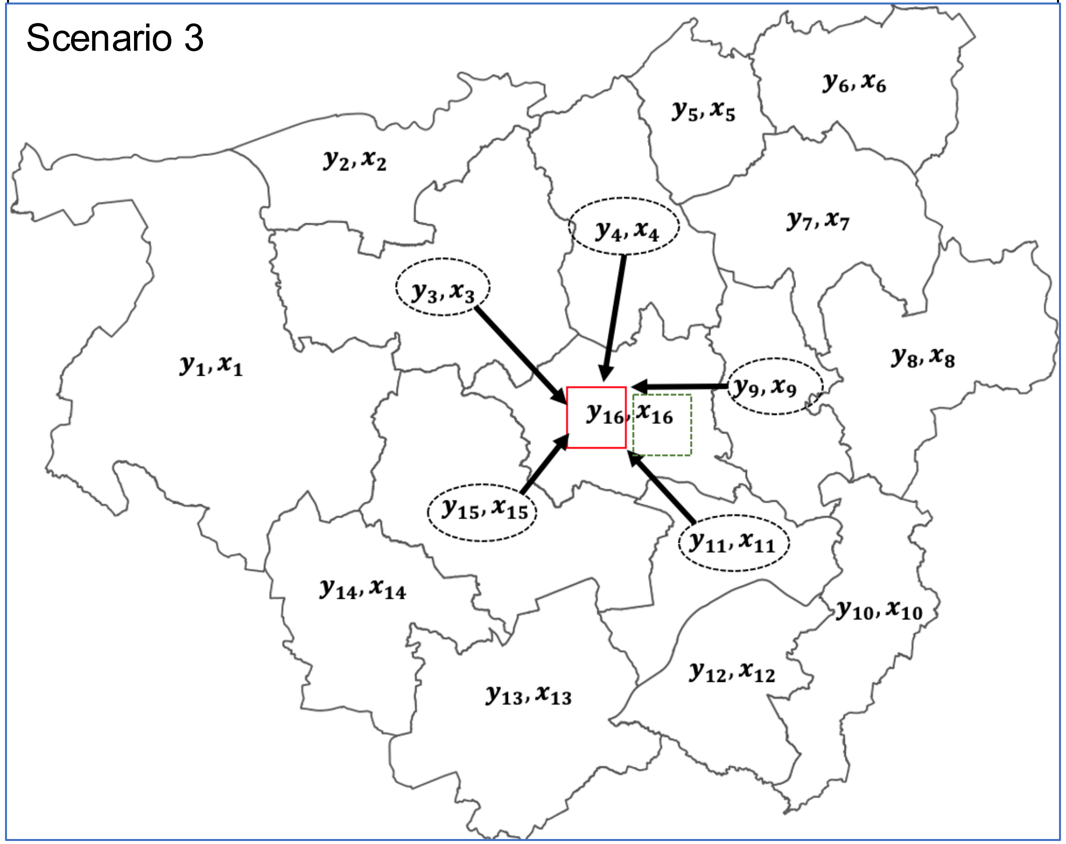
Recap

Multivariable Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

3. Spatial Lag Model (lagged on both the dependent and independent variable)

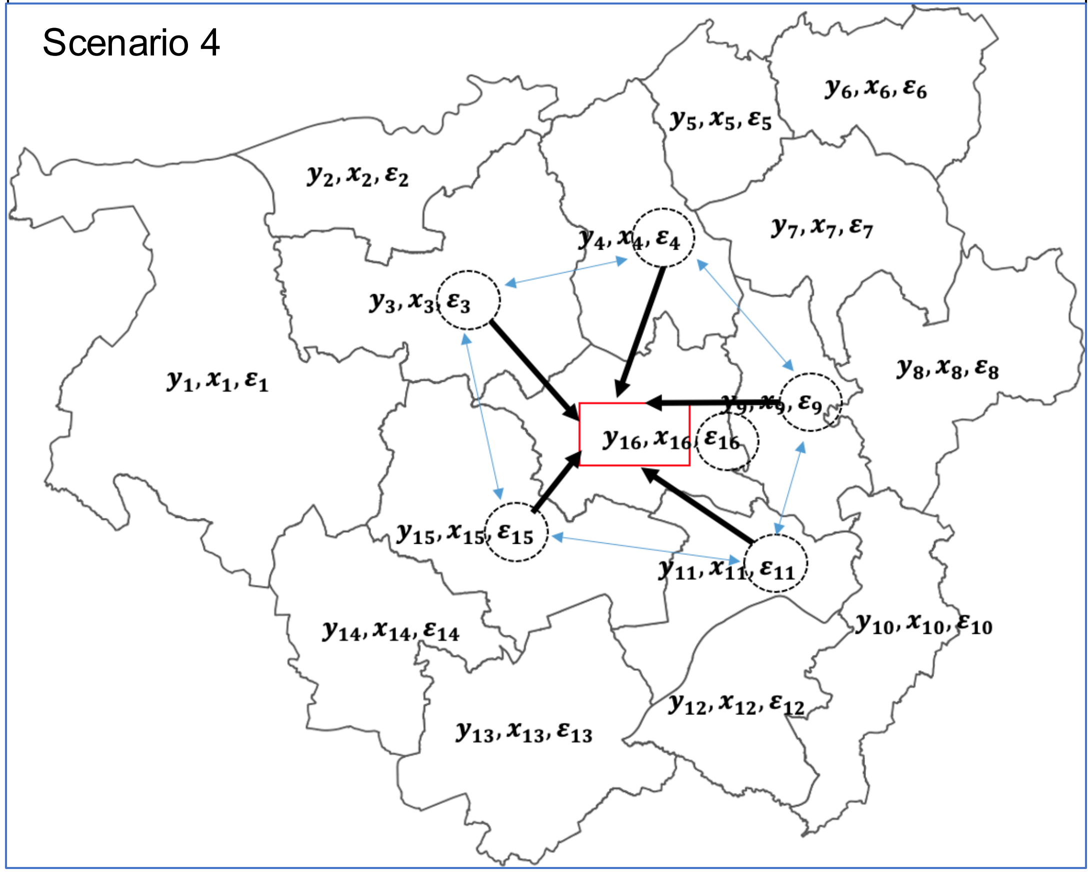
$$y = \rho WY \ + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \theta WX + \varepsilon$$



- Refer to slide number 3 to see the meanings of each parameters

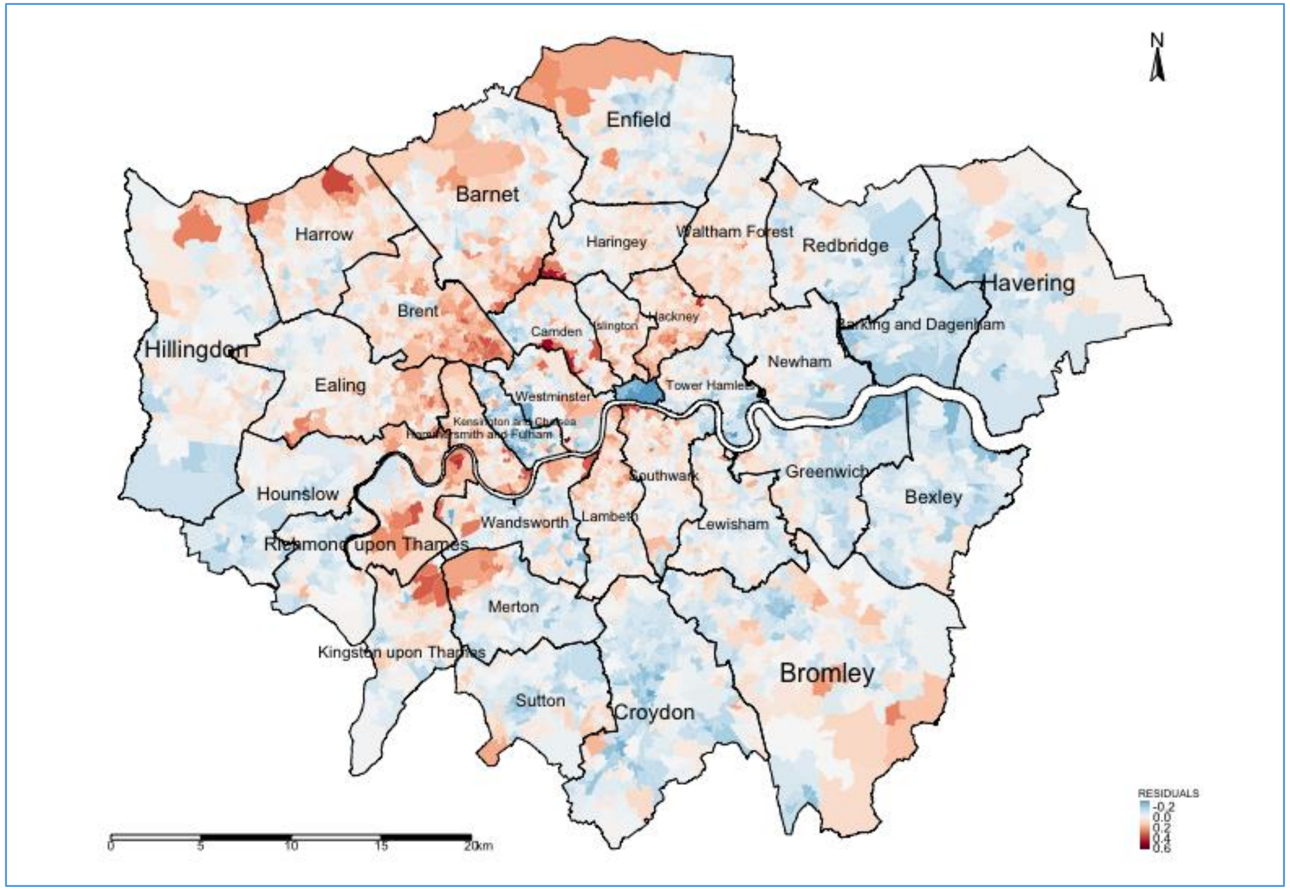
4. Spatial Error Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \lambda W\mathbf{u} + \varepsilon$$


Scenario 4
y_6, x_6, ε_6
y_5, x_5, ε_5
y_2, x_2, ε_2
y_4, x_4, ε_4
y_7, x_7, ε_7
y_3, x_3, ε_3
y_1, x_1, ε_1
y_8, x_8, ε_8
y_9, x_9, ε_9
$y_{16}, x_{16}, \varepsilon_{16}$
$y_{15}, x_{15}, \varepsilon_{15}$
$y_{11}, x_{11}, \varepsilon_{11}$
$y_{14}, x_{14}, \varepsilon_{14}$
$y_{10}, x_{10}, \varepsilon_{10}$
$y_{12}, x_{12}, \varepsilon_{12}$
$y_{13}, x_{13}, \varepsilon_{13}$

- u are the correlated spatial error terms
- λ “lambda” estimated coefficient for the product W and u.

Recap



Examining the residuals for determining spatial patterning and evidence of spatial autocorrelation. Moran's I test was 0.475 ($p < 0.001$) meaning that the residuals are clustered.

Broadly, there's an over-estimation in the house-price and spatial aspects needs to be accounted for.

Modelled results from Linear and Spatial Lag (Y) regression model

Variable(s)	Linear Model	Lag (Y) Model (Total effects)
log(Income)	2.036*	2.217*
log(Deprivation)	0.136*	0.079*
log(PTAL)	0.031*	0.019*
AIC	**-8510.8**	**-9863.3**
R^2	0.7889 (78.89%)	N/A

- While we have accounted for spatial configuration using the spatial model, as well as accounted for spatial autocorrelation, we were able to determine the **Global** relationships between dependent and independent variables.
- What about if we want to investigate further patterns but a much **local-level**?
- LM, and any of the spatial Lag and error models cannot solve this problem

Spatial Models

Week 7

Spatial Lag and Error Models

Week 8

Geographically Weighted Regression

Week 9

Spatial Risk Models

What is a Geographically Weighted Regression

Definition of Geographically Weighted Regression (GWR) model:

GWR is a statistical model which can indicate where non-stationarity may take place across space; it can be used to identify how **locally** weighted regression coefficients may vary across the study area (unlike its counterpart i.e., the **Spatial Lagged and/or Error Models** which provides **global coefficients**)

We use GWRs to:

1) Determine **area-specific relationships** or local **associations** between a specified **outcome** (i.e., **dependent variable**) with one or more **predictors** (i.e., **independent variable(s)**)

2) Find out whether those area-specific relationship or local associations are statistically significant across geographic space.

GWRs fall under the family of linear regression models. Recall last week the various model types and families?

Here is a board overview:

Distribution of dependent variable	Suitable Model
Continuous measures: e.g., average income in postcode (£); concentrations of ambient particular matter (PM2.5); Normalised Vegetative Difference Index (NDVI) etc.,	**Linear regression**
Binary measures (1 = "present" or 0 = "absent"): e.g., Person's voting for a candidate, Lung cancer risk, house infested with rodents etc.,	**Logistic Regression**
Binomial measure (or proportion): e.g., prevalence of houses in a postcode infested with rodents, percentage of people in a village infected with intestinal parasitic worms, prevalence of household on a street segment victimised by crime etc.,	**Logistic Regression**
Counts or discrete measures: e.g., number of reported burglaries on a street segment, number of riots in a county etc.,	**Poisson Regression**
Time-to-event binary measures: e.g., Lung cancer risk due to chronic exposure to environmental levels of indoor radon. Risk of landslide and time dependence of surface erosion etc.,	**Survival Analysis with Cox regression**

Multivariable Linear Regression Model

$$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} + \cdots + \boldsymbol{\varepsilon}$$

Variables

- $\boldsymbol{y}$ is the dependent variable
- $\boldsymbol{x_1, x_2, x_3, \ldots, x_k}$ are the independent variables

Parameters

- $\boldsymbol{\beta_0}$ is the intercept
- $\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}, \ldots, \boldsymbol{\beta_k}$ are the slopes (or coefficients) for the corresponding variables $\boldsymbol{x_1, x_2, x_3, \ldots, x_k}$
- $\boldsymbol{\varepsilon}$ is the error term

Geographical Weighted Regression Model

$$y_i = \beta_{i,0}(u_i, v_i) + \beta_{i,1}(u_i, v_i)x_{i,1} + \beta_{i,2}(u_i, v_i)x_{i,2} + \cdots \beta_{i,k}(u_i, v_i)x_{i,k} + \varepsilon_i$$

Variables

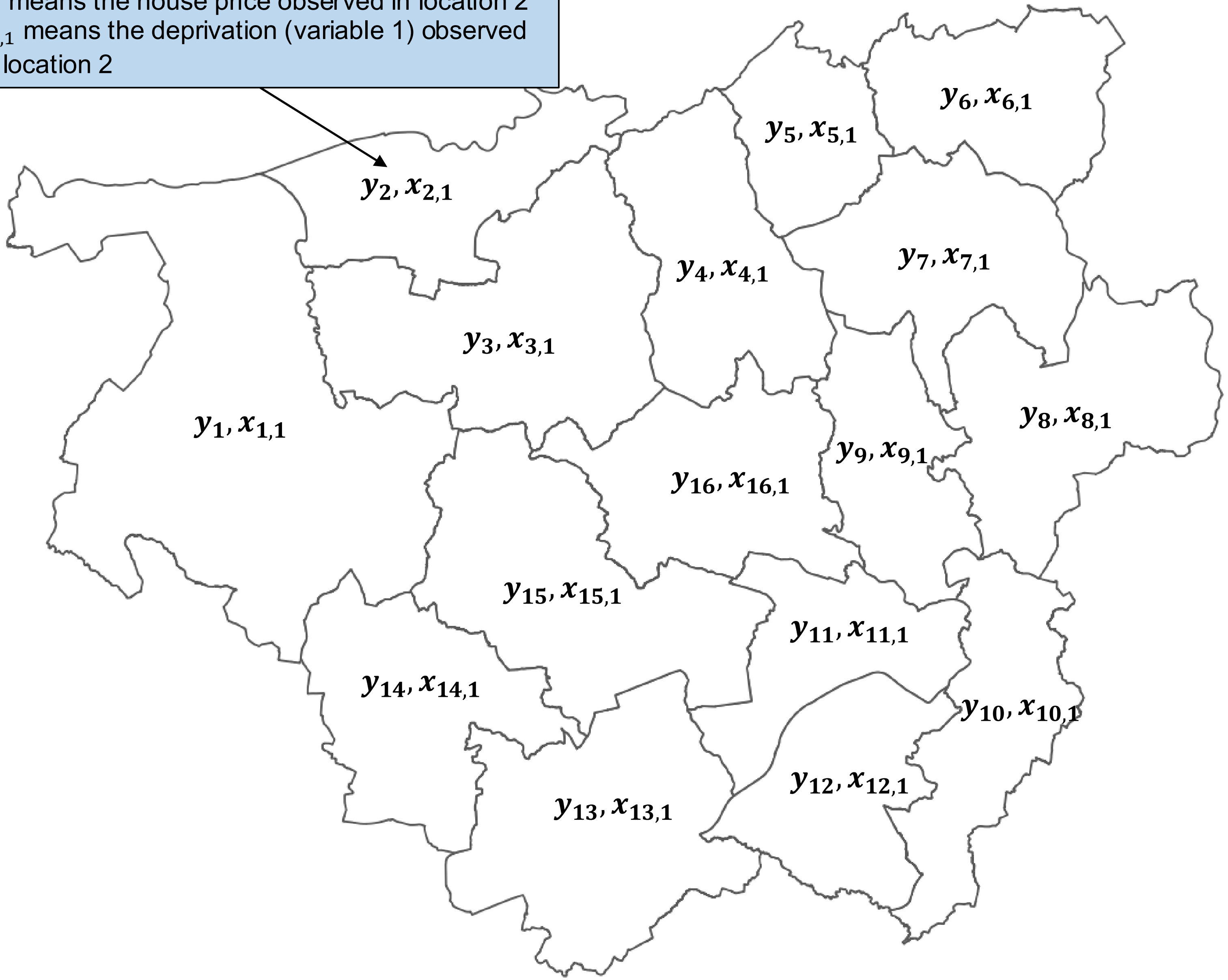
- $\boldsymbol{y_i}$ is the dependent variable indexed at observation (or location) i
- x_{i1}, x_{i2} and so to x_{ik} are the k-number of independent variables indexed at at

Parameters

- $\boldsymbol{\beta_{i0}}(u_i, v_i)$ is the intercept as a function of a geographic location (i.e., coordinates on a grid)
- $\boldsymbol{\beta_{i1}}(u_i, v_i), \boldsymbol{\beta_{i2}}\ (u_i, v_i)$, $\boldsymbol{\beta_{i3}}(u_i, v_i)$**,…,** $\boldsymbol{\beta_{ik}}(u_i, v_i)$ are the slopes (or coefficients) for the corresponding variables x_{i1}, $x_{i2}, x_{i3}, \dots, x_{ik}$ which are function of a geographic location (u_i, v_i)
- $\boldsymbol{\varepsilon}$ is the error term

Suppose we have a hypothetical study area with 16 areas

y_2 means the house price observed in location 2
$x_{2,1}$ means the deprivation (variable 1) observed in location 2



Notes:

Let Y be some dependent variable that is continuous and normally distributed, where there are 16 observation for Y at some i-location (i.e. $\boldsymbol{y_1}, \boldsymbol{y_2} \dots \boldsymbol{y_{16}}$)

- For example: Averaged house price (£)

Let X be some k^th independent variable $x_{i,k}$, (in this case k = 1) where there are 16 locations for X (i.e. $\boldsymbol{x_{1,1}}, \boldsymbol{x_{2,1}} \dots \boldsymbol{x_{16,1}}$)

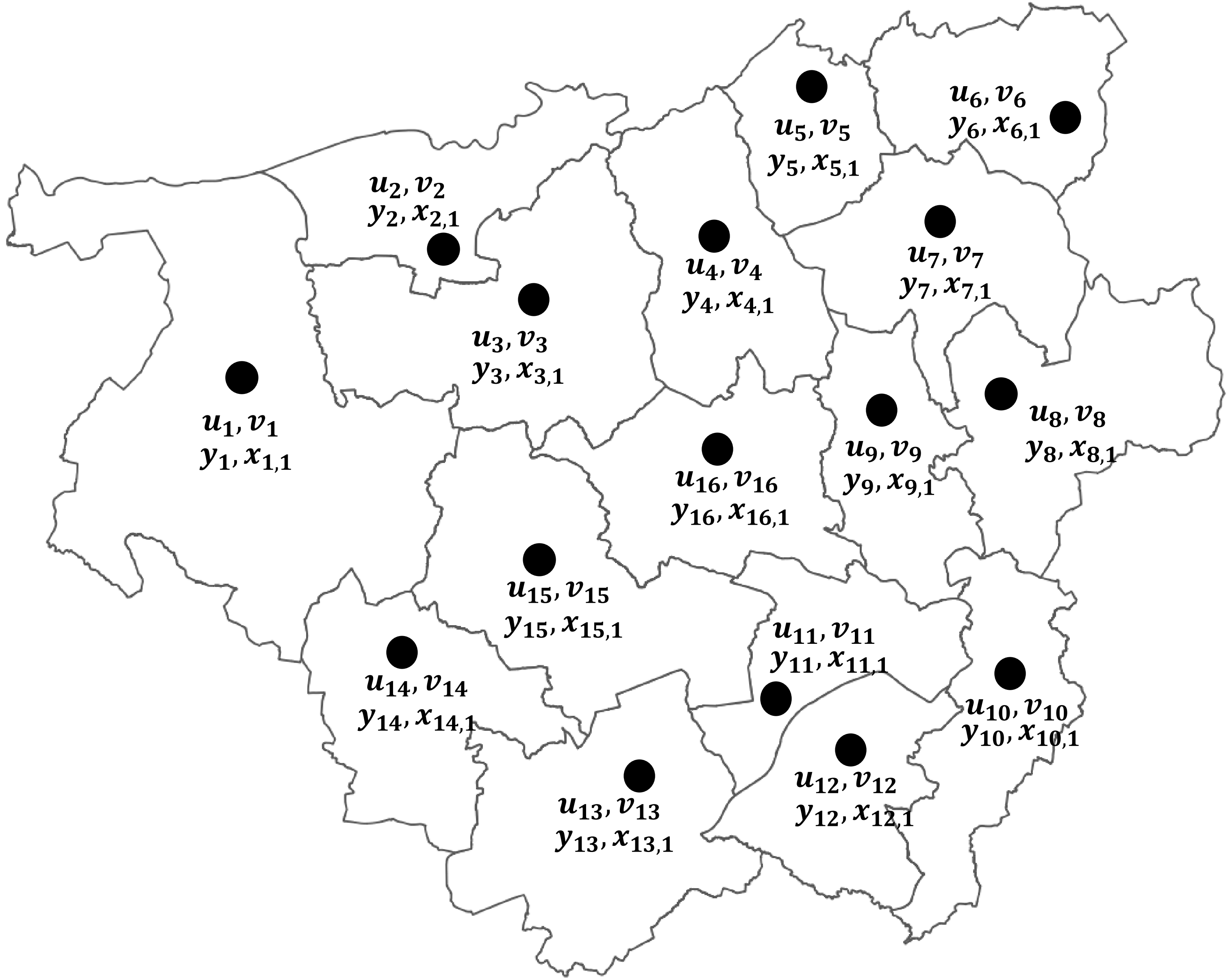
- For example: Socioeconomic deprivation score

Research question: To investigate the **geospatial** impacts of socioeconomic deprivation on house price in each area in this hypothetical study area.

Insufficient to use the typical linear regression model for this context

$$\boldsymbol{y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\varepsilon}$$

Suppose we have a hypothetical study area with 16 areas



Notes:

We want to model the relationship of $y_i, x_{i,1}$ at location (u_i, v_i)

Because $y_i, x_{i,1}$ is calibrated on (u_i, v_i) as a function, we are able to use some model (i.e., GWR) to compute coefficients at each location of i.

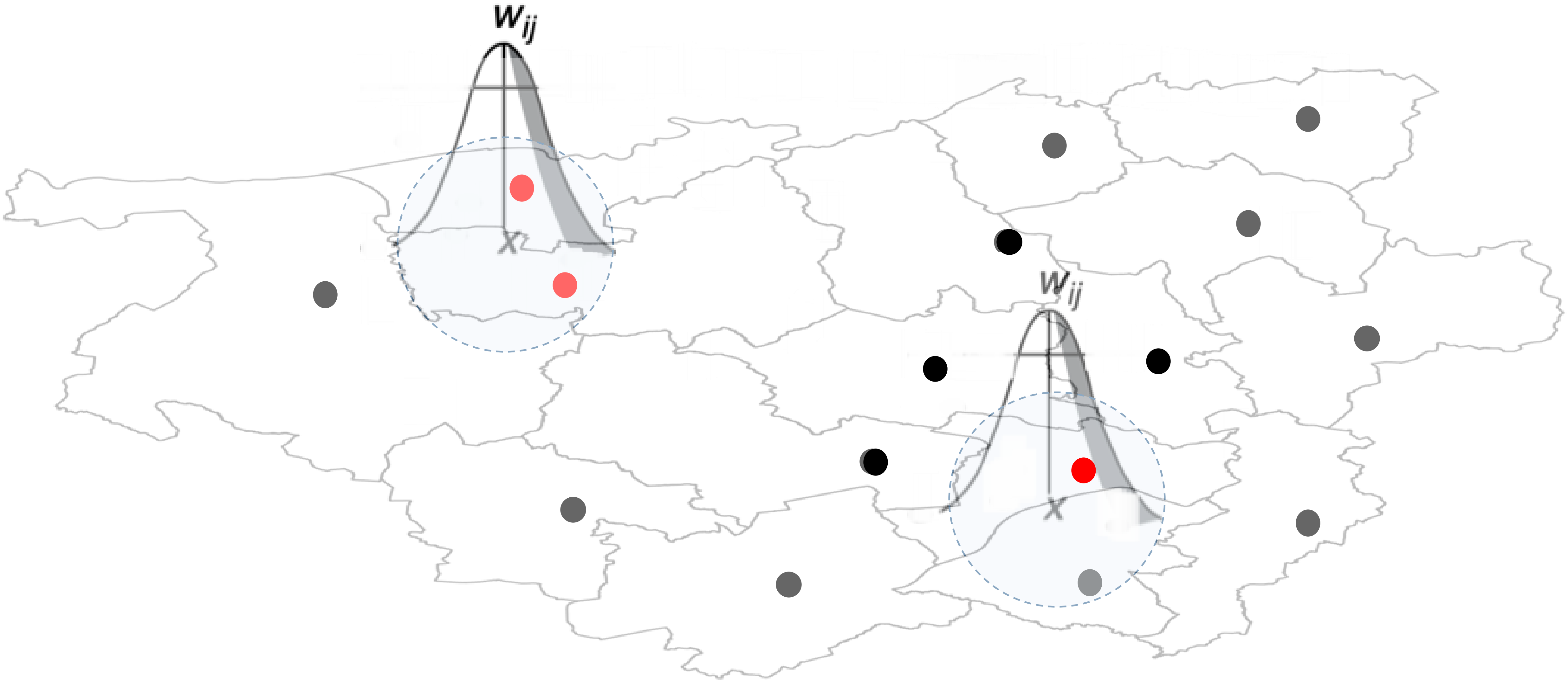
In this example for IMD, these coefficients are represented as $\boldsymbol{\beta_{i,1}}(u_i, v_i)$

The GWRs implicitly use distance-based weights through **spatial kernels** or **bandwidths**. Hence, it relies on points.

Centroids are extracted from areas and used for such analysis.

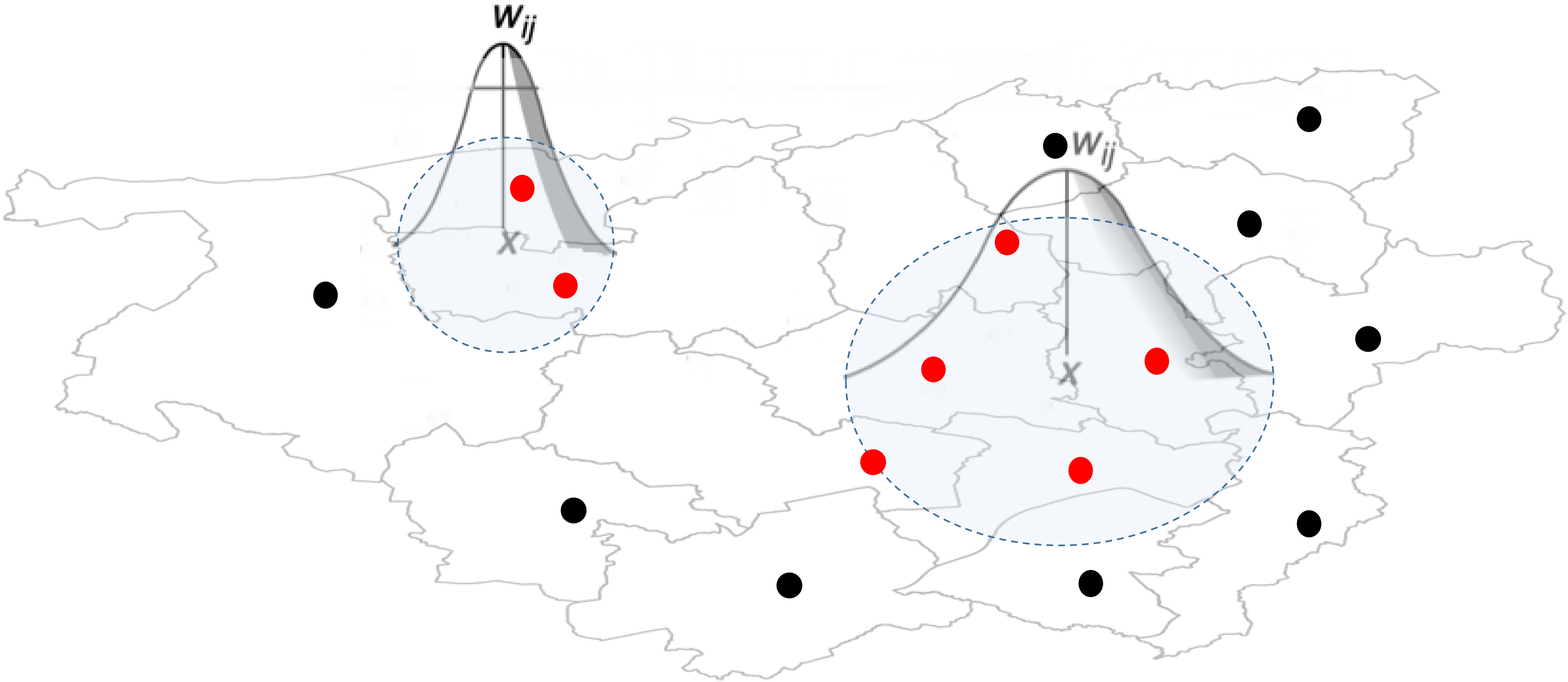
$$\boldsymbol{y_i} = \boldsymbol{\beta_{i,0}}(u_i, v_i) + \boldsymbol{\beta_{i,1}}(u_i, v_i)\boldsymbol{x_{i,1}} + \boldsymbol{\varepsilon_i}$$

GWR with fixed spatial kernel (or bandwidth)



Note: GWRs are distance-based models. It uses bandwidth to consider nearest neighbours when accounting spatial configuration.

GWR with adaptive spatial kernel (or bandwidth)



Use the Adaptive spatial kernel for building your spatial weights! It is much better than using the fixed bandwidth

Workflow for GWR modelling

Modelling process using GWR

When you want to conduct evidence-based analysis with spatial data – especially if the outcome is from a continuous distribution – you might want to follow these steps:

- STEP 1: Carry some descriptive analysis to understand the underlying spatial distribution
- STEP 2: Perform a Linear regression in order to assess the residuals for the model output to determine whether not the assumptions of independence have been violated.
- You can check for multicollinearity among independent variable using the Variance Inflation Factor (VIF < 10)

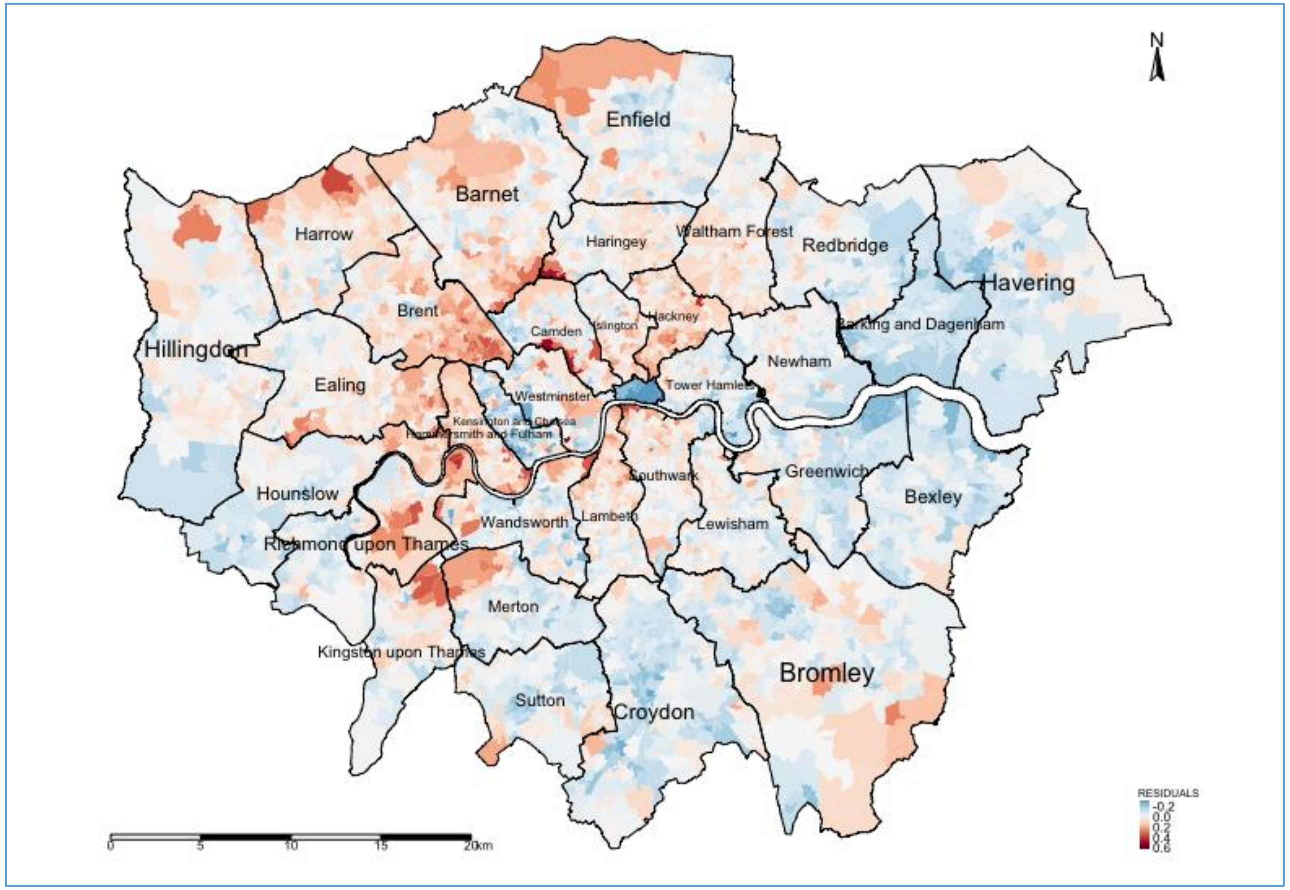
If there's a violation:

- STEP 3: Examine whether there is evidence of spatial dependence in the residuals using Moran's I test.
- If Moran's I test is not significant – STOP! NO EVIDENCE OF SPATIAL DEPENDENCE HENCE NO NEED OF SPATIAL APPROACH

If Moran's I test is significant and positive – then map the residuals accordingly to examine the spatial patterning of the residual.

- STEP 4: Extract the centroids of the areas and use them for computing the kernel bandwidths. [I] Highly recommend to use the adaptive bandwidths, which is flexible than the fixed for estimating the optimal bandwidth.
- STEP 5: The estimated bandwidth is fitted into the GWR model to estimate the following quantities: 1.) Local R-squared, 2.) area-specific coefficients and 3.) standard errors for significance test for each areas.
- STEP 6: Extract the coefficients and desired results and map them accordingly to examine the spatial variation in the relationship between dependent and independent variables.
- STEP 7: Interpretation

Example using the house price data for London



Examining the residuals for determining spatial patterning and evidence of spatial autocorrelation. Moran's I test was 0.475 ($p < 0.001$) meaning that the residuals are clustered.

Broadly, there's an over-estimation in the house-price and spatial aspects needs to be accounted for.

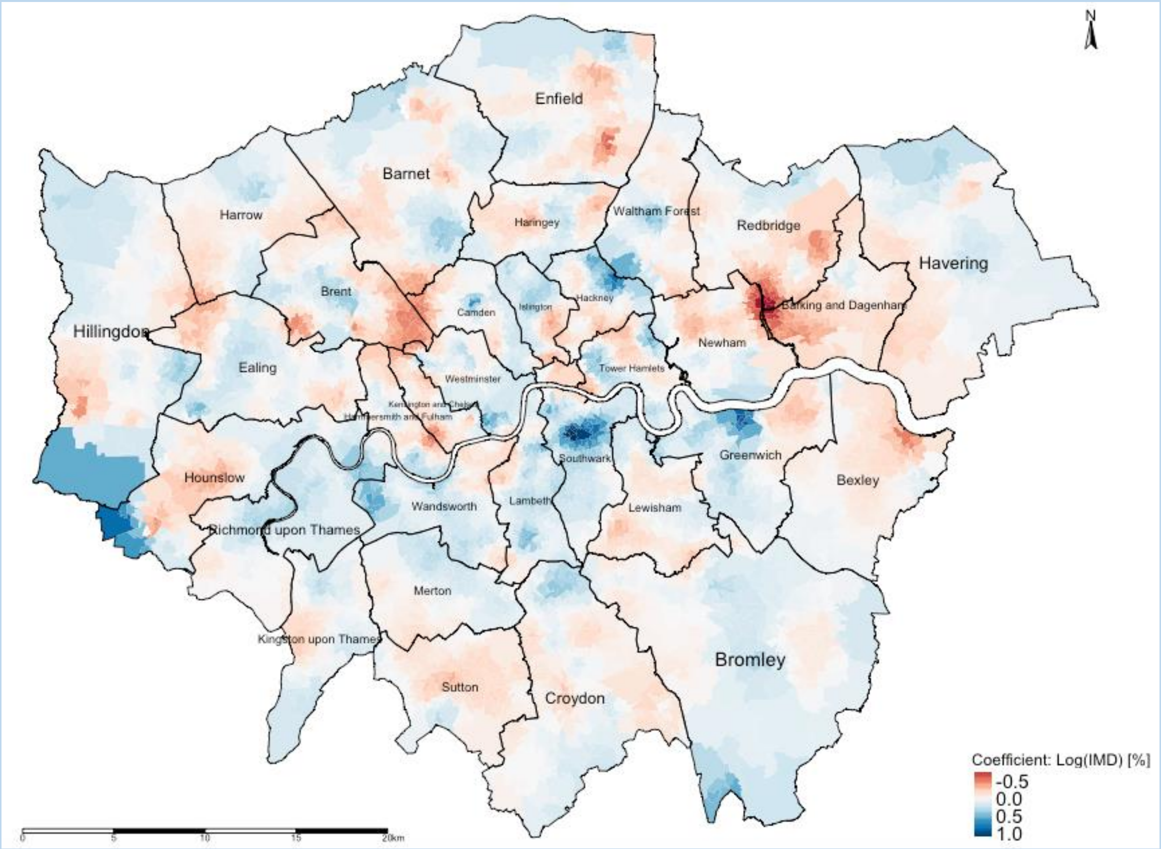
Reporting the Global estimates

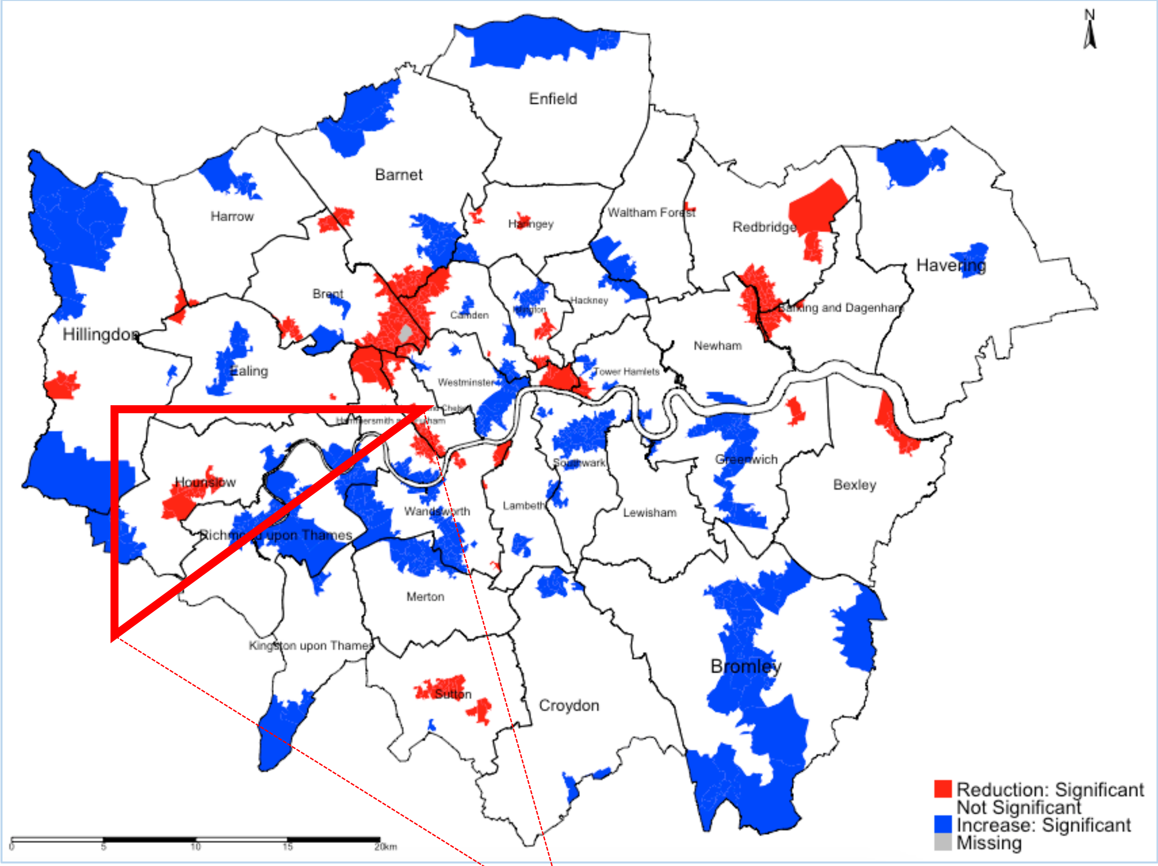
Modelled results in table are from Linear, Spatial Lag (Y) and GWR regression model

Variable(s)	**Linear**	**Lag (Y)**	**GWR**
log(Income)	2.036*	1.267*	2.036
log(Deprivation)	0.136*	0.045*	0.137
log(PTAL)	0.031*	0.011*	0.030
AIC	**-8510.8**	**-9863.3**	**-11242**
R^2	0.7889 (78.89%)	N/A	0.9318 (93.18%)

- The GWR model is better than the linear and Spatial Lag regression. We take the model with the highest R-squared value, as well as the lowest AIC value.

Reporting the local estimates (using socioeconomic deprivation (adjusted for other risk factors) as a motivating example)

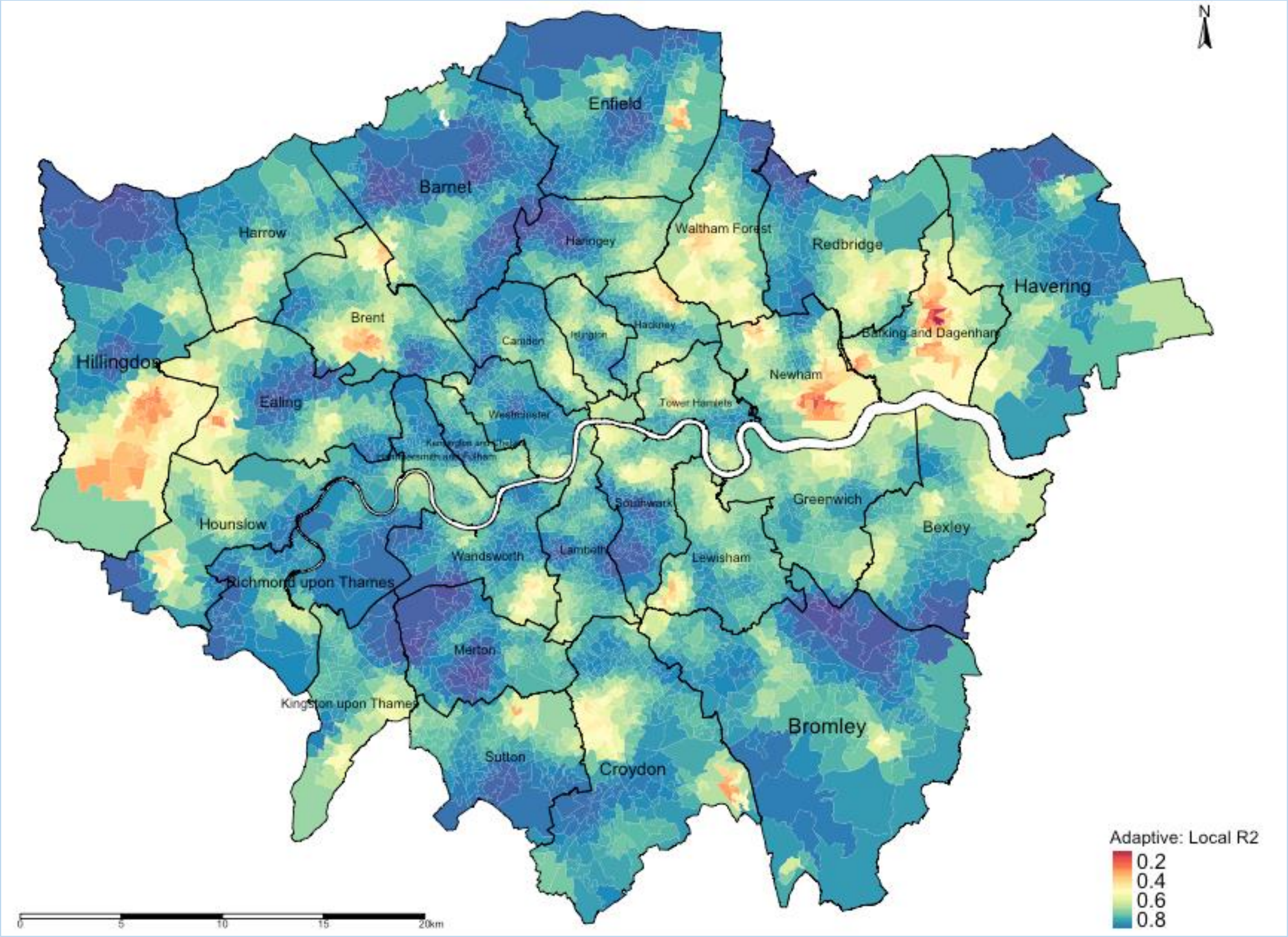




Interpretation: There is spatial variability in the relationship between our variable socioeconomic deprivation (transformed) and averaged house price (transformed) in London. The GWR outputs reveals that local coefficients range from a minimum value of -0.946 to a maximum value of 1.085, indicating that one percentage point increase in the levels of deprivation in LSOAs of London is associated with a reduction of 0.946% in house prices in some LSOAs and (weirdly) an increase of 1.085% in others. Broadly, the relationship are opposing.

Interpretation: For instance, in the **Borough of Hounslow**, we can see a significant reduction in house prices in relation to increased levels of socioeconomic deprivation (adjusted for income and accessibility). Such reduction are clustered in the mid-section of Borough of Hounslow which were coloured red. Note that in far north eastern section of the Borough of Hounslow with pockets of LSOA's coloured blue shows a significant increase in house price in relationship to IMD which is difficult to explain and thus can be interpreted as a chance finding. All sections that are coloured white are not significant.

Reporting the local R-squared to assess the model's performance for each areas



Interpretation: The areas that are going towards the shade of dark reds (i.e., value of 0) are local regression models that have broadly performed poorly in its prediction for house price and its association with the three variables (income, deprivation and PTAL). Likewise, the areas that are going towards the shade of dark blues (i.e., value of 1) are local regression models that have broadly performed very well in its prediction for house price and its association with the three variables (income, deprivation and PTAL).

Note: These results are essential as the local R2 values of each area show the model’s ability to predict the explained variance in house prices caused by deprivation, income and accessibility for specific areas.

Any questions?