UCL

# GEOG0114: PRINCIPLES OF SPATIAL ANALYSIS

## WEEK 10: SPATIAL MODELS (PART 2)

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

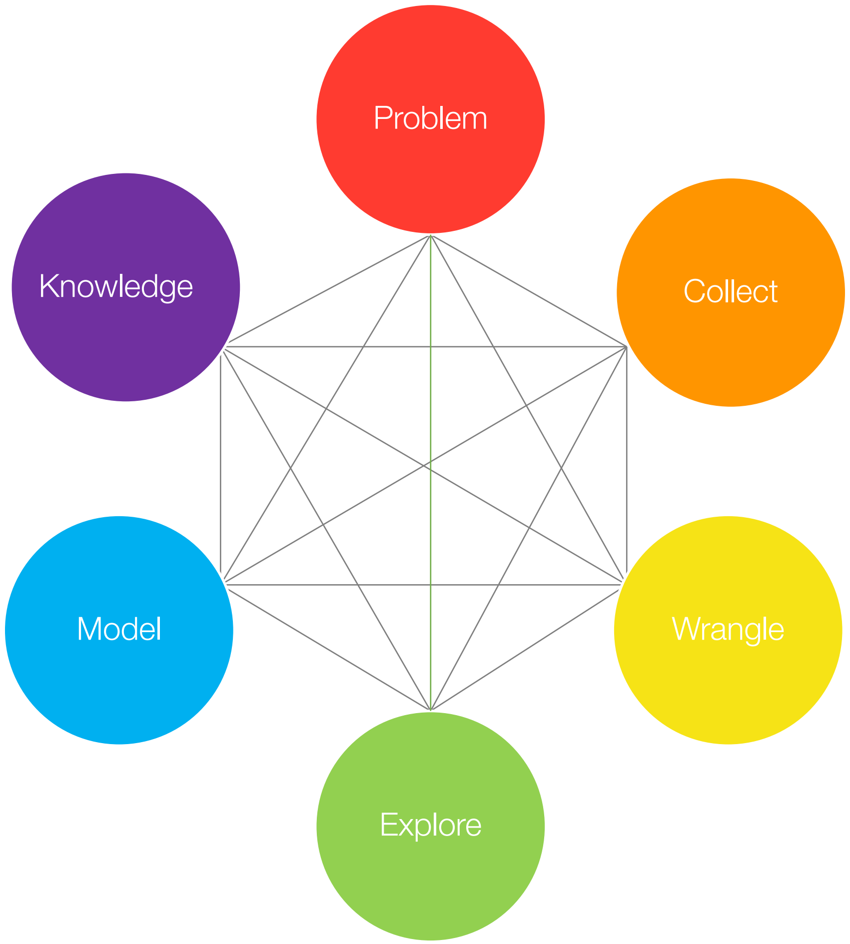
# Contents

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

4. Overarching summary of the GEOG0114



## Recap

# Spatial Models

**Week 9**

Spatial Lag and Error Models

**Week 10**

Geographically Weighted Regression (GWR) Models

In Week 9, we learnt special types of spatial model which accounts for spatial configuration of areal data

# Recap

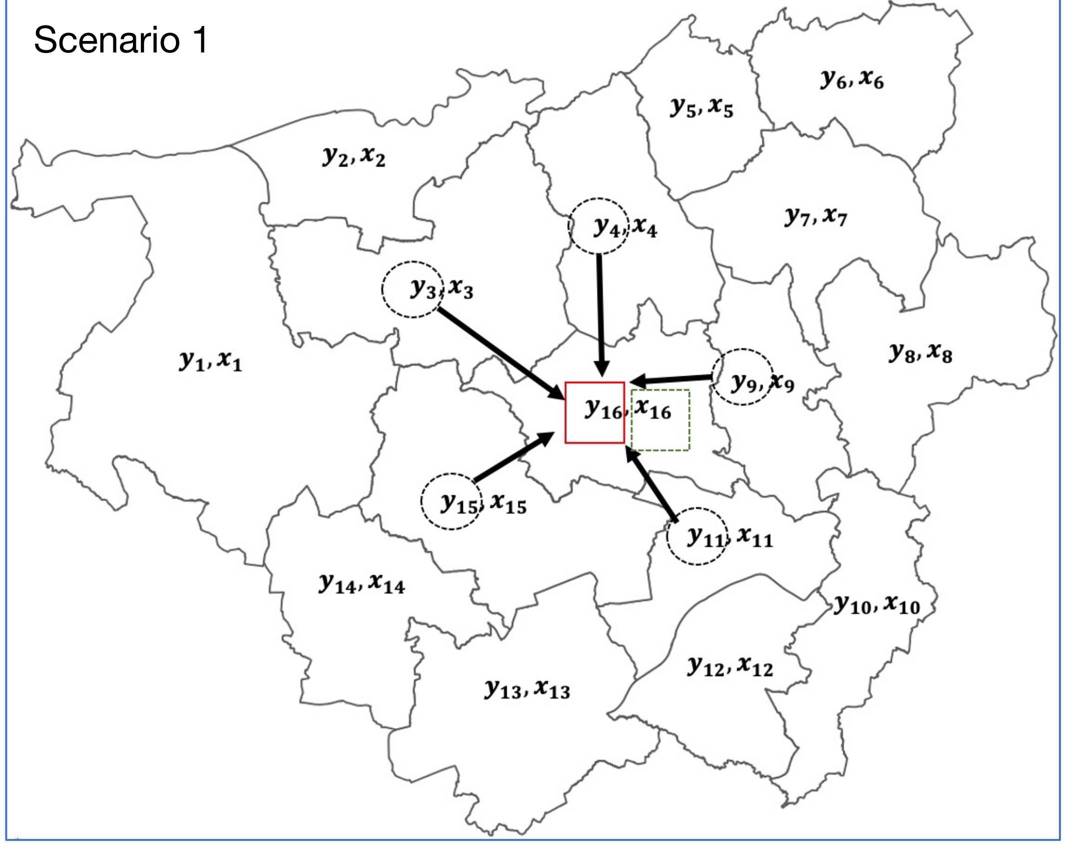
**Multivariable Linear Regression**

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

**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

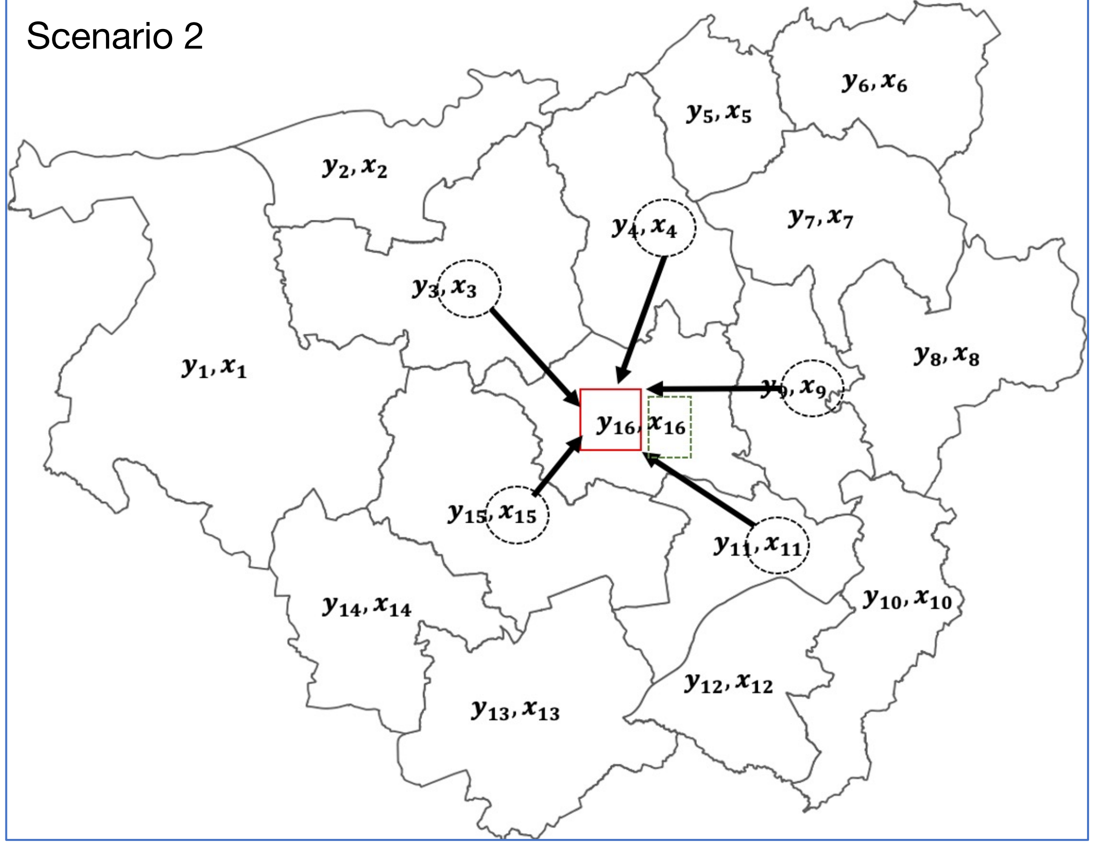



- W is a spatial weights matrix (contiguity – based)
- Y in the right hand of the equation represents the observed outcome from other areas neighbouring that influences what we're trying to predict
- $\rho$ "Rho" is the degree of how our predicted outcome are influenced by its neighbouring Y measures.

**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
- $\theta$ "theta" is the degree of how our predicted outcome are influenced by its neighbouring X measures.

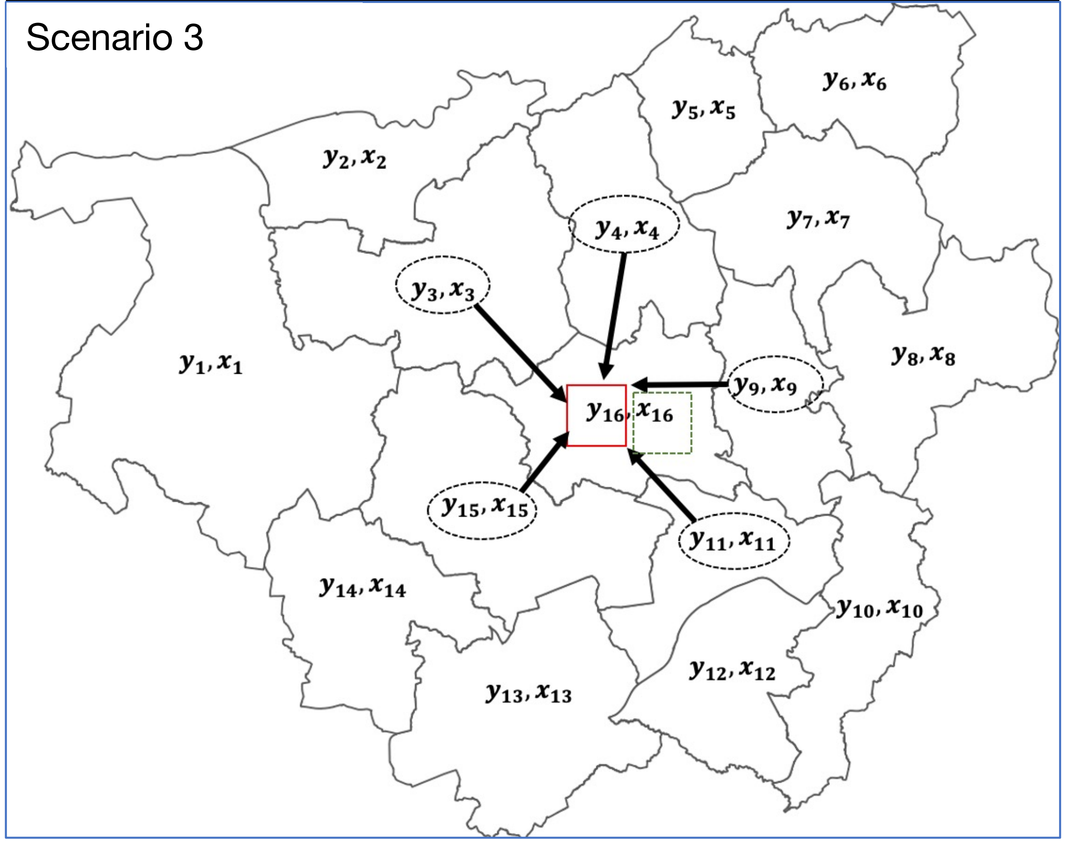
## Recap

**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)**

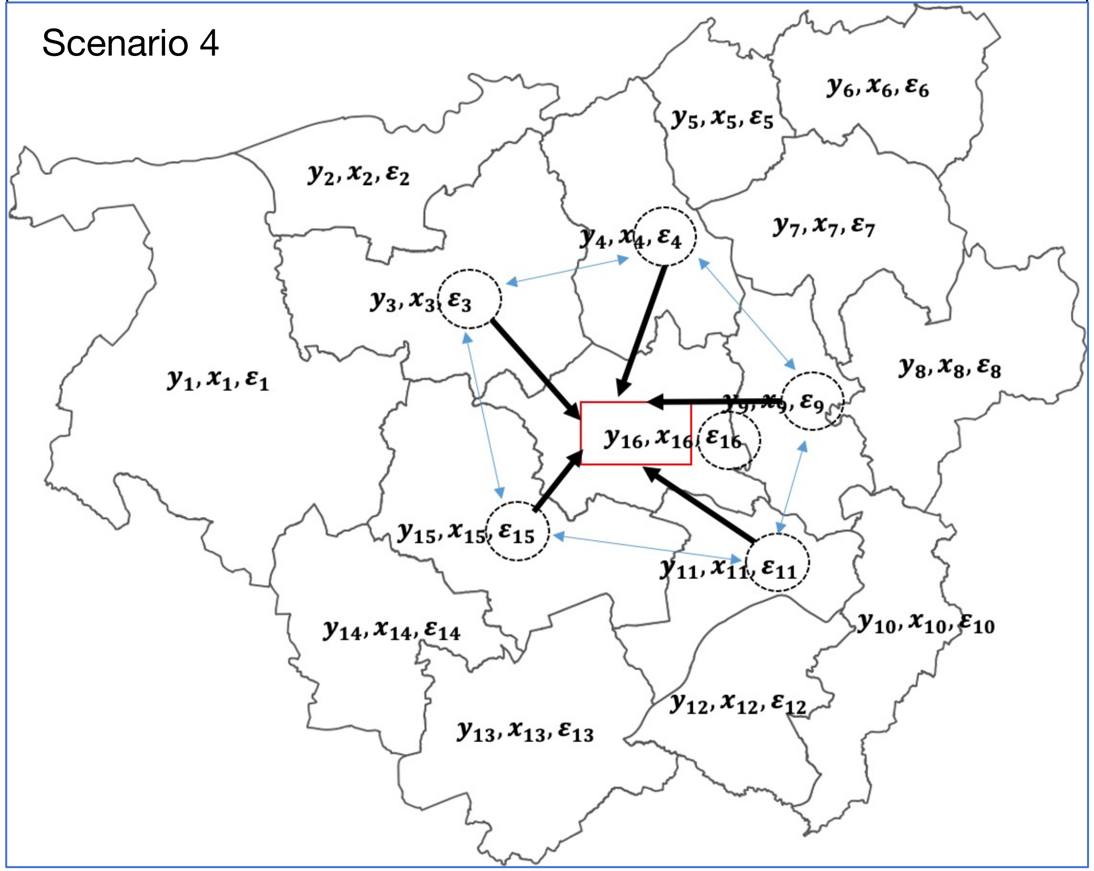
$$y = \rho WY + \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \theta WX + \varepsilon$$




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

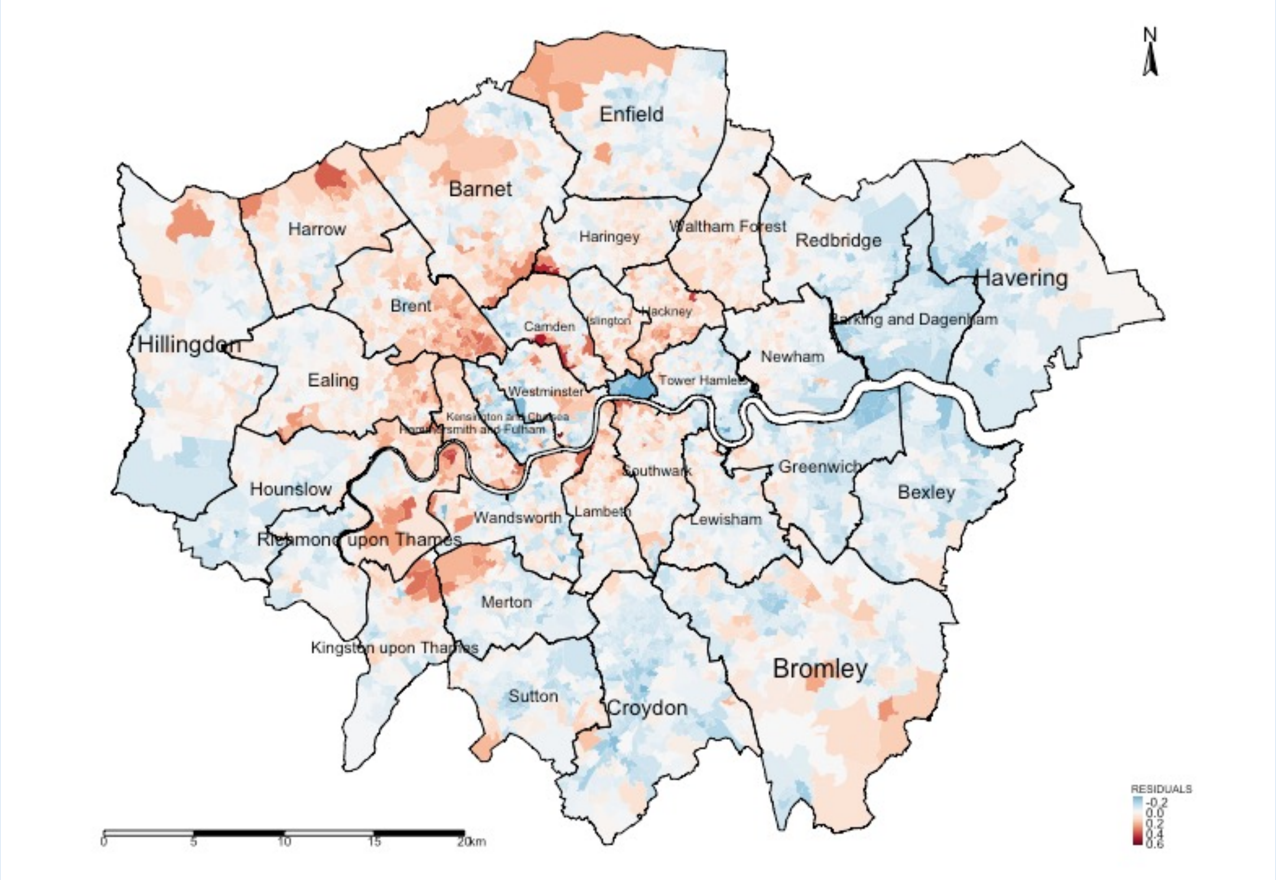
**4. Spatial Error Model**

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




- u are the correlated spatial error terms
- $\lambda$ "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 9**

Spatial Lag and Error Models

**Week 10**

Geographically Weighted Regression (GWR) 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 relationship** or **association** 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 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

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

**Variables**

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

**Parameters**

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

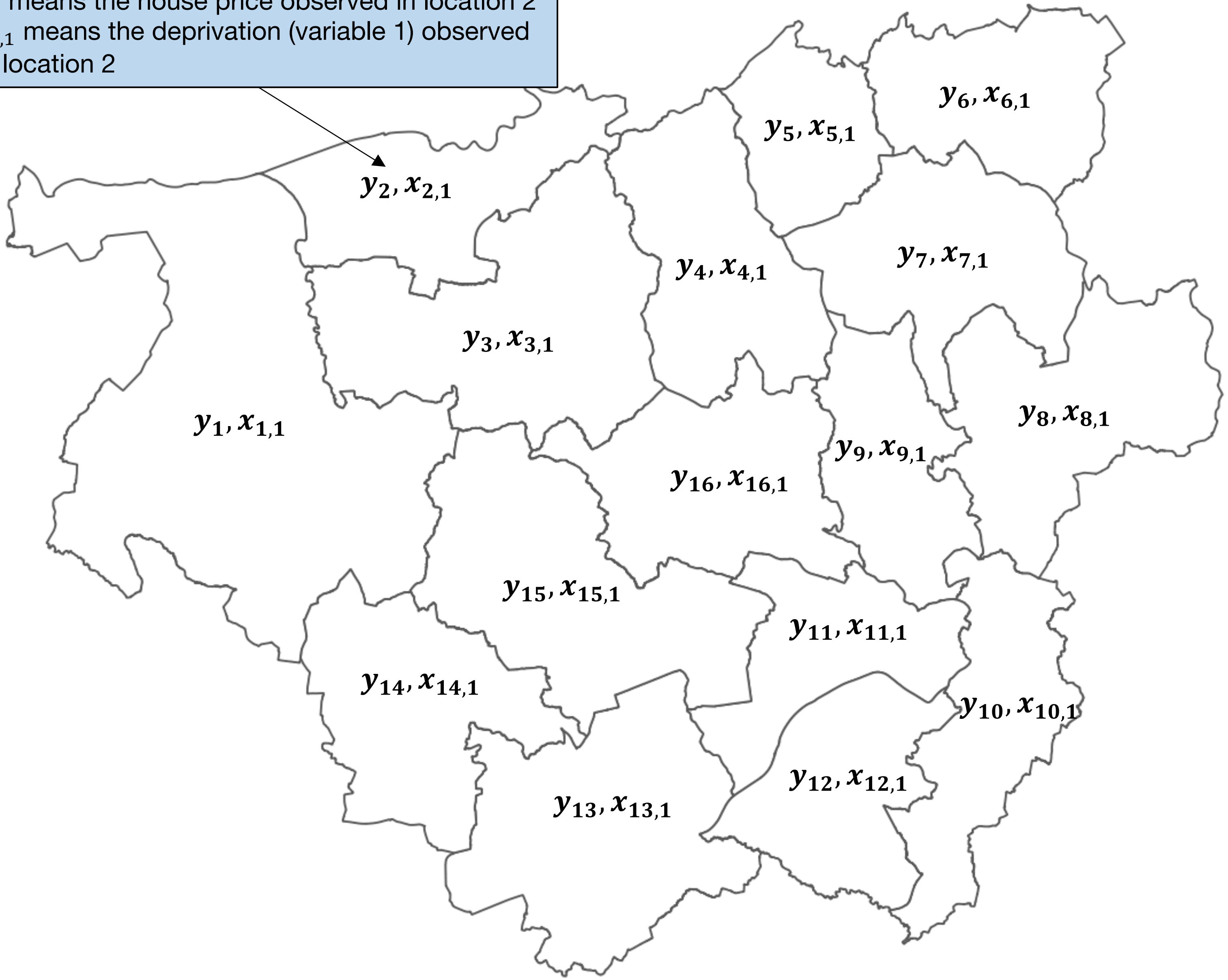
- $\boldsymbol{y_i}$ is the dependent variable indexed at observation (or location) $i$
- $x_{i1}$, $x_{i2}$ and so to $x_{ik}$ are the k-number of independent variables indexed at at

**Parameters**

- $\boldsymbol{\beta_{i0}}(u_i, v_i)$ is the intercept as a function of a geographic location (i.e., coordinates on a grid)
- $\boldsymbol{\beta_{i1}}(u_i, v_i), \boldsymbol{\beta_{i2}}\ (u_i, v_i)$, $\boldsymbol{\beta_{i3}}(u_i, v_i)\boldsymbol{,\ldots,}\ \boldsymbol{\beta_{ik}}(u_i, v_i)$ are the slopes (or coefficients) for the corresponding variables $x_{i1}, x_{i2}, x_{i3}, \ldots, x_{ik}$ 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. $\boldsymbol{x_{1,1}}, \boldsymbol{x_{2,1}} \dots \boldsymbol{x_{16,1}}$)

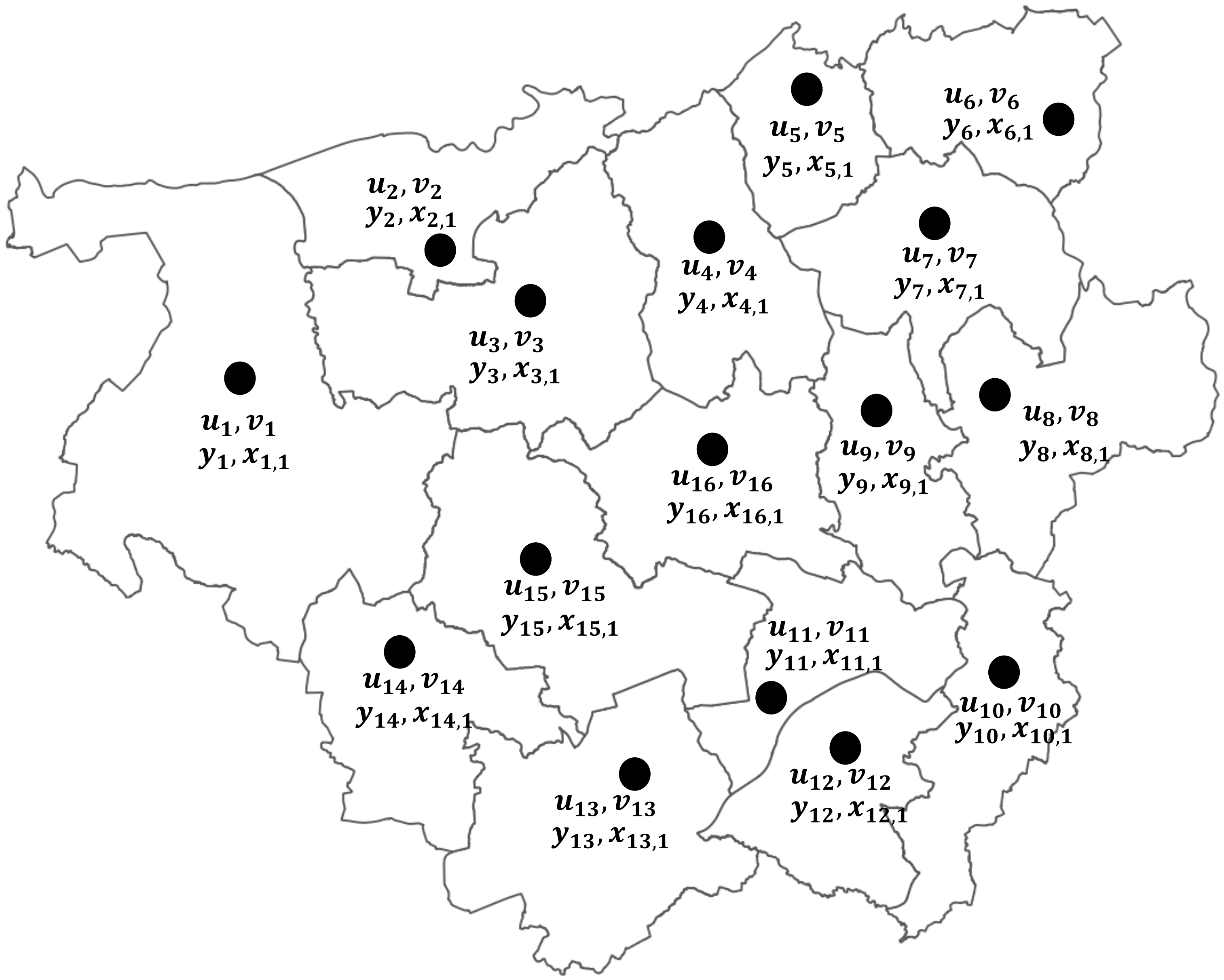
- For example: Socioeconomic deprivation score

**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 $i$.

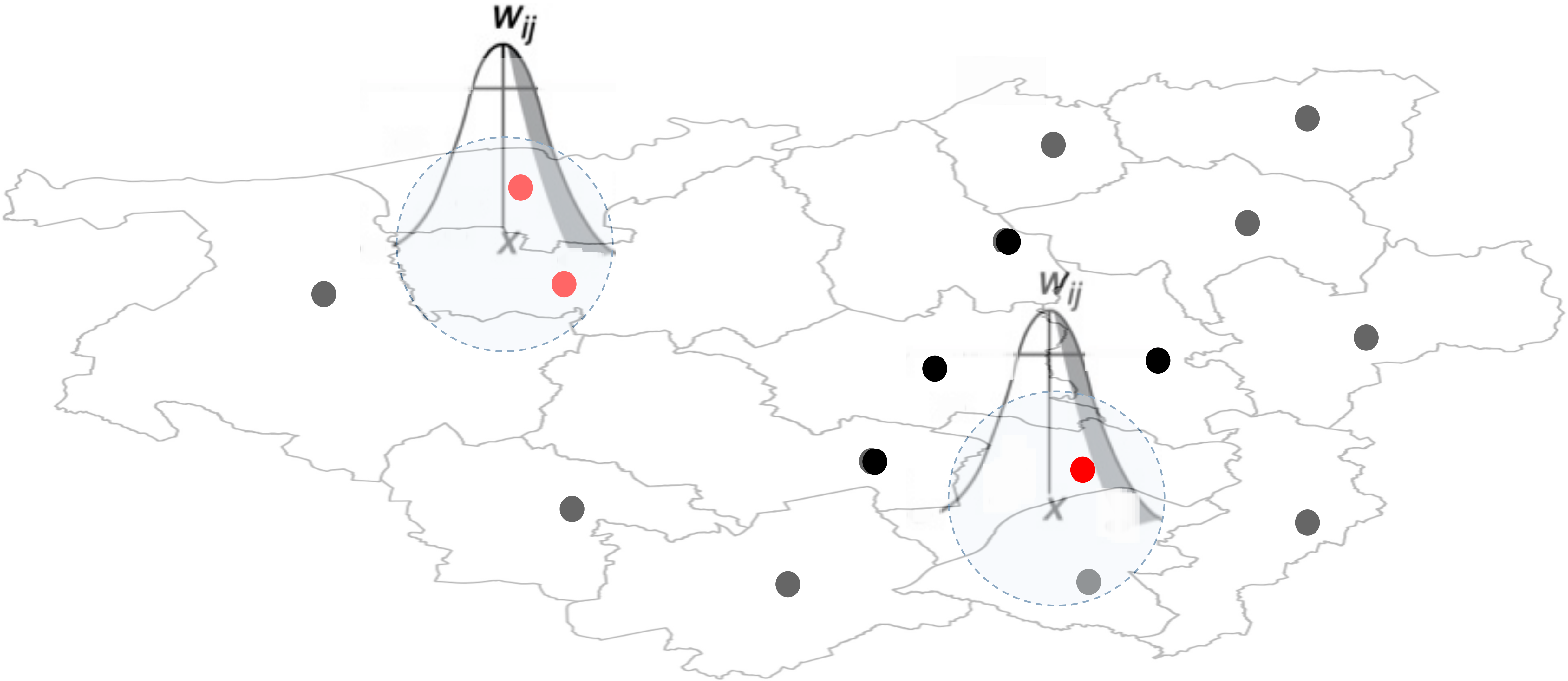
In this example for IMD, these coefficients are represented as $\boldsymbol{\beta_{i,1}}(u_i, v_i)$

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.

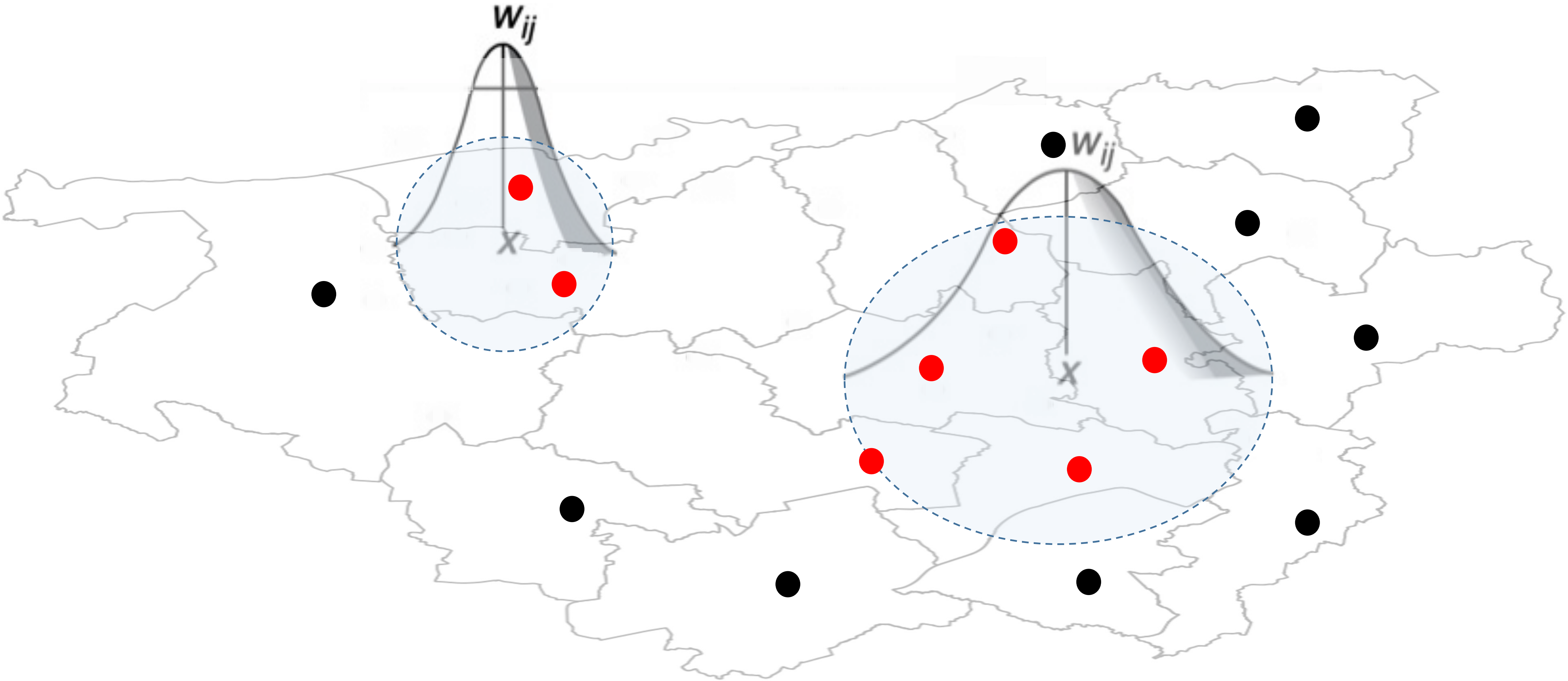
$$\boldsymbol{y_i} = \boldsymbol{\beta_{i,0}}(u_i, v_i) + \boldsymbol{\beta_{i,1}}(u_i, v_i)\boldsymbol{x_{i,1}} + \boldsymbol{\varepsilon_i}$$

# 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)

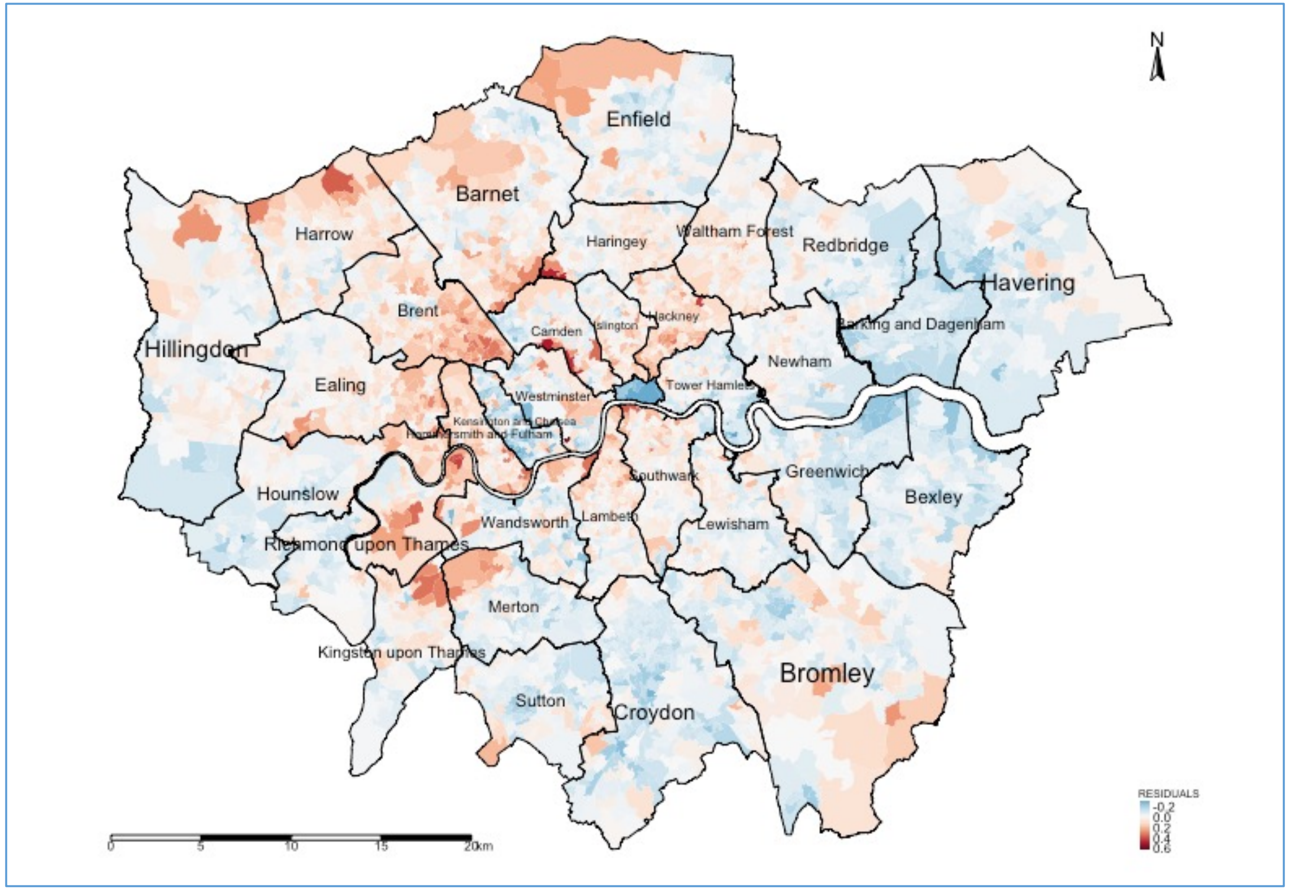
If there's a violation:

- **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.

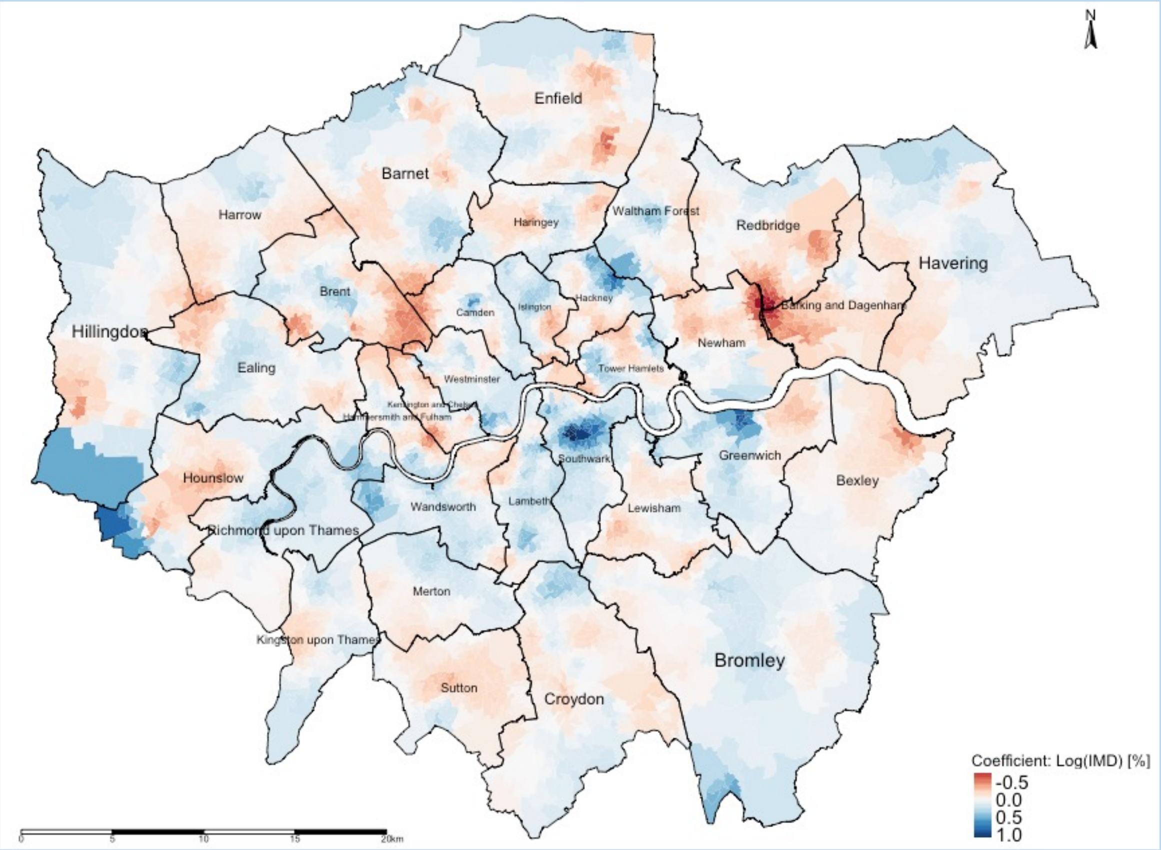
## Reporting the Global estimates

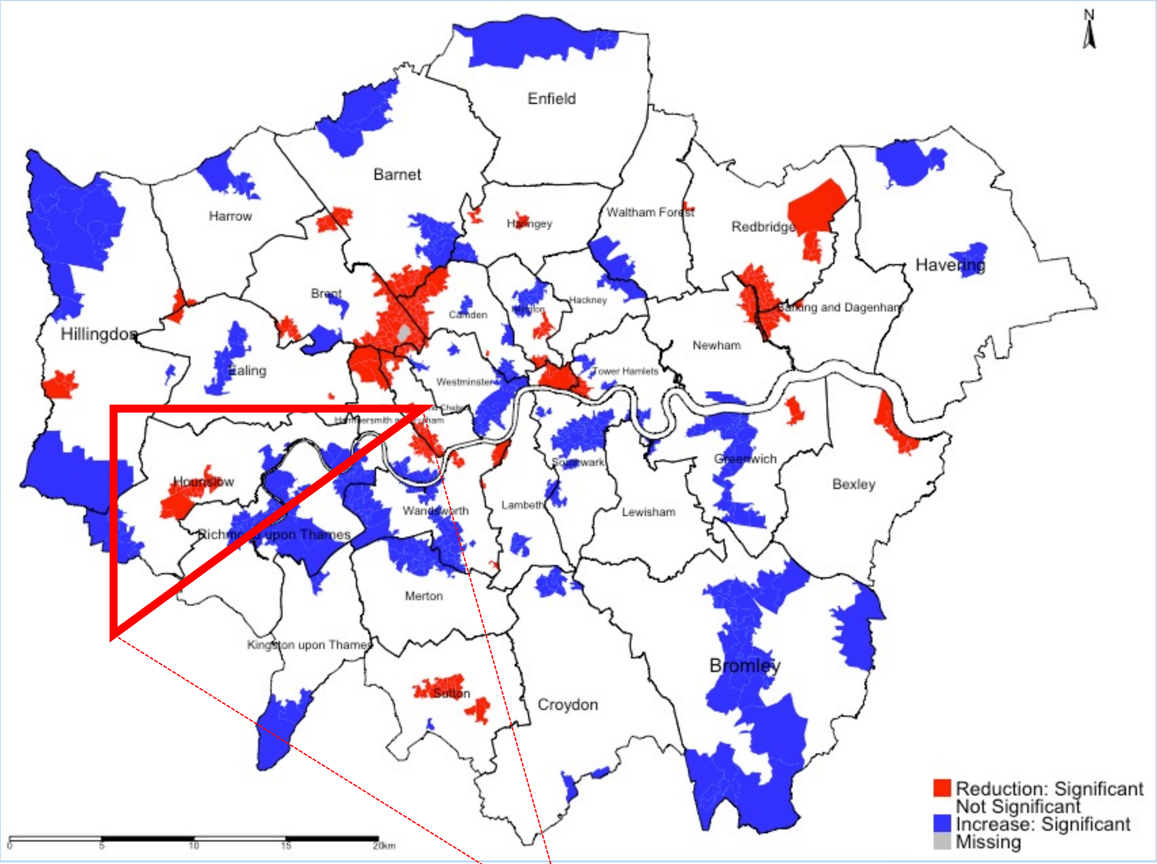
Modelled results in table are from Linear, Spatial Lag (Y) and GWR regression model

| **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 an motivating example)

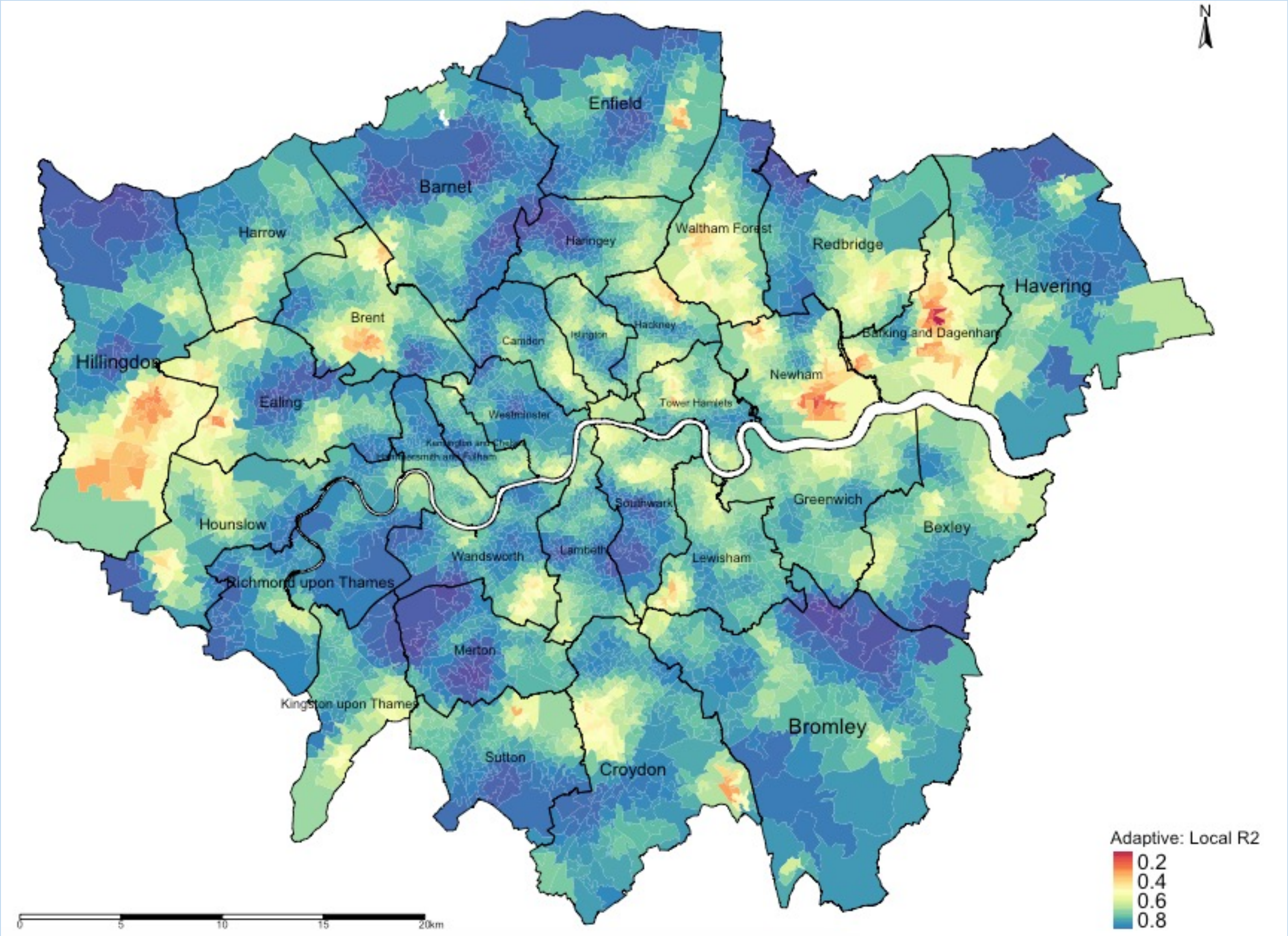



**Interpretation:** There is spatial variability in the relationship between our variable socioeconomic deprivation (transformed) and averaged house price (transformed) in London. The GWR outputs reveals that local coefficients range from a minimum value of -0.946 to a maximum value of 1.085, indicating 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?

# Summary of GEOG0114 & Assessment

# What have we covered in the last 10 weeks…

| Week | Weekly Topics |
|---|---|
| 1 | Spatial Analysis for Data Science |
| 2 | Graphical Representation of Spatial Data |
| 3 | Spatial Autocorrelation |
| 4 | Suitability Mapping: Part I (Qualitative approach) |
| 5 | Suitability Mapping: Part II (Quantitative approach) |
| | Reading Week (Assessment) |
| 6 | Geostatistics using Kriging |
| 7 | Geodemographics |
| 8 | Transport Network Analysis |
| 9 | Spatial Models: Part I (Spatial Lag & Error Models) |
| 10 | Spatial Models: Part II (Geographically Weighted Regression) |

## Tips/Strategy for the Assignment

**Overarching descriptive analysis of the raster or area dataset**

Basic summary measures
Spatial distribution and patterning
Description about the dataset

**Areal data**

**Moran's I test**

**Regression-Based Models**

**Geodemographic Classification**

**Linear Regression Models**

**Spatial Lag and Error Models**

**Geographically Weighted Regression**

**Raster-based analysis**

What process does the point data follow?

**Geostatistical**

**Ordinary/Regression Kriging**

**Point Process**

**Ecological Niche Model**

Standalones

Unique aggregated data for street segments?

**Transport Network Analysis**

Raster only?

**Analytical Hierarchical Process**

# Go to Moodle's Assessment

**Assignment Documentation: Guidance notes and Cover Sheet**

GEOG0114 Spatial Data Analysis Project - Instructions & Guidance Notes 199.9 KB PDF document

Please download the worksheet for the Spatial Data Analysis Project. For further guidance, you find this information contained in this documentation.

GEOG0114 Spatial Data Analysis Project - Cover Sheet template 42.1 KB Word 2007 document

Please download the cover sheet template for your Spatial Data Analysis Project. This document must be submitted as the front page of your 3,000 word report. It must contain your **candidate ID**, **title of report**, **the word count**, and **link(s) to R-scripts and data sets** (e.g., OneDrive, Google Drive or GitHub repository).

**Below are the GREAT examples of past projects:**

## PLEASE GO THROUGH THE DOCUMENTATION!

**See Best examples of past projects from 2021/22 and 2022/23**

- Thermal inequities in Chicago neighborhoods (2021/22) 558.4 KB PDF document
- How social determinants of health effect health conditions in London (2021/22) 4.3 MB PDF document
- A spatial analysis of gun violence in the city of New York (2021/22) 453.1 KB PDF document
- Understanding relationships between cultural accessibility and educational deprivation in Inner London (2022/23) 2.2 MB PDF document
- Disparities in COVID-19 death rates among ethnic groups in England and Wales (2022/23) 2.2 MB PDF document
- Influence of the spatial variations of soil properties on soil-transmitted helminths transmission in Nigeria (2022/23) 10.6 MB PDF document

Spatial Data Analysis Project
Assignment's deadline is <u>Monday 8th January 2024 (noon)</u>

**Example Data source(s)**

UK Census 2021: https://www.nomisweb.co.uk/sources/census_2021_bulk
UK Deprivation 2019: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019
UK population 2010-2020 estimates: https://shorturl.at/BCKX1
Police: https://data.police.uk/
**CDRC: https://data.cdrc.ac.uk**

Mosquito/species data:

- https://experience.arcgis.com/experience/7228a5a27442468494caec2934c2b73d/page/Page/
- https://www.gbif.org

Wildfire FIRMS: https://firms.modaps.eosdis.nasa.gov

Disease data: https://espen.afro.who.int/
https://espen.afro.who.int/tools-resources/cartography-database
https://espen.afro.who.int/tools-resources/download-data

Domestic fires LSOA:
https://assets.publishing.service.gov.uk/media/64be67369c2df000129402cf/low-level-geography-dataset-270723.ods
https://www.gov.uk/government/statistics/fire-statistics-incident-level-datasets/low-level-geography-dataset-guidance

**Inspiration for writing methodology**

**Musah et al. (2020): https://doi.org/10.1016/j.apgeog.2019.102126**
**Todd et al. (2022): https://doi.org/10.1177/23998083211001836**
**Li et al. (2022): https://doi.org/10.1016/j.apgeog.2022.102718**

Any questions about the assignment?

## GEOG0114: Course Evaluation & Student Feedback (Week 7-10)

# https://forms.gle/55AydSPLrV8cBcwn7

Dear Students,

As part of the Continuous Module Dialogue, we are conducting this survey to gauge the levels of student satisfaction with the learning experience in module **GEOG0114: Principles of Spatial Analysis**. We would like to receive your feedback, which would be greatly appreciated. This will help us make improvements to the course. The survey should only take up to 5 or 10 minutes, and your responses are completely anonymous.

Thank you,

Anwar and Justin.

# See you in term 2