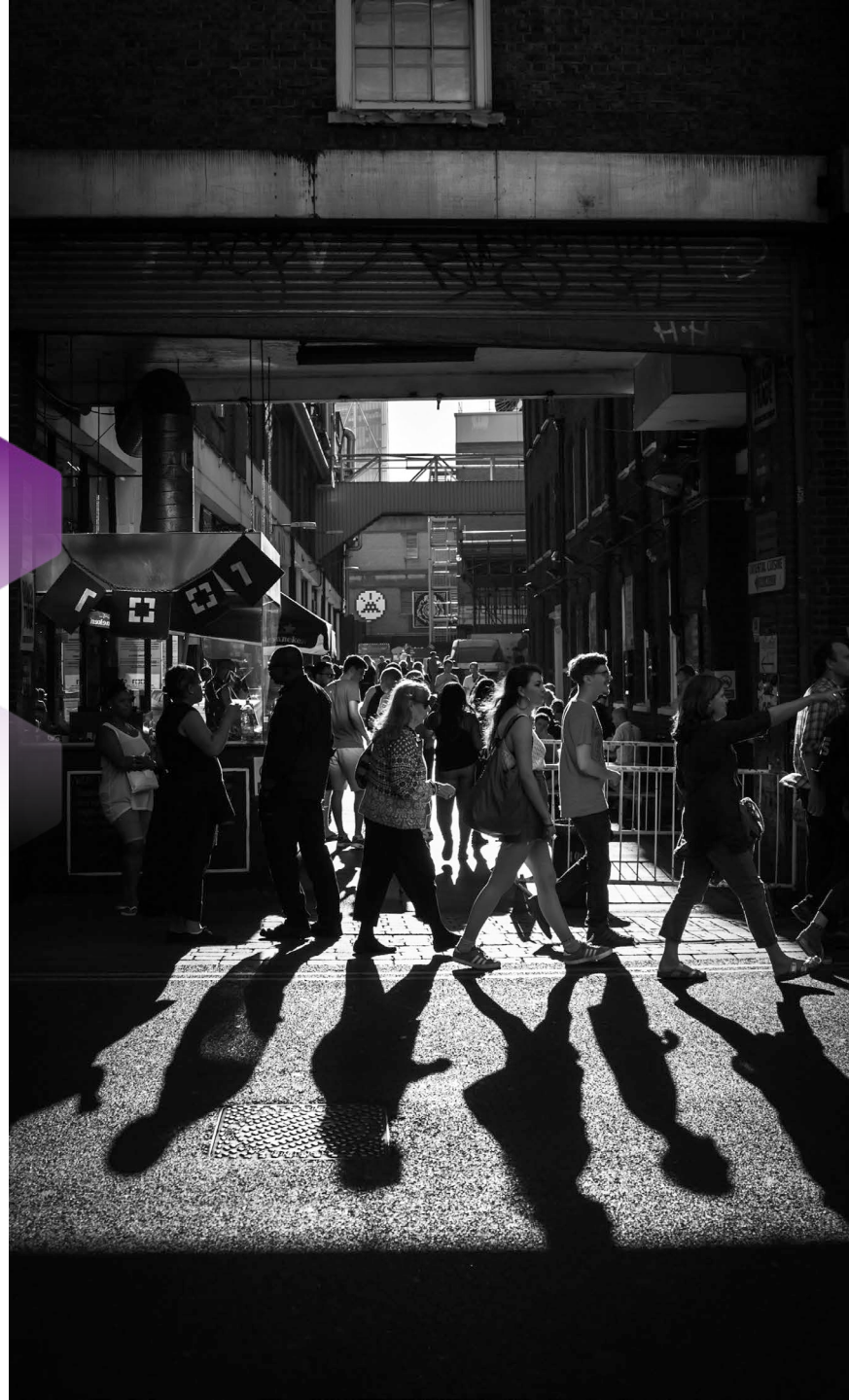
UK Data Service

Placing the pandemic

An ecological study of Covid-19



Economic and Social Research Council

UK Data Service

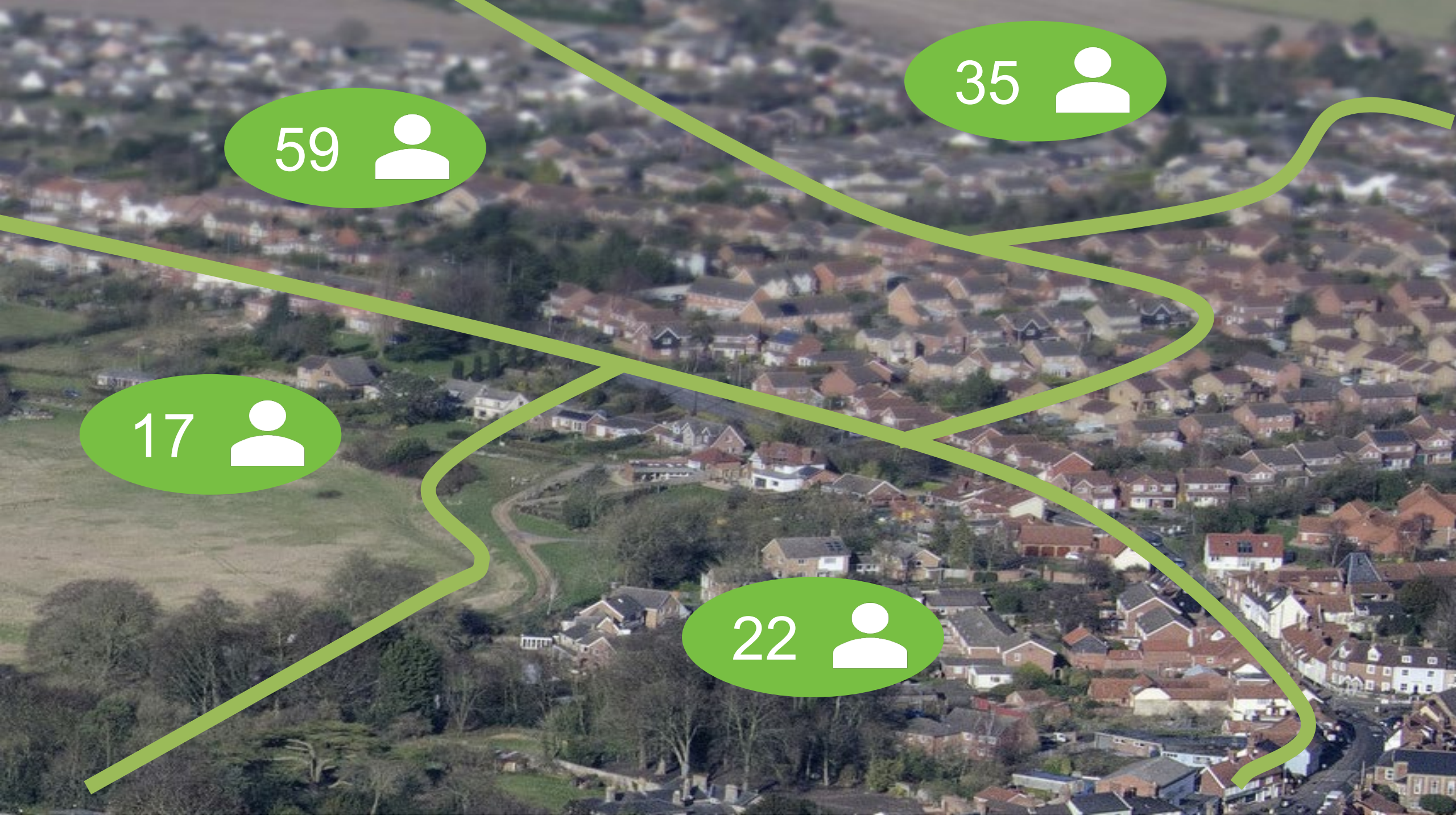
A Quick Tour of UK Census Data

David Rawnsley
UKDS Senior Data Co-ordinator, Aggregate Data

Research Methods Festival 27 October 2021

Quiz time

- A – 2,467
- B – 27,229
- C – 176,632
- D – 390,870
- E – 1,156,221


59
35
17
22

The Census of Population

- Every 10 years
- First modern census in 1841
- Data held secure for 100 years
- Digitised censuses from 1971 (1961 ongoing)
- Last ever census?

A UK Census?

- Different nations, different governments, different ideas.
  - Office for National Statistics
  - National Records of Scotland (formerly GROS)
  - Northern Ireland Statistics and Research Agency

  - Different Geographies
  - Different Questions
  - Different Outputs
  - Lots of harmonisation

England and Wales Census
21st March 2021

- 97 % response
- Average completion time – 23 minutes
- Census Coverage Survey
- Census Quality Survey
- Output consultation

Northern Ireland Census
21st March 2021

- Census Coverage Survey
- Outputs consultation
- Add geographies
- Statistical Disclosure Control
- Quality Assurance

Scotland Census
1st March 2022

- Census rehearsal
- Data processing rehearsal
- Work ongoing to ensure that people are not counted twice or not counted at all.

Census Outputs

- Should start to appear 1 year after census day
- Headline figures first
- Ready made tables
- Tables that flex
- Build your own tables
- Links to administrative data

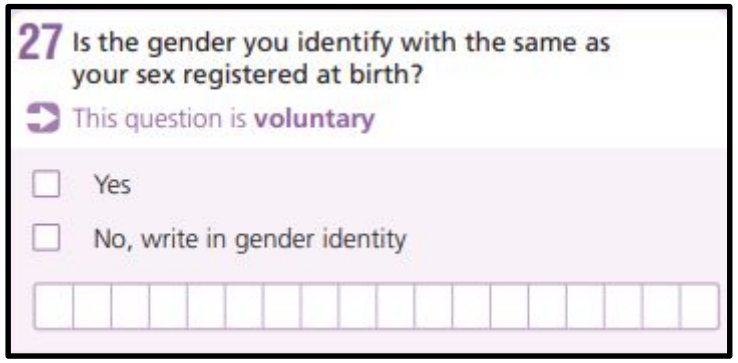
What can the census aggregate data tell us

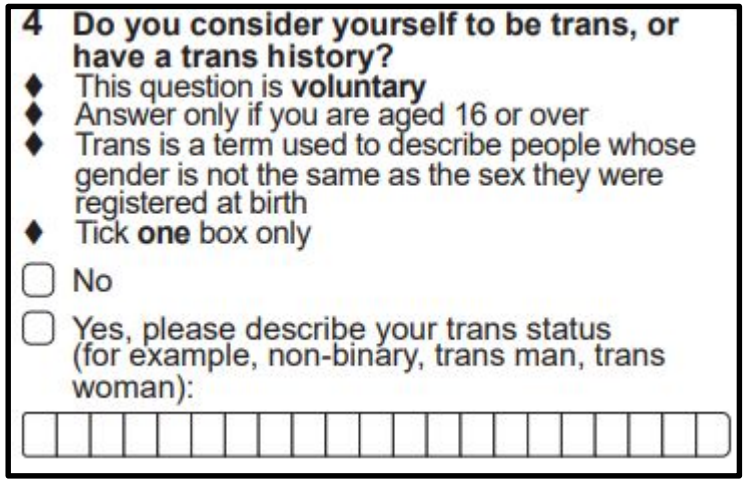
- The most complete source of information about demographic and socio-economic characteristics of the UK population.

  - Population
  - Employment, occupation, socio-economic class, qualifications
  - Ethnicity, religion, national identity, language
  - Housing, tenure, residence type
  - Household and family composition, occupancy
  - Carers and caring
  - Long term health problems or disability
  - Travel to work, migration

New questions

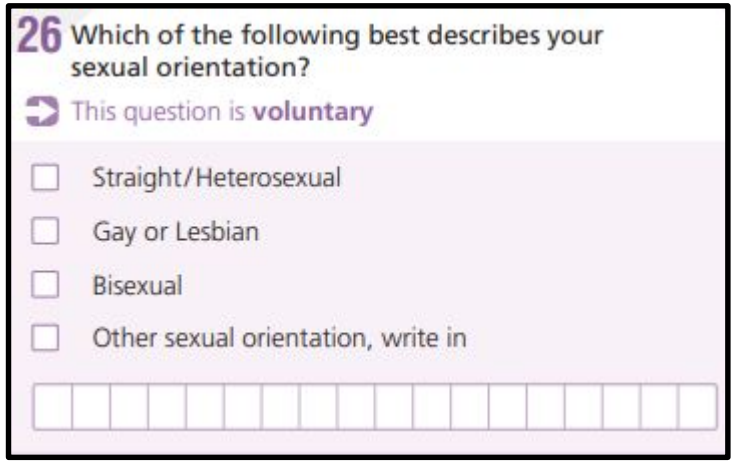
- Gender – England, Wales, Scotland
- Voluntary

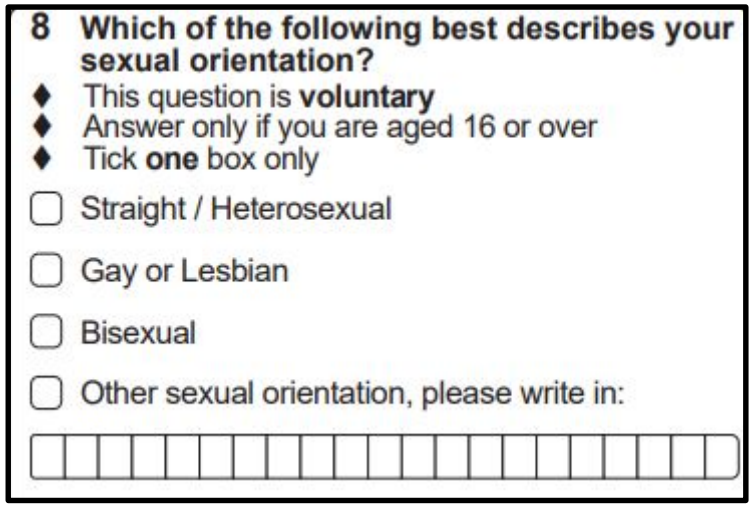




New questions

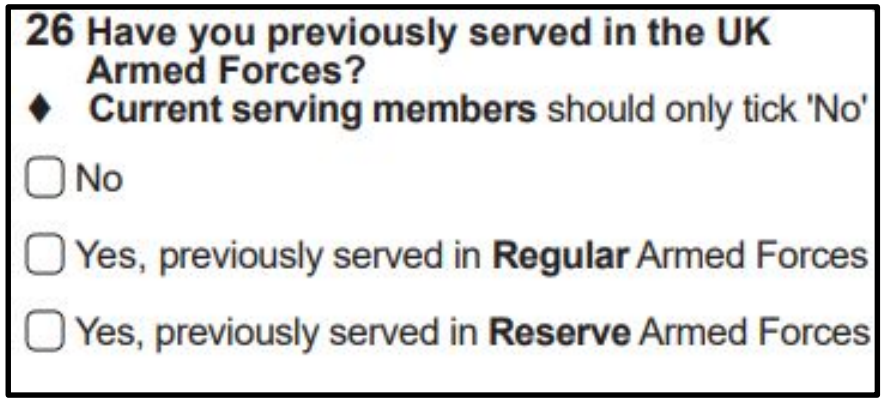
- Sexuality – England, Wales, Scotland, Northern Ireland
- Voluntary





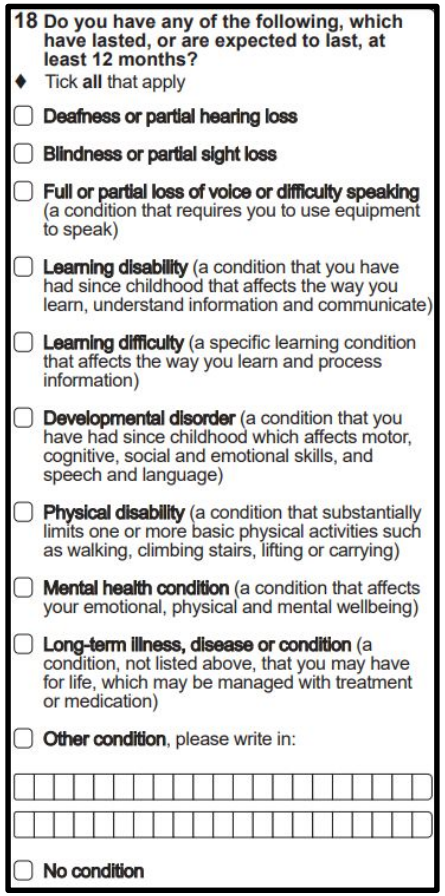
New questions

- Veteran status – England, Wales, Scotland



New questions

- Health conditions – Scotland, Northern Ireland

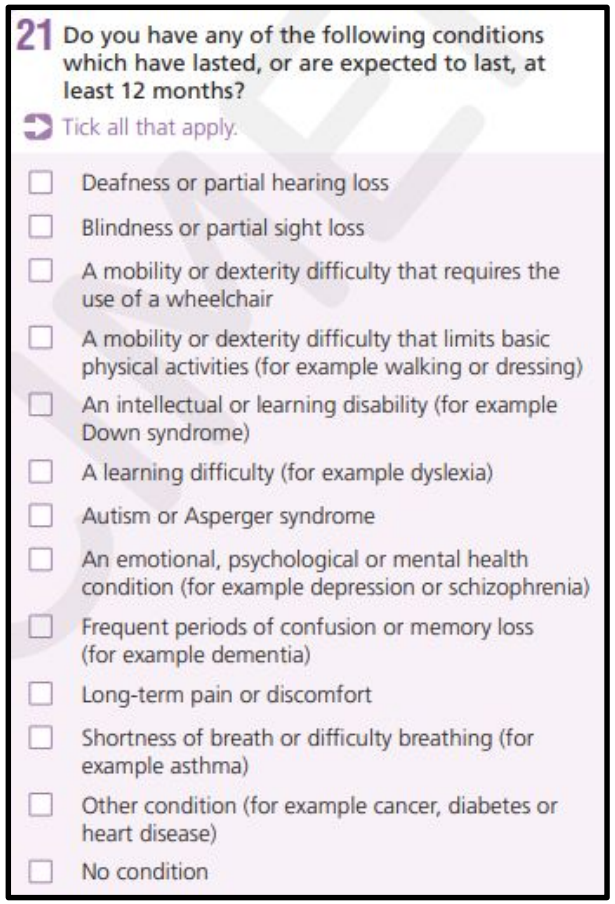


18 **Do you have any of the following, which have lasted, or are expected to last, at least 12 months?**

♦ Tick **all** that apply

- ☐ **Deafness or partial hearing loss**
- ☐ **Blindness or partial sight loss**
- ☐ **Full or partial loss of voice or difficulty speaking** (a condition that requires you to use equipment to speak)
- ☐ **Learning disability** (a condition that you have had since childhood that affects the way you learn, understand information and communicate)
- ☐ **Learning difficulty** (a specific learning condition that affects the way you learn and process information)
- ☐ **Developmental disorder** (a condition that you have had since childhood which affects motor, cognitive, social and emotional skills, and speech and language)
- ☐ **Physical disability** (a condition that substantially limits one or more basic physical activities such as walking, climbing stairs, lifting or carrying)
- ☐ **Mental health condition** (a condition that affects your emotional, physical and mental wellbeing)
- ☐ **Long-term illness, disease or condition** (a condition, not listed above, that you may have for life, which may be managed with treatment or medication)
- ☐ **Other condition**, please write in:

☐ **No condition**



21 **Do you have any of the following conditions which have lasted, or are expected to last, at least 12 months?**

Tick all that apply.

- ☐ Deafness or partial hearing loss
- ☐ Blindness or partial sight loss
- ☐ A mobility or dexterity difficulty that requires the use of a wheelchair
- ☐ A mobility or dexterity difficulty that limits basic physical activities (for example walking or dressing)
- ☐ An intellectual or learning disability (for example Down syndrome)
- ☐ A learning difficulty (for example dyslexia)
- ☐ Autism or Asperger syndrome
- ☐ An emotional, psychological or mental health condition (for example depression or schizophrenia)
- ☐ Frequent periods of confusion or memory loss (for example dementia)
- ☐ Long-term pain or discomfort
- ☐ Shortness of breath or difficulty breathing (for example asthma)
- ☐ Other condition (for example cancer, diabetes or heart disease)
- ☐ No condition

What the census cannot tell us

- Wealth and income
  - Derived deprivation data

- Personal identification
    - Data blurring and obfuscation

The Tricky Issue of Census Geography

- Building block is the Output Area
- These are used to create Super Output Areas (SOA)
  - Two types – Lower layer SOA (LSOA) and Middle layer SOA (MSOA)
  - In Scotland – Data Zones (DZ) and Intermediate Geographies (IG)
  - In Northern Ireland – just LSOA

- None of the above relate to anything ‘real’.

- Regions, Counties, Local Authorities, Wards and Electoral Divisions
- No postcode geography
- http://geoconvert.digitalresources.jisc.ac.uk/

Census Boundary Data

- UK Borders Easy Download
  https://borders.ukdataservice.ac.uk/easy_download.html
- 2011 – 1981 & 2021 when released
- Mapinfo, Shapefile, KML, CSV
- Casweb has 2001/1991 data and boundaries bundled together
- UK Borders Boundary Data Selector
  https://borders.ukdataservice.ac.uk/bds.html

Accessing census aggregate data

- Infuse - http://infuse.ukdataservice.ac.uk/
  - Data from 2001 and 2011 censuses
- Casweb - http://casweb.ukdataservice.ac.uk/
  - Data from 1971, 1981, 1991 and 2001 censuses
  - Boundary data with 1991 & 2001 censuses
- DKAN - https://www.statistics.digitalresources.jisc.ac.uk/
  - Currently only 2011 but currently loading 2001 and plans for censuses back to 1971

Deprivation data

- No questions on income or wealth.
- Derived from room occupancy, house ownership/tenancy, car availability, employment status etc…
- A number of recipes
  - Carstairs
  - Townsend
  - Index of Multiple Deprivation/Indices of Deprivation
  - https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019

Matching Geographies

- GeoConvert - http://geoconvert.digitalresources.jisc.ac.uk/
  - Uses address points to calculate an area population
  - Clever proportioning
  - Caveats
    - Think about what you are attempting to do.
    - Postcodes change
- Supporting documentation - http://geoconvert.digitalresources.jisc.ac.uk/help/documentation.html

Census bulk data

- DKAN - https://www.statistics.digitalresources.jisc.ac.uk/
- Whole tables for geographical areas
- Easy to search
- Metadata
- Expanding all the time

For more information …

- https://ukdataservice.ac.uk/learning-hub/census/

UK Data Service

Ecological analysis

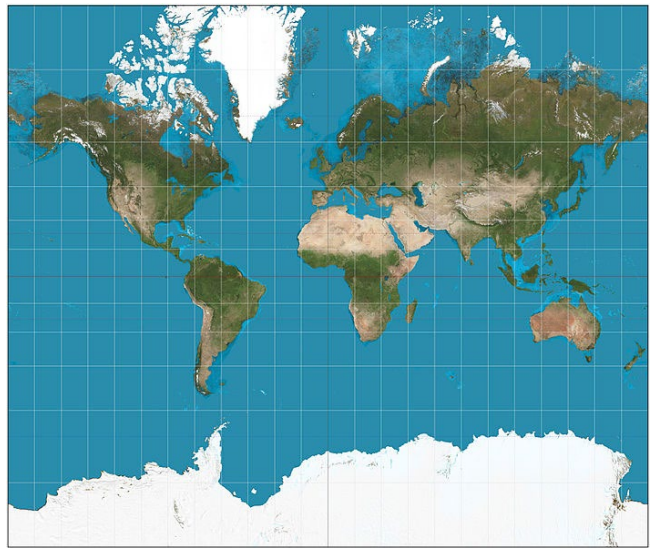
Nigel de Noronha
UKDS Co-ordinator, Census microdata

Research Methods Festival 27 October 2021

Understanding spatial data

- projection
- shape
- data
  - areal
  - point

Projection - from a sphere to a flat surface

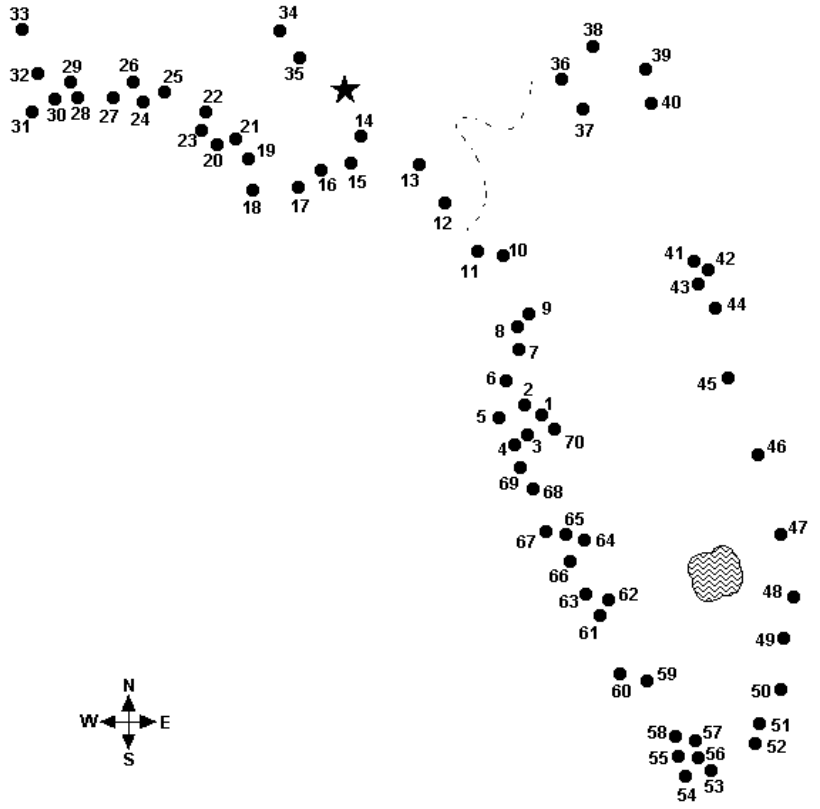


Co-ordinates

- Latitude and Longitude (WGS 1984)
- British National Grid (Eastings & Northings)

Shape files

Commonly used system is based on ESRI developed by Arc-GIS developed by joining up the dots



Data types

Areal data: collected for each geographical unit in the areas you wish to represent

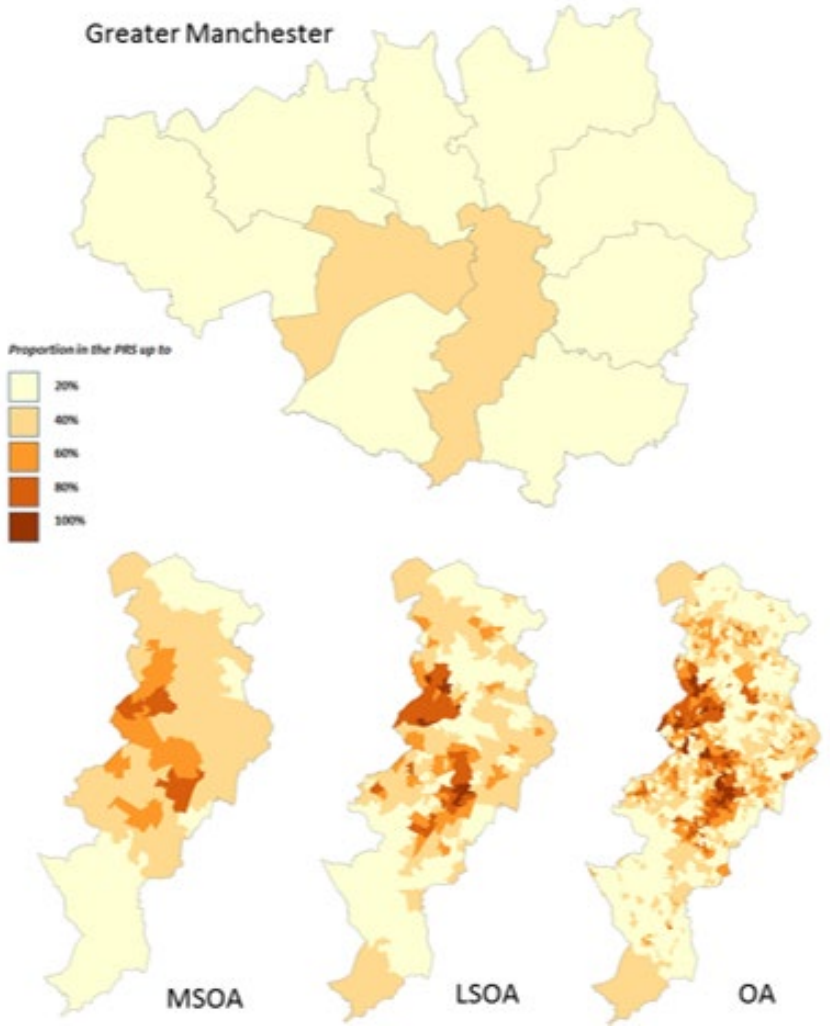
	A	B	C	D	E	F	G	H
1	LSOA	pop	empft	hwind	owned	sochse	prs	sp
2	Bolton 001	537	312	0.703911	0.928741	0.007126	0.064133	0.123515
3	Bolton 001	566	337	0.717314	0.763485	0.002075	0.23444	0.232365
4	Bolton 001	435	251	0.705747	0.900524	0.015707	0.08377	0.157068
5	Bolton 001	560	354	0.723214	0.865116	0.002326	0.132558	0.176744
6	Bolton 001	519	290	0.697495	0.869658	0.004274	0.126068	0.215812

Point data: collected for each location you want to highlight

	A	B	C	D	E	
1	X	Y	Label_Text	Phase	Condition	TIF
2	382360.6	403798.8	Heaton Pa	1 & 2	functional	
3	383530.3	403163.7	Bowker Va	1 & 2	functional	

Choosing the best geographical unit

Figure 1 – proportion of households in the private rented sector (PRS)



Local authority - Manchester is represented by a single shade – PRS under 30%

MSOA (Mid layer super output area) - 57 MSOAs in Manchester and patterns of concentration in city centre and north and south corridors – PRS > 10% and < 70%

LSOA (Lower layer super output areas) - 282 LSOAs in Manchester – PRS > 5% < 90%

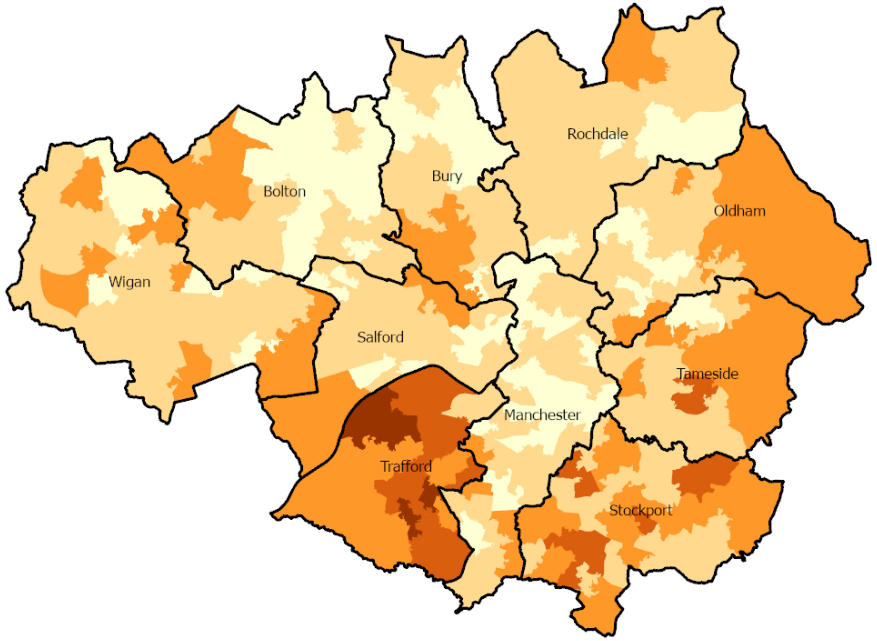
OAs (output areas) – 1530 OAs in Manchester – PRS = 0% and < 97%

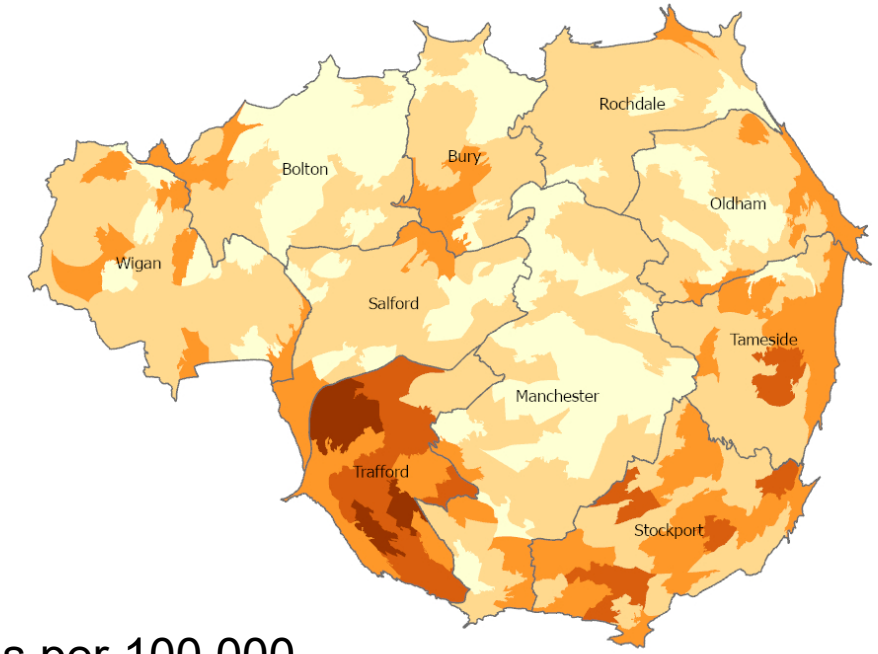
Selected geography

MSOA data (lowest level available for Covid case rates)

Greater Manchester

Representing the data





Covid cases per 100,000 at MSOA level in first week of October 2021

Physical

Cartogram

(resized to reflect even distribution of the population)

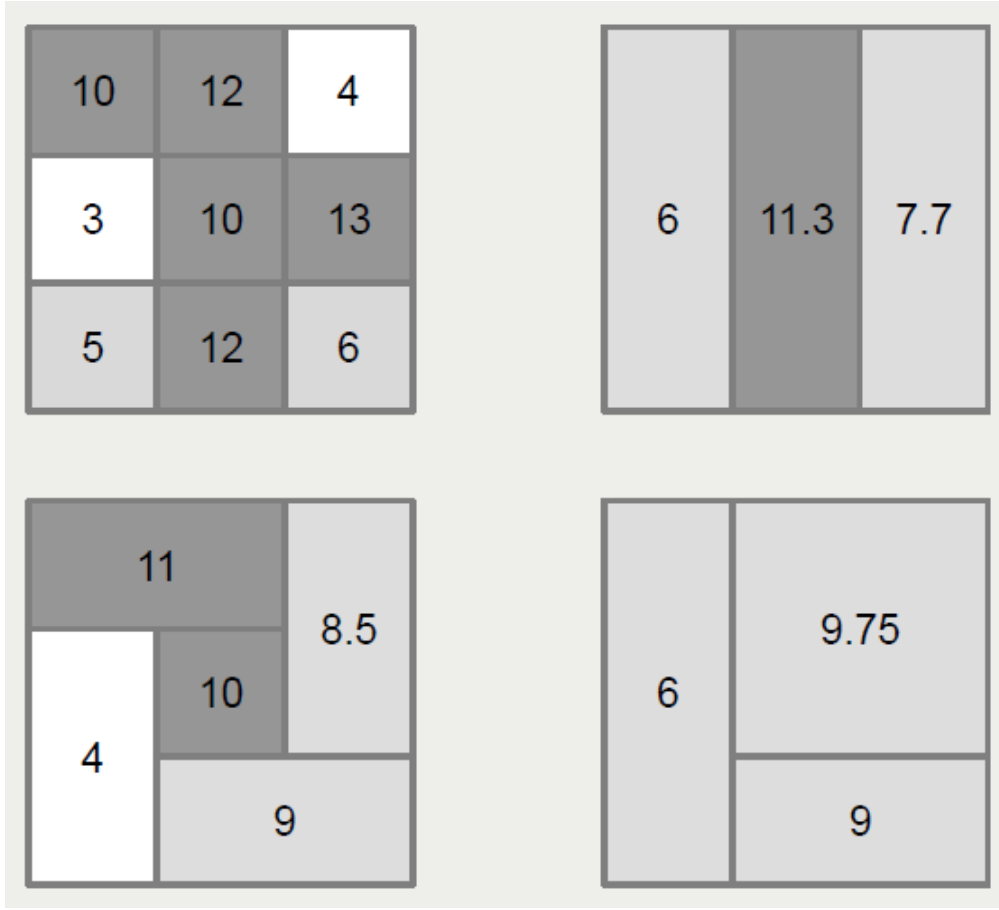
Principles of spatial analysis

- Modifiable areal unit problem
- Ecological and individual fallacy
- Spatial dependence and heterogeneity
- Spatial autocorrelation
- Defining neighbours
- Spatial statistics
- Regression models to take account of spatial effects

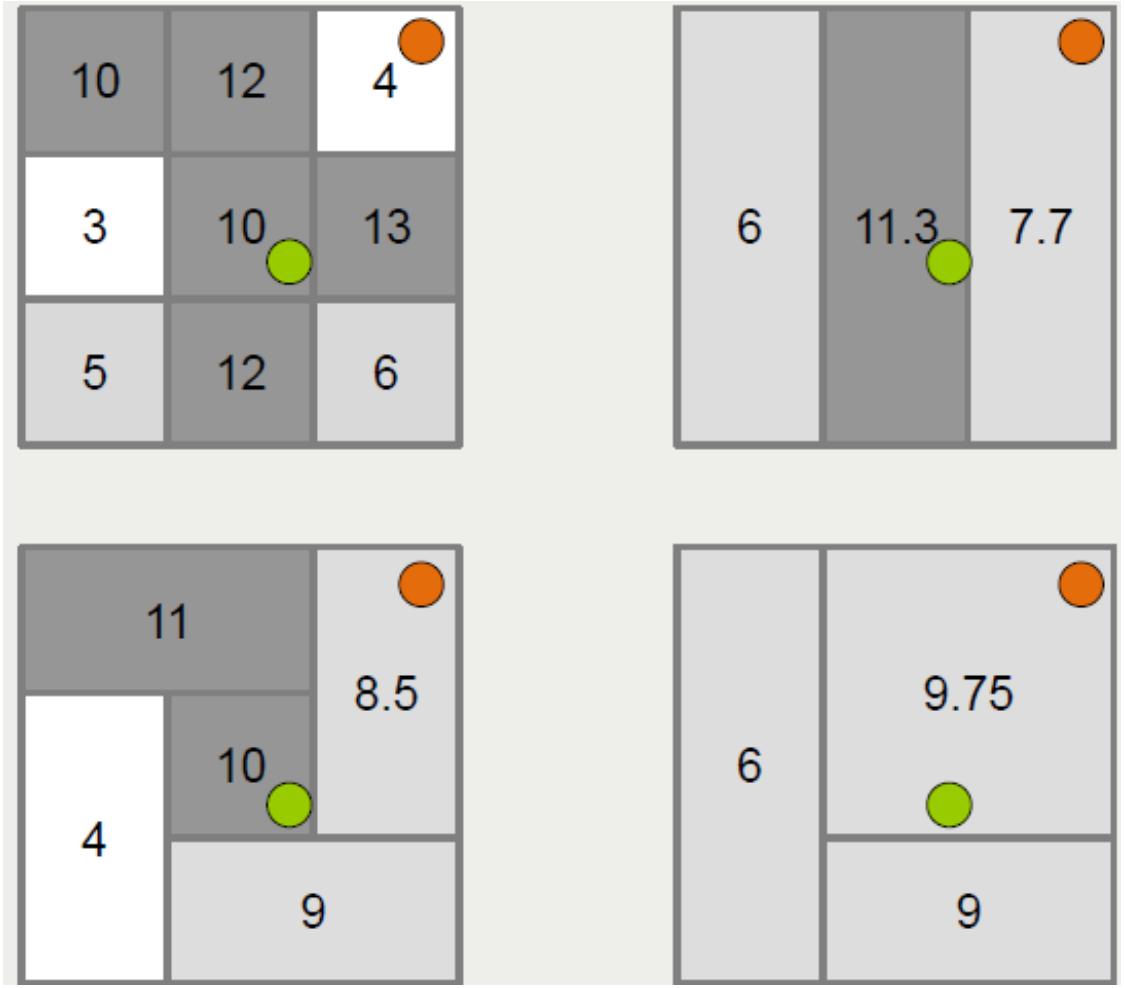
Modifiable areal unit problem

“a problem arising from the imposition of artificial units of spatial reporting on continuous geographical phenomenon resulting in the generation of artificial spatial patterns” (Heywood, 1998)

“States and other forms of socio-political organization […] exercise their power in part through the ability to draw and redraw boundaries inside and around their territories” (Agnew, 2000)



… can introduce error



Spatial dependence and heterogeneity

Spatial Dependence (2nd order spatial process)

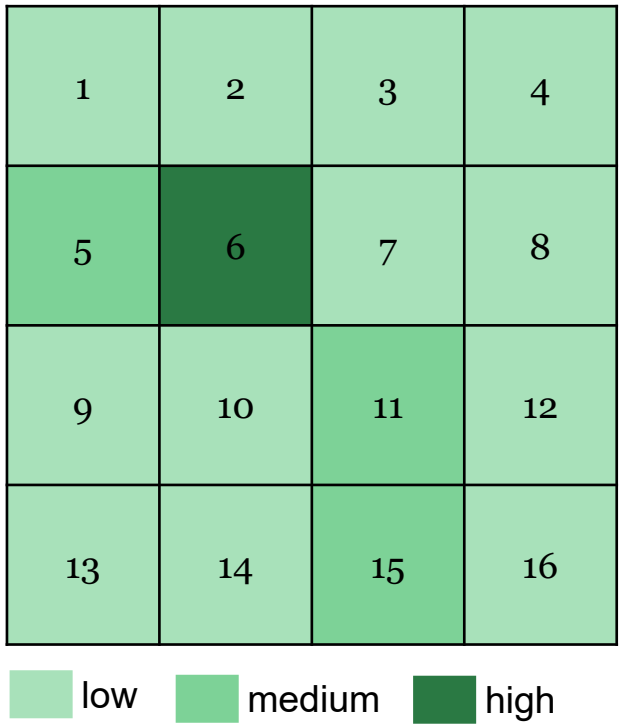
- Anselin (1988): “the existence of a functional relationship between what happens at one point in space and what happens elsewhere.“

Spatial Heterogeneity (1st order spatial process)

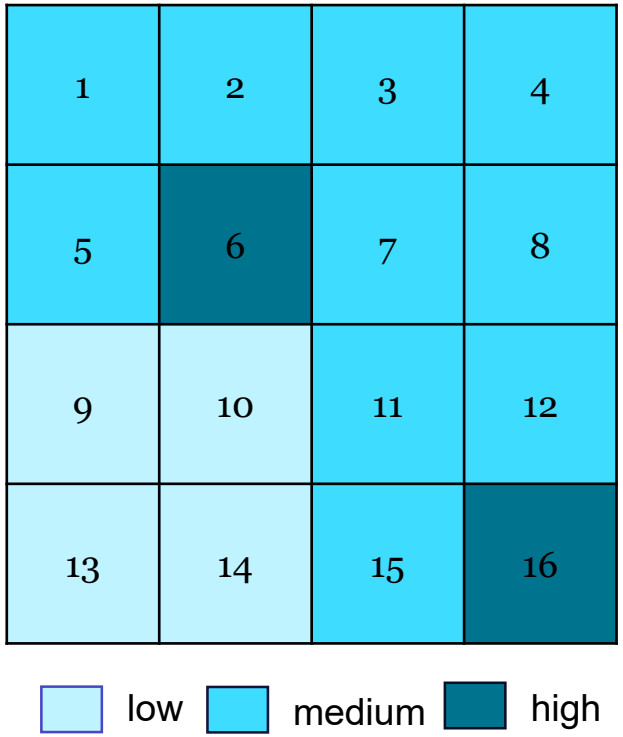
- Not generated by spatial interaction. It refers to variation in relationships over space caused by the uniqueness of location or by spatially autocorrelated omitted variables.

Spatial dependence

Prevalence of burglary

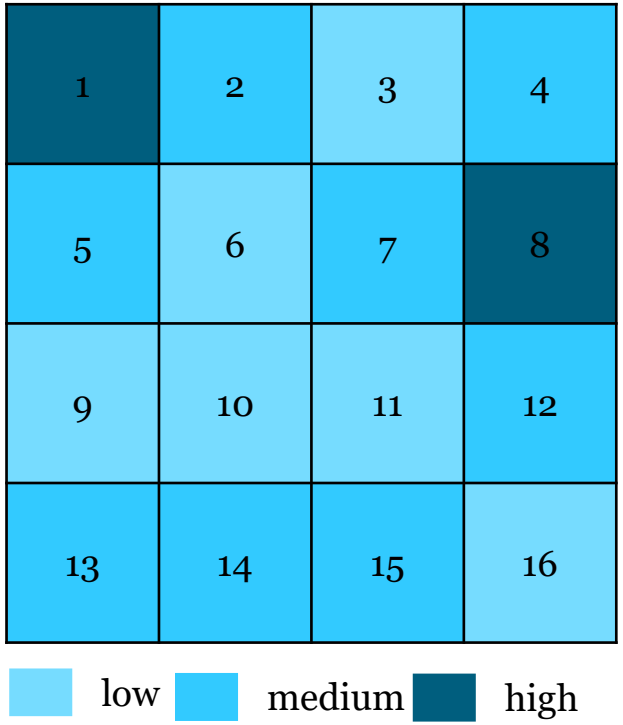


Income per household

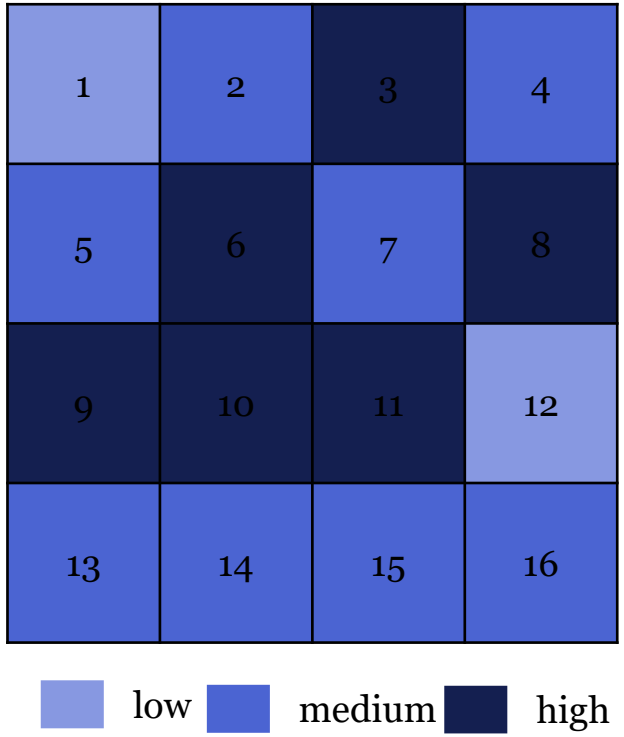


Spatial heterogeneity

Fertility levels



Female labour force participation



Spatial autocorrelation

Spatial Autocorrelation can be

- positive: clustering of high or low values
- negative: clustering of neighbours with high and low values (chessboard pattern)

Effect of Spatial Autocorrelation in the error term is twofold:

- it makes OLS-estimates of the t-test values unreliable
- Positive Spatial Autocorrelation inflates the value of the r^2-statistic, negative Spatial Autocorrelation deflates it.

Defining neighbours

To decide whether there is a spatial effect we need to calculate a neighbourhood weight. This can be base on:

- Contiguity (adjacent spatial units)
- Distance (between centroids of polygons)
- Limited number of nearest neighbours

For this workshop we will use measures of contiguity

Measures of contiguity

First order rook

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

First order queen

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Global measures of spatial autocorrelation

Moran's I (1950)

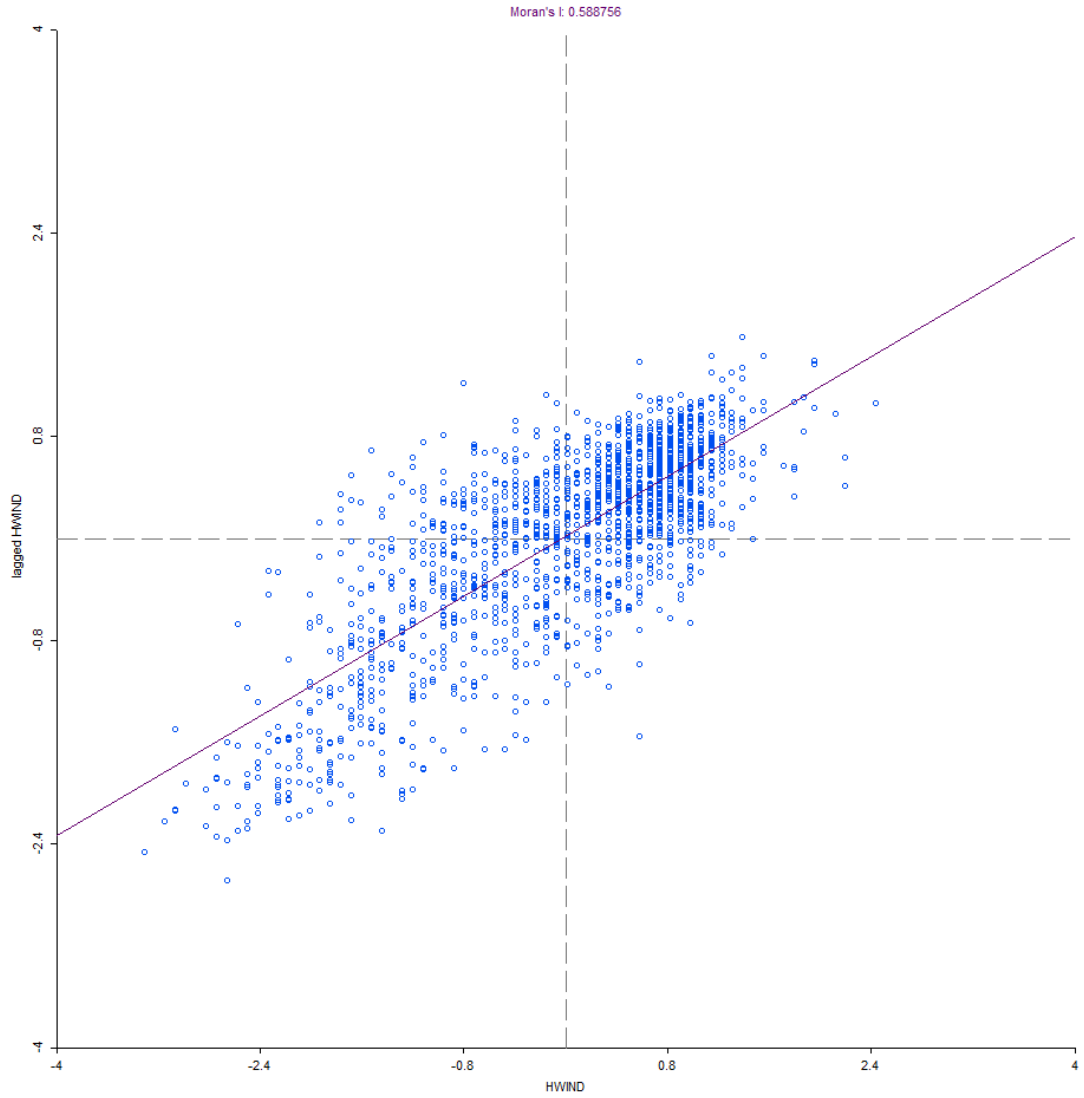
$$I = \left(\frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \right) \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Returns values of

1	**0**	**-1**
Positive	None	Negative

Spatial autocorrelation

Scatterplot of positive spatial autocorrelation



Shows the value of each area on the x-axis and the neighbourhood value on the y-axis

Local indicators of spatial association (LISA)

Local Indicators of Spatial Association (Anselin, 1995)

$$I_i = \frac{z_i}{m_2} \sum_j w_{ij} z_j$$

with

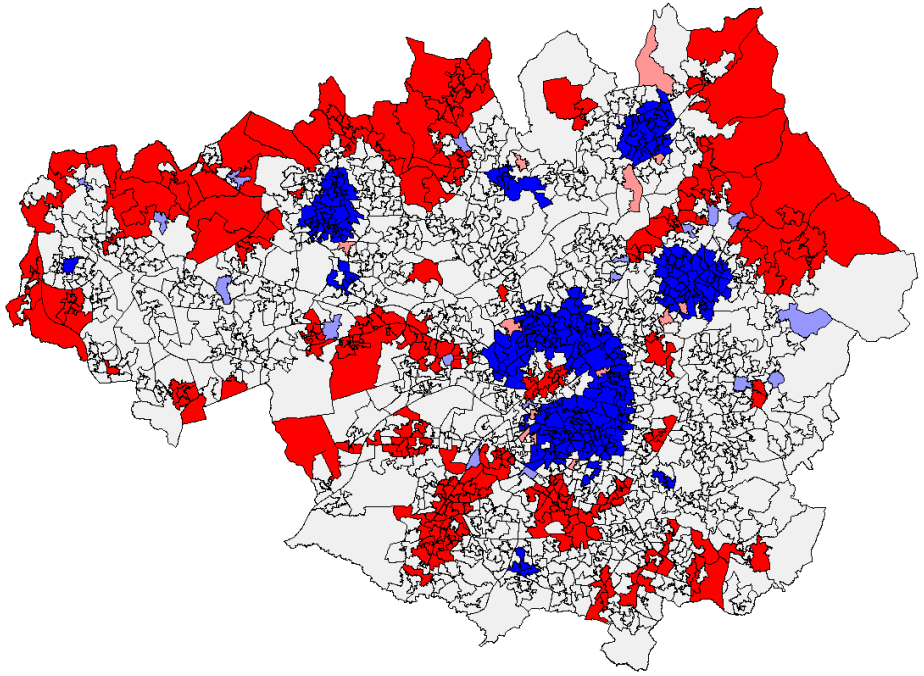
$$m_2 = \frac{\sum_i z_i^2}{n}$$

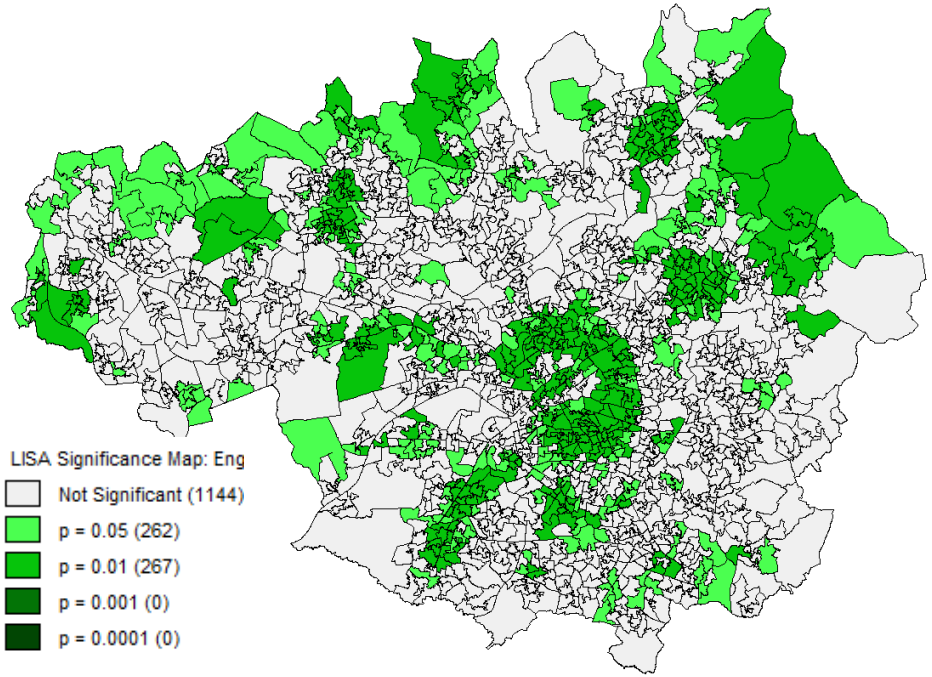
(Variance of variable of interest)

z_i/z_j: deviation of variable of interest from the mean in i/ neighboring regions j of i.

w_{ij}: spatial weights matrix

Local indicators of spatial association LISA maps





White	not significant
Red	High surrounded by High
Blue	Low surrounded by Low
Light blue	Low surrounded by High
Pink	High surrounded by Low

Adapting the basic regression model

The basic linear regression model is

- $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n + e$ (the error term)

To account for spatial autocorrelation we assume that the regression formula may vary between the spatial units we are looking at

- $y_i = b_{0i^*} + b_{1i}x_{1i} + b_{2i}x_{2i} + \ldots + b_{ni}x_{ni} + e_i$
- e.g. survival probabilities in the event of a heart attack will depend on distance to the hospital. We could interact with the severity of the heart attack.

* we may decide to let the intercept vary in our model

Addressing spatial effects in regression

After including all meaningful available variables there may still be a significant result in the Moran's I on the residuals due to:

- spatially correlated omitted variables
- spatially correlated errors in variable measurement
- spatially correlated interactions not considered

To address this we can use a spatial lag or spatial error model.

Spatial lag and spatial error terms

To examine the effects of spatial lag we introduce the lagged value of the dependent variable as a predictor in the regression. The result is called Rho (ρ).

To examine the effects of spatial error we use two error terms. The first accounts for spatially correlated error and the second for spatially uncorrelated error. The result for the spatially correlated error term is called Lambda (λ).

ρ and λ are estimated by maximum likelihood.

Deciding on spatial lag or error?

Theoretical considerations or exploratory data analysis (i.e. Moran's I and LISA maps).

Diagnostic tests giving the Lagrange Multiplier Test and the Robust Lagrange Multiplier Test which identify the existence of lag and/or error.

Lag or error?

The null hypothesis is that there is no spatial autocorrelation: $\rho = 0$ and $\lambda = 0$ in which case we can use a standard regression model

Step 1

- Lagrange Multiplier Lag significant ➔ spatial lag model
- Lagrange Multiplier Error significant ➔ spatial error model

If both tests are significant

- Robust Lagrange Multiplier Lag significant ➔ spatial lag model
  - *(tests for spatial lag where spatial error is significant)*
- Robust Lagrange Multiplier Error significant➔ spatial error model
  - *(tests for spatial error where spatial lag is significant)*

Workshop introduction

We will use Geoda to

- load the data
- produce neighbourhood weights
- explore spatial autocorrelation, clustering and sparsity using LISA maps for the selected variables
- conduct a regression that takes account of spatial autocorrelation and interpret the results