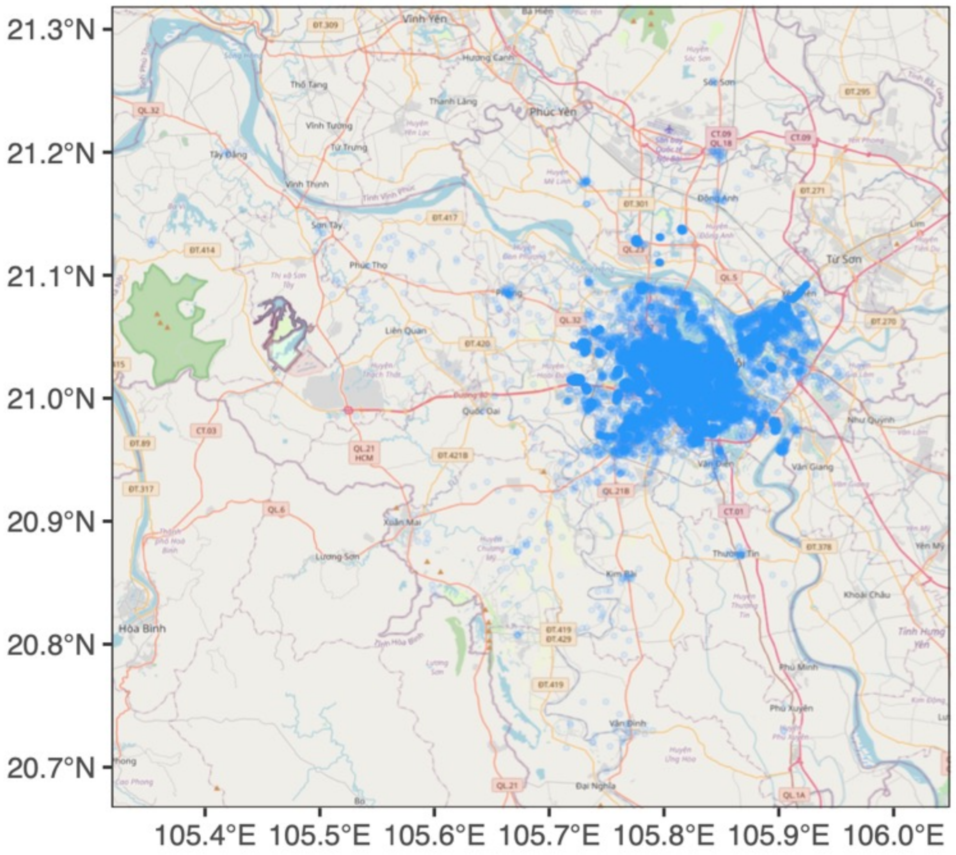
Method Development within the Project

Alexis **Comber**, Nick **Malleson**, Hang **Nguyen Thi Thuy**, Thanh **Bui Quang**, Minh **Kieu**, Phe **Hoang Huu**

University of Leeds, VNU Vietnam Japan University, Hanoi, Vietnam, VNU University of Science, Hanoi, Vietnam, University of Auckland, R&D Consultants

Pre-amble

- Project has generated an amazing dataset
  - ~26,000 respondents, ~140 answers / variables
- Survey data have been used to
  - understand travel **behaviours**
    - by age, occupation, trip purpose, mode & distance
  - analyse travel **attitudes**
    - eg to potential motorbike ban
  - **link** to census data (to add explanatory power)
  - work by Minh Kieu, Nick Malleson and others
- I have used the survey to develop **some novel methods**



Outline

I have developed **novel methods** in 3 main areas

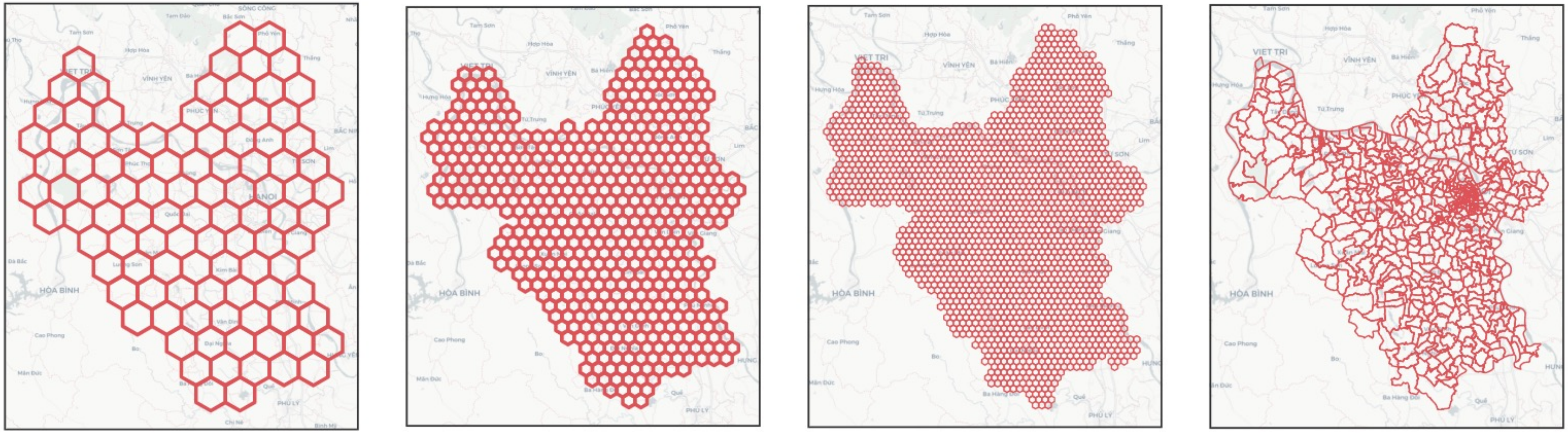
1. Determining optimal aggregation scale
   - handling the Modifiable Areal Unit Problem

2. Multiscale GW Discriminant Analysis
   - Parameter specific, scale local classification

3. Methods for Under-sampling
   - resampling your sample

1. Optimal aggregation scale

- We want to **link** survey data to **other data**
  - e.g. demographic, environmental, social, economic, etc
- BUT other data are reported at a various **different scales**
- Key question: which scale is **appropriate**?

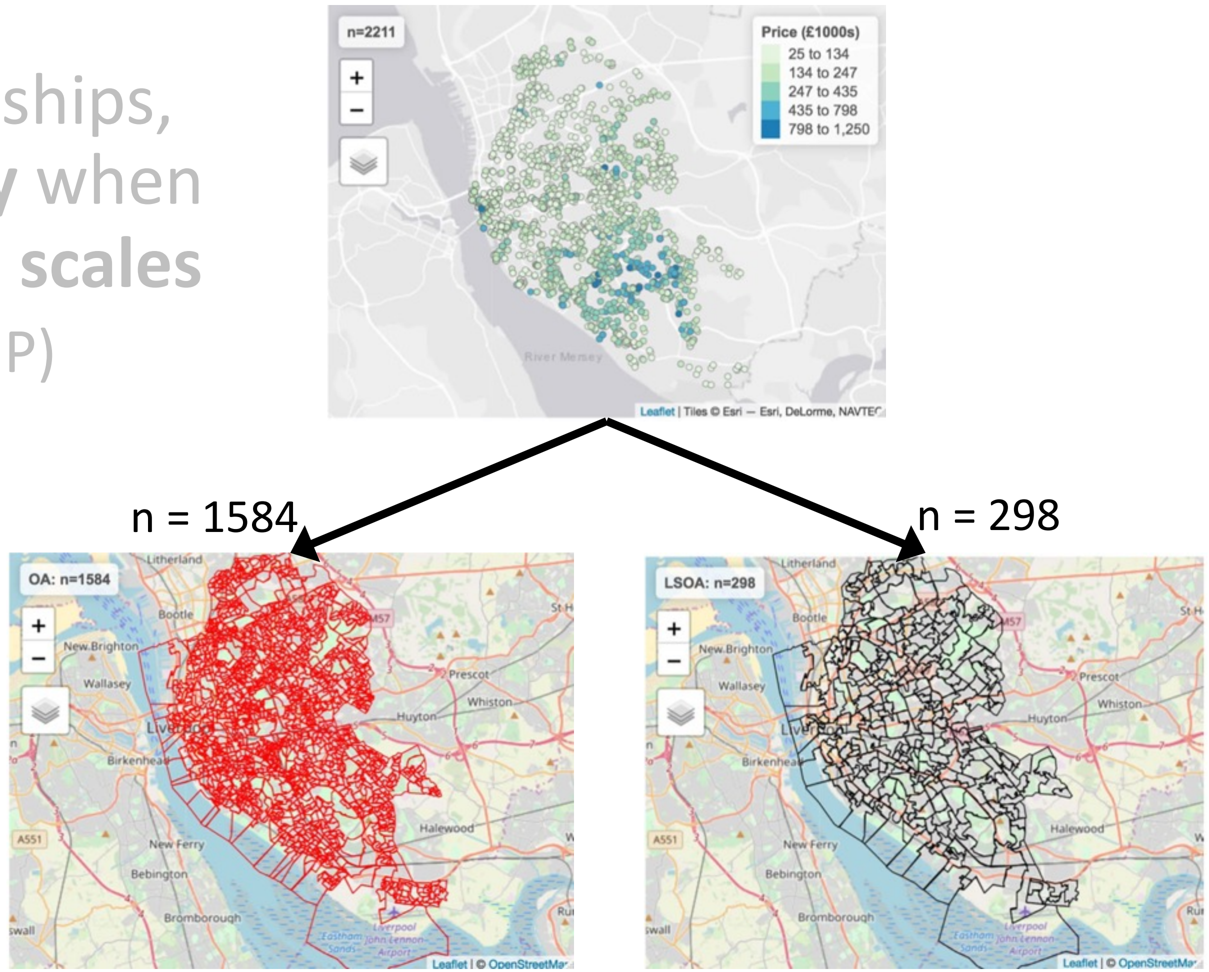


1. Optimal aggregation scale

- Why is this a key question?
- Simply because: **statistical** relationships, trends and correlations trends **vary** when data are **aggregated** over **different scales**
  - Modifiable Areal Unit Problem (MAUP)
  - Known in Geography for a long time

1. Optimal aggregation scale

- Why is this a key question?
- Simply because: **statistical** relationships, trends and correlations trends **vary** when data are **aggregated** over **different scales**
  - Modifiable Areal Unit Problem (MAUP)
  - Known in Geography for a long time

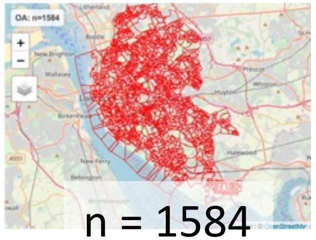


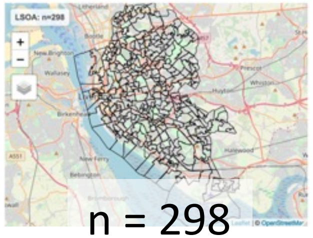
- House price example
  - 2 **scales** of aggregation
  - create 2 **models** of house price with demographic data
  - counts are the **same** but spread **over different areas**

1. Optimal aggregation scale

- Why is this a key question?
- Simply because: **statistical** relationships, trends and correlations trends **vary** when data are **aggregated** over **different scales**
  - Modifiable Areal Unit Problem (MAUP)
  - Known in Geography for a long time

- House price example
  - 2 **scales** of aggregation
  - create 2 **models** of house price with demographic data
  - counts are the **same** but spread **over different areas**





Covariate	OA	LSOA
(Intercept)	33.986	−43.505
gs_area	0.875	0.412
u25	1.977	2.882
u45	0.714	1.962
u65	5.368	5.543
o65	3.481	6.967
unmplyd	−8.246	−10.850

1. Optimal aggregation scale

- Recent optimising Ecosystem Service
- Suggested that **best aggregation scale** can be determined identifying scales at which the **processes** are **stable**
- Find stability of **variances**, **covariances** and **higher moments** in context of the subsequent data analyses
  - i.e. variance etc within intended statistical model
- Evaluated **6 variances** to find **optimal** aggregation **scale**


land MDPI

Article

The Importance of Scale and the MAUP for Robust Ecosystem Service Evaluations and Landscape Decisions

Alexis Comber [1,*] and Paul Harris [2]

1 School of Geography, University of Leeds, Leeds LS2 9JT, UK
2 Sustainable Agriculture Sciences, Rothamsted Research, North Wyke, Okehampton EX20 2SB, UK; paul.harris@rothamsted.ac.uk
* Correspondence: a.comber@leeds.ac.uk

Abstract: Spatial data are used in many scientific domains including analyses of Ecosystem Services (ES) and Natural Capital (NC), with results used to inform planning and policy. However, the data spatial scale (or support) has a fundamental impact on analysis outputs and, thus, process understanding and inference. The Modifiable Areal Unit Problem (MAUP) describes the effects of scale on analyses of spatial data and outputs, but it has been ignored in much environmental research, including evaluations of land use with respect to ES and NC. This paper illustrates the MAUP through an ES optimisation problem. The results show that MAUP effects are unpredictable and nonlinear, with discontinuities specific to the spatial properties of the case study. Four key recommendations are as follows: (1) The MAUP should always be tested for in ES evaluations. This is commonly performed in socio-economic analyses. (2) Spatial aggregation scales should be matched to process granularity by identifying the aggregation scale at which processes are considered to be stable (stationary) with respect to variances, covariances, and other moments. (3) Aggregation scales should be evaluated along with the scale of decision making (e.g., agricultural field, farm holding, and catchment). (4) Researchers in ES and related disciplines should up-skill themselves in spatial analysis and core paradigms related to scale to overcome the scale blindness commonly found in much research.

Keywords: spatial support; land use; genetic algorithm

check for updates

Citation: Comber, A.; Harris, P. The Importance of Scale and the MAUP for Robust Ecosystem Service Evaluations and Landscape Decisions. *Land* **2022**, *11*, 399. https://doi.org/10.3390/land11030399

Academic Editor: Benjamin Burkhard

Received: 4 February 2022
Accepted: 3 March 2022
Published: 9 March 2022

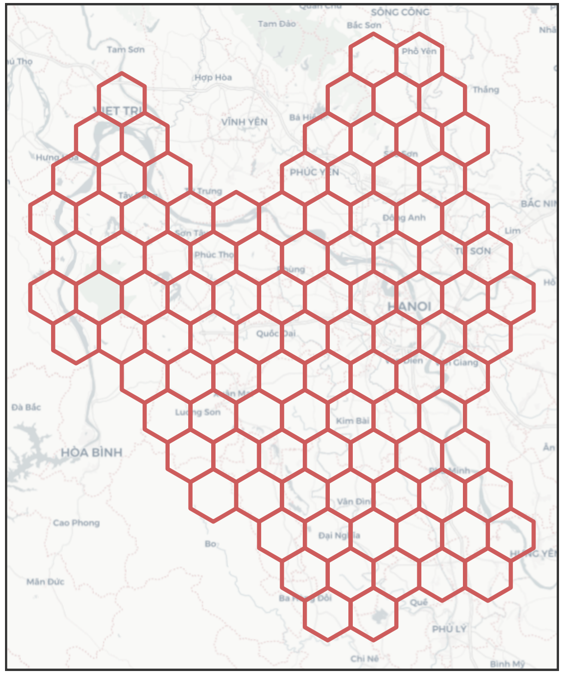
Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

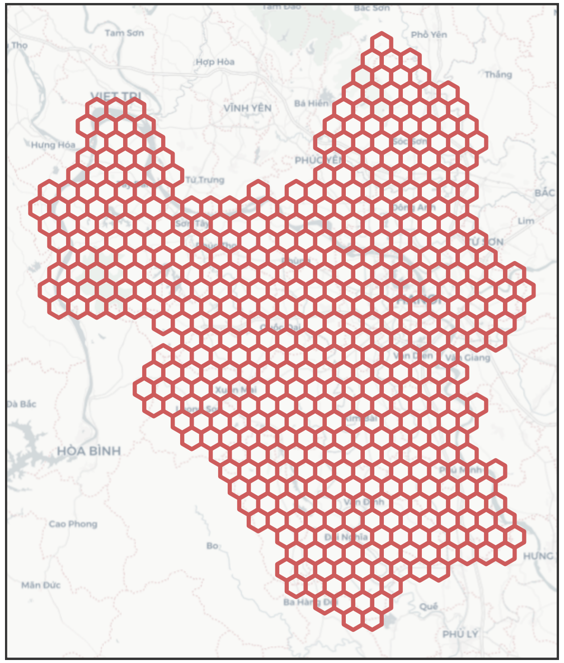
1. Introduction

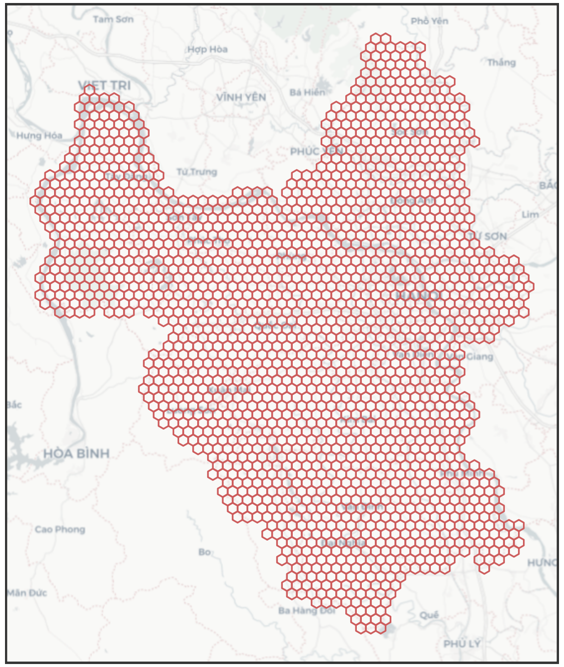
Spatial scale—the spatial scale of measurement or in geostatistics, spatial support—has huge impacts on spatial analyses, model outputs and, thus, process understanding. The impacts of scale are well understood in quantitative social sciences to the point where any research in this domain is expected to be able to describe the impacts of their choice of aggregation scale on their analysis, results and derived understanding [1]. However, little land use research and related studies of the goods and services provided by land based systems such as agricultural production, biodiversity, flood protection and other elements related to concepts of Natural Capital (NC) and Ecosystem Service (ES) has considered the impacts of spatial data scales on their analyses. In fact, there are many examples of blindness to the analytical impacts of scale, where processes captured at one scale are applied to another without considering the inferential impacts of these differing scales. For example, Spake et al. [2] applied forest models captured over stands (a specific spatial unit in forestry) to 10 km gridded data and Finch et al. [3] used a nutrient delivery model constructed over a 50 m grid to make inferences on 1 km squares. Such scale mismatches affect the robustness of the results and have implications for the reliability of any policy or planning recommendations arising from them. This paper seeks to highlight the importance of considering and evaluating the impact of scale using a hypothetical ES optimisation problem. In so doing, it addresses this key methodological gap in current approaches to

1. Optimal aggregation scale

- To find **optimal** aggregation **scale**
  - Created aggregation grids at **different scales** ($n = 80$)
  - Survey data **aggregated over grids** and a statical **model** created



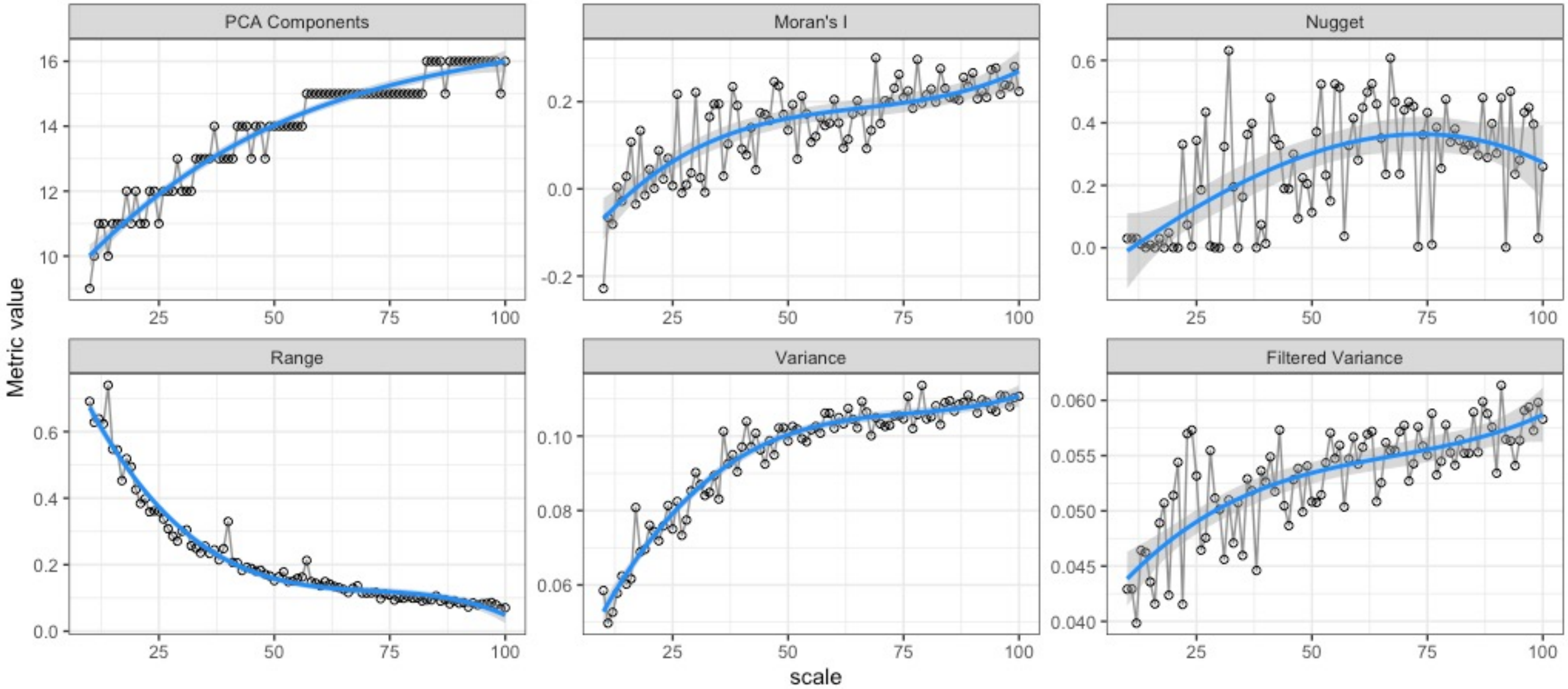




1. Optimal aggregation scale

- Evaluated 6 variances to find optimal aggregation scale
  - **Variance** of target variable
  - **Filtered Variance** (eg > 5 respondents)
  - model residual **Variogram**
    - the Nugget effect from a linear model fitted with a spatially autocorrelated error term
  - residual variogram correlation **Range**
  - number of PCA Components that explain 80% of variation
  - **Moran's I** (spatial clustering) model residuals

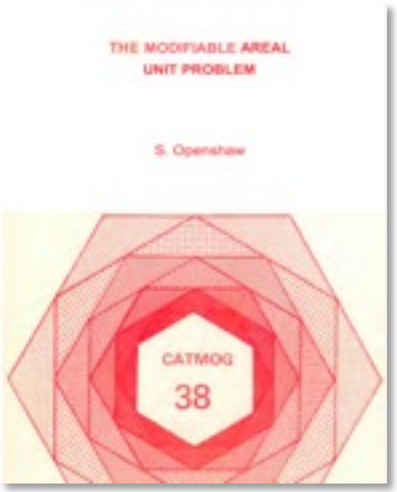
1. Optimal aggregation scale

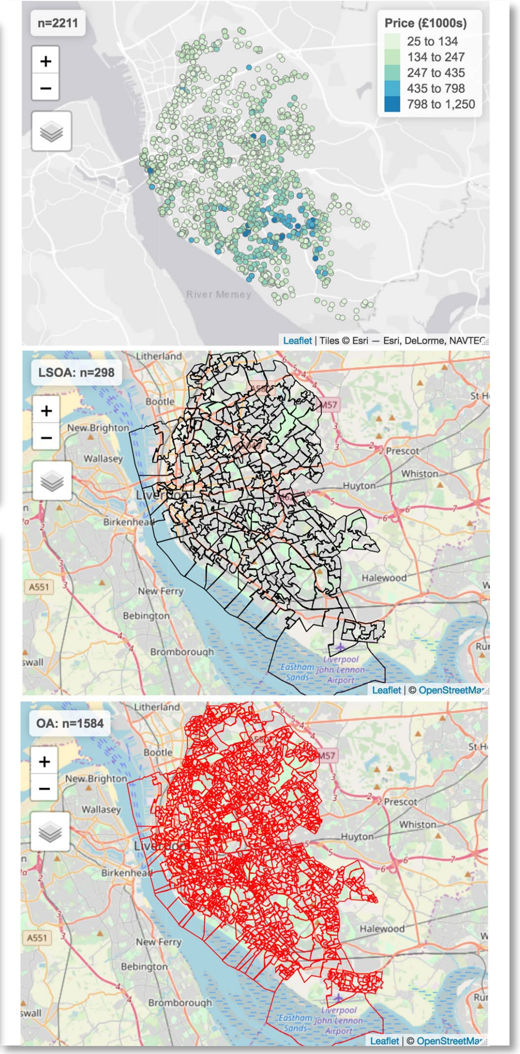


- Indicates **optimum** aggregation scale of 50-70 (**2km^2 to 1km^2**)
  - Some stability (Variance, PCA, Morans' I, Filtered Variance), some highly variable (Nugget)

1. Optimal aggregation scale

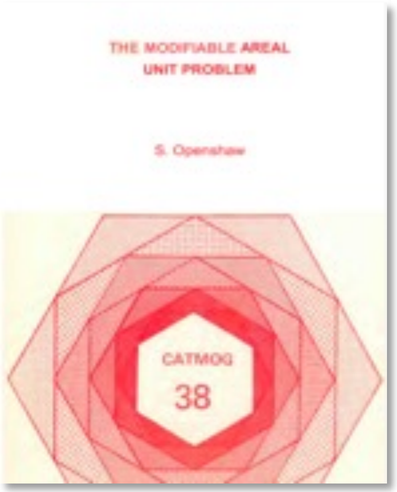
- **Why** is this a **key question**?
- Modifiable Areal Unit Problem (MAUP)
  - Simply that **statistical** relationships, trends and correlations trends **vary** when data are **aggregated** over **different scales**
- Scale **changes** our process understanding
  - model **outputs vary** when constructed from data aggregated over different areas

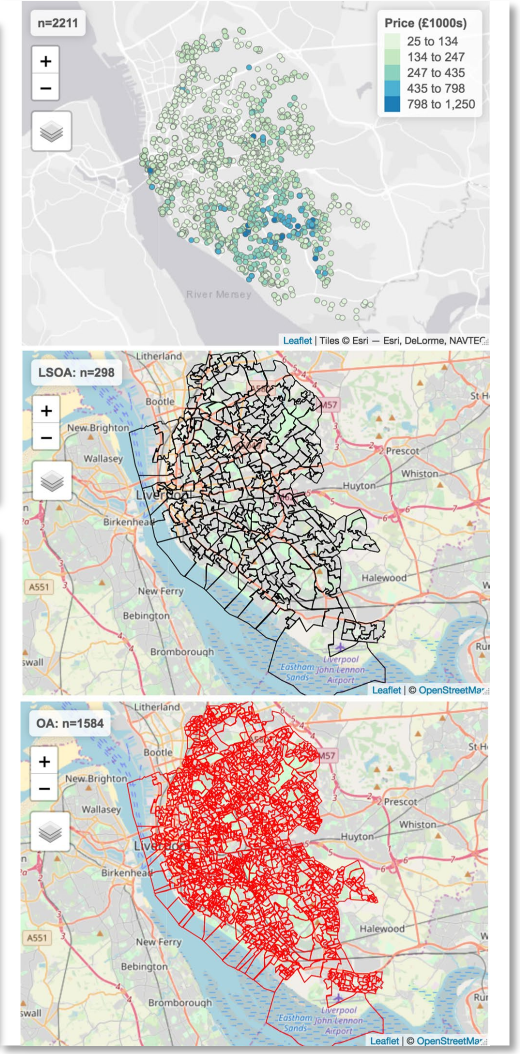




1. Optimal aggregation scale

- **Why** is this a **key question**?
- Modifiable Areal Unit Problem (MAUP)
  - Simply that **statistical** relationships, trends and correlations trends **vary** when data are **aggregated** over **different scales**
- Scale **changes** our process understanding
  - model **outputs vary** when constructed from data aggregated over different areas
- MAUP applies to **ALL** data
  - remember **all data are spatial**
    - collected some-**where**
- implications for Data Science, AI, ML etc





Outline

I have developed **novel methods** in 3 main areas

1. Determining optimal aggregation scale
   - handling the Modifiable Areal Unit Problem

2. Multiscale GW Discriminant Analysis
   - Parameter specific, scale local classification

3. Methods for Under-sampling
   - resampling your sample

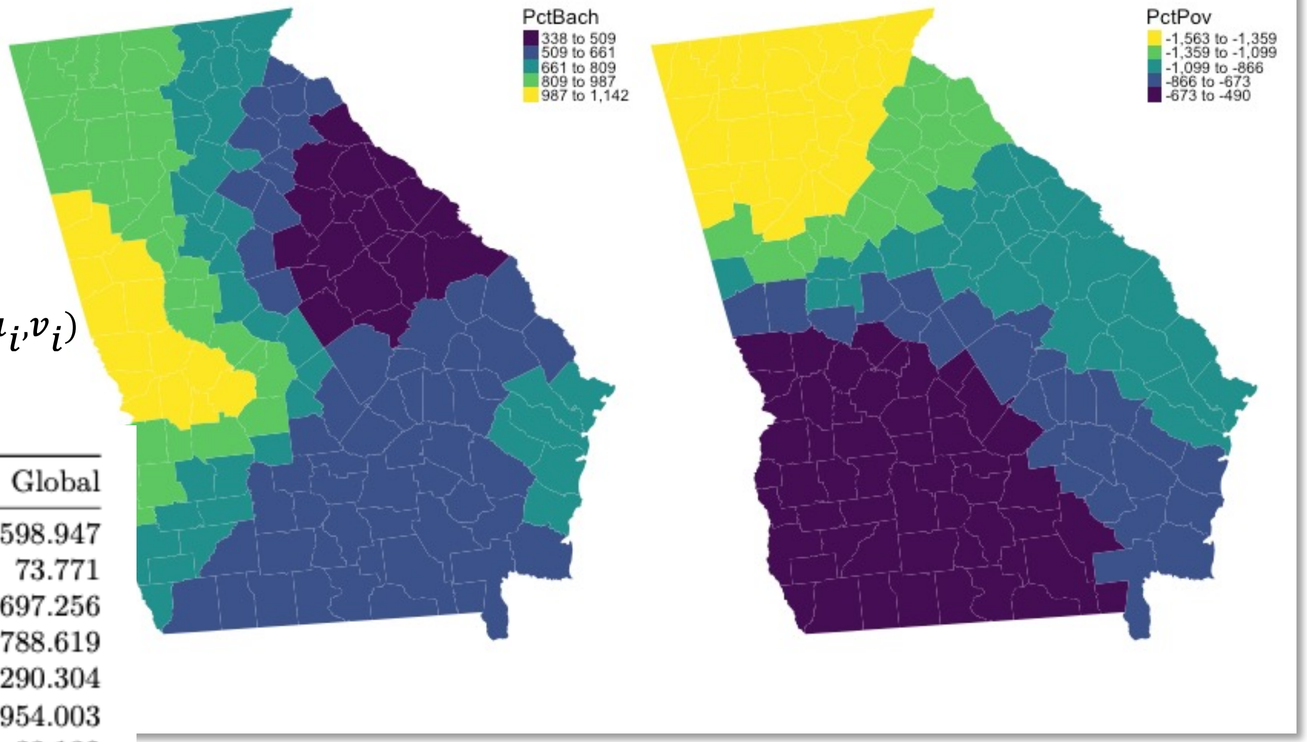
Geographically Weighted models

- Create many **local** models (ie local **coefficients**)
- These **vary spatially**
- For example **Regression**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots \beta_n x_n$$

$$y = \beta_{0(u_i,v_i)0} + \beta_1 x_{1(u_i,v_i)} + \beta_2 x_{2(u_i,v_i)} \dots \beta_n x_{n(u_i,v_i)}$$

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Global
Intercept	43565.247	48436.098	51904.667	52557.332	56896.243	61630.683	52598.947
PctRural	25.725	53.867	77.671	78.477	105.007	138.930	73.771
PctBach	338.066	596.111	672.064	710.142	855.772	1141.588	697.256
PctEld	-1155.696	-948.323	-856.672	-859.909	-787.134	-525.598	-788.619
PctFB	-2931.360	-2100.136	-1268.050	-1410.017	-750.505	151.845	-1290.304
PctPov	-1562.544	-1243.883	-874.675	-931.347	-593.802	-489.711	-954.003
PctBlack	-196.510	-27.919	30.547	17.437	77.950	160.202	33.132



Multiscale GWDA (MGWDA)

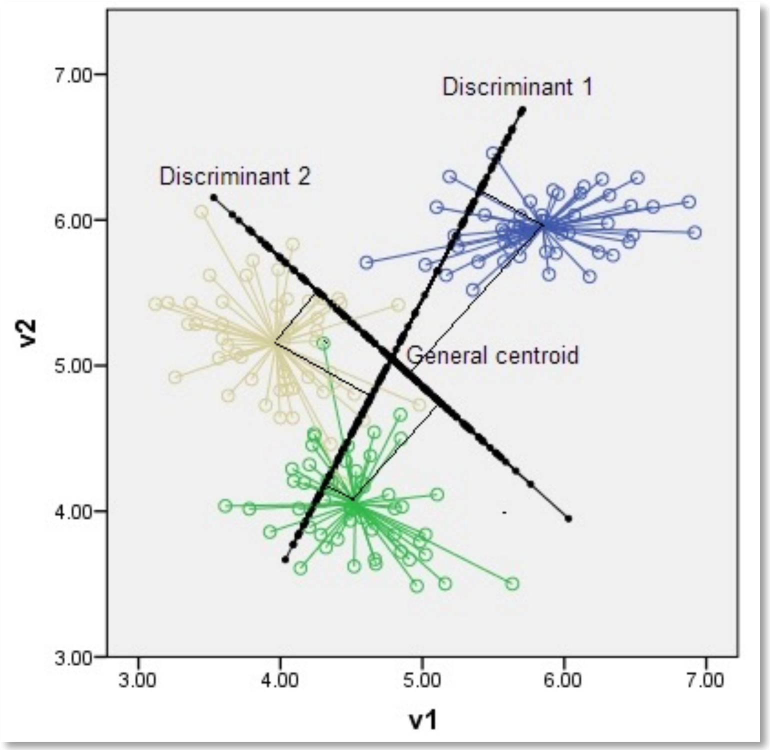
- Ordinary Discriminant Analysis (DA)
  - used to predict **class membership**
    - alternative to multinomial logistic regression
  - very popular in **machine learning** communities
    - used as information learning technique eg pattern recognition.

Multiscale GWDA (MGWDA)

- Ordinary Discriminant Analysis (DA)
  - used to predict **class membership**
    - alternative to multinomial logistic regression
  - very popular in **machine learning** communities
    - used as information learning technique eg pattern recognition.
- Conceptually, in a DA
  - data are considered to be **drawn from** different populations
  - for each class

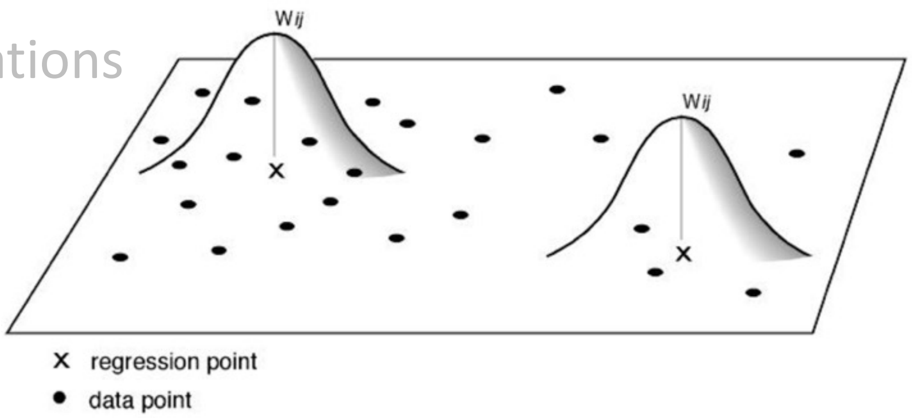
Multiscale GWDA (MGWDA)

- Ordinary Discriminant Analysis (DA)
  - used to predict **class membership**
    - alternative to multinomial logistic regression
  - very popular in **machine learning** communities
    - used as information learning technique eg pattern recognition.
- Conceptually, in a DA
  - data are considered to be **drawn from** different populations
  - for each class
- DA generates **Discriminant Functions**
  - used to generate class membership probabilities



Multiscale GWDA (MGWDA)

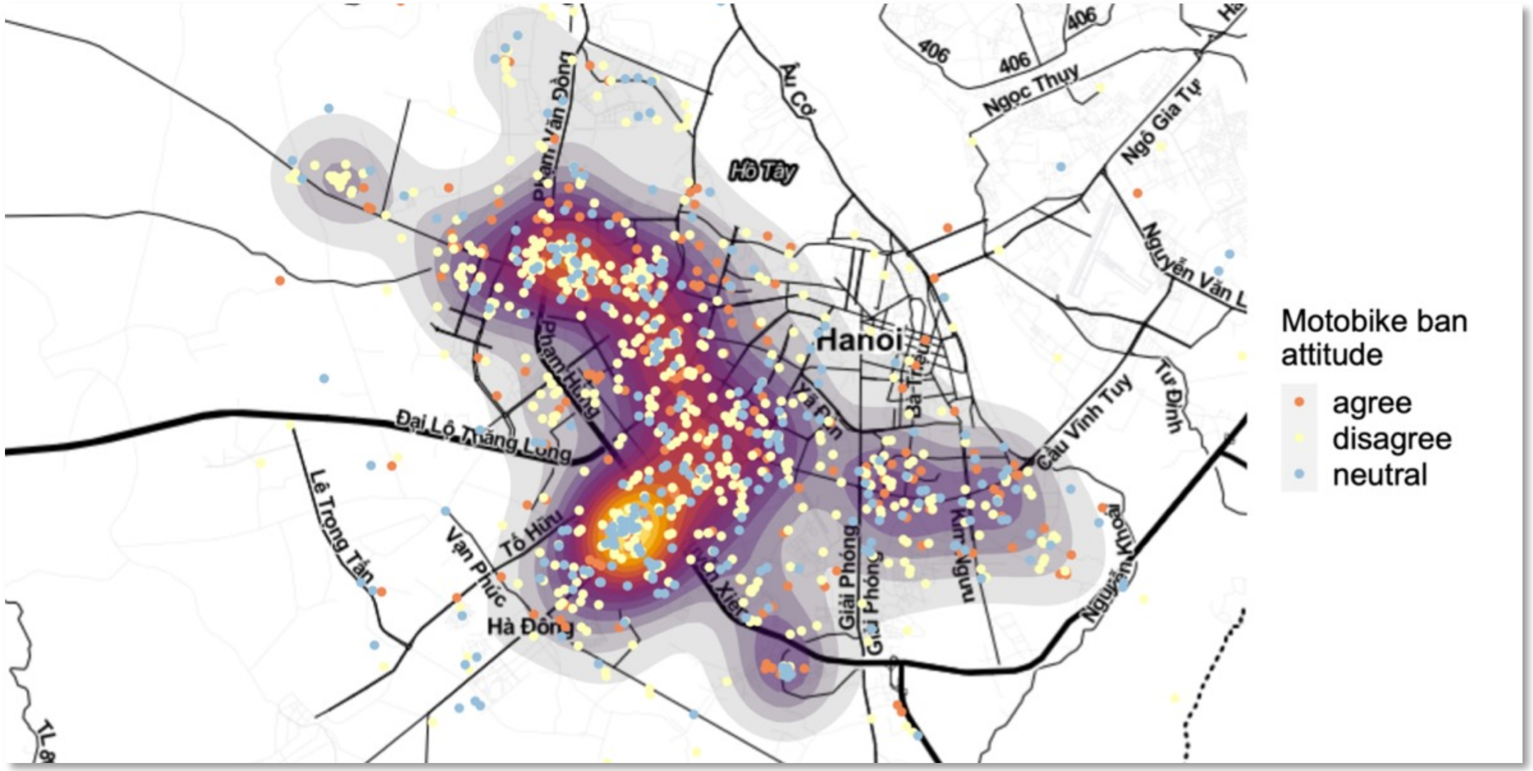
- Ordinary Discriminant Analysis (DA)
  - used to predict **class membership**
    - alternative to multinomial logistic regression
  - very popular in **machine learning** communities
    - used as information learning technique eg pattern recognition.
- Conceptually, in a DA
  - data are considered to be **drawn from** different populations
  - for each class
- DA generates **Discriminant Functions**
  - used to generate class membership probabilities



- DA under a **Multiscale GW framework**
  - multiple **local models** (kernel / moving window)
  - determine optimal kernel **size** for each variable → scale of **relationship**

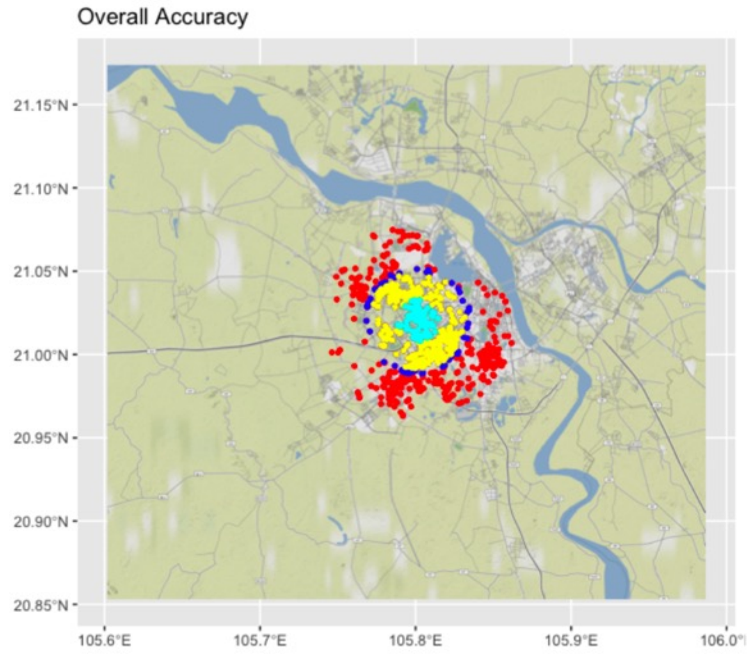
Multiscale GWDA (MGWDA)

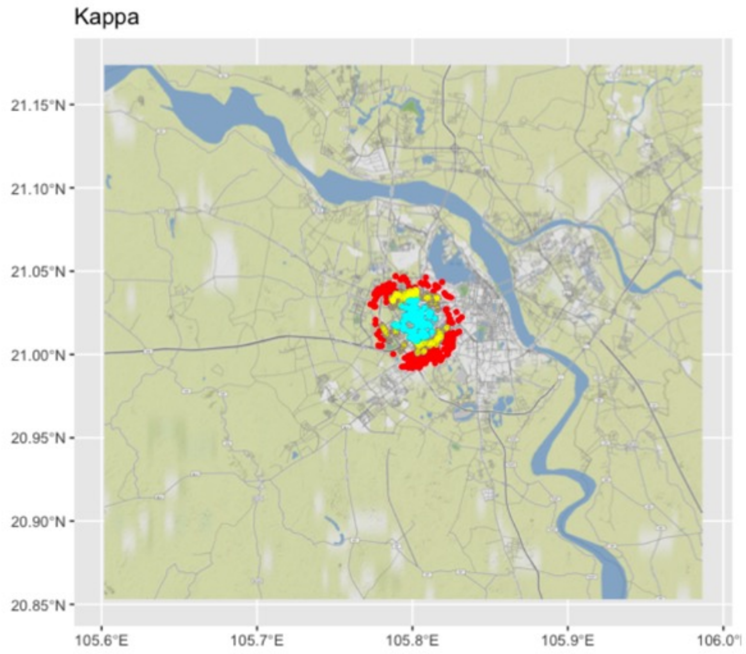
- Survey: attitudes to proposed **motorbike ban**
- MGWDA model against age, gender, trip purpose, trip distance



Multiscale GWDA (MGWDA)

- MGWDA of ban attitudes
  - Shows different **scales of process** and **statistical relationship**
    - Some **highly localised**, others **near global**
  - But depends on evaluation (Overall Accuracy and Kappa)





<u>Percentage of data included in **each local model**</u>

Variable	**Overall**	**Kappa**
Gender (red)	80%	40%
Purpose (blue)	50%	20%
Age (yellow)	40%	10%
Distance (cyan)	20%	10%

Shows the **varying** scales of **influence** of different factors

Multiscale GWDA (MGWDA)

- MSGWDA
  - improves classification accuracy
    - From standard DA to Geographically Weighted DA to Multi-scale GWDA
  - indicates **variation** in **scales of relationship** between inputs & outcome
    - the **gender** variable tends towards the global (**same everywhere**)
    - the **trip purpose**, **age** and **distance** highly localised in their effect (**locally varying**)

Multiscale GWDA (MGWDA)

- MSGWDA
  - improves classification accuracy
    - From standard DA to Geographically Weighted DA to Multi-scale GWDA
  - indicates **variation** in **scales of relationship** between inputs & outcome
    - the **gender** variable tends towards the global (**same everywhere**)
    - the **trip purpose**, **age** and **distance** highly localised in their effect (**locally varying**)
- Has **policy** implications
  - potential for local **targeted** strategies / policy for **specific groups**
  - and for what groups a **one-size-fits all policy** will work

Multiscale GWDA (MGWDA)

- MSGWDA
  - improves classification accuracy
    - From standard DA to Geographically Weighted DA to Multi-scale GWDA
  - indicates **variation** in **scales of relationship** between inputs & outcome
    - the **gender** variable tends towards the global (**same everywhere**)
    - the **trip purpose**, **age** and **distance** highly localised in their effect (**locally varying**)
- Has **policy** implications
  - potential for local **targeted** strategies / policy for **specific groups**
  - and for what groups a **one-size-fits all policy** will work
- This **local process** understanding is a key advantage of **spatially varying** statistical models – I work with these a lot!!

Outline

I have developed **novel methods** in 3 main areas

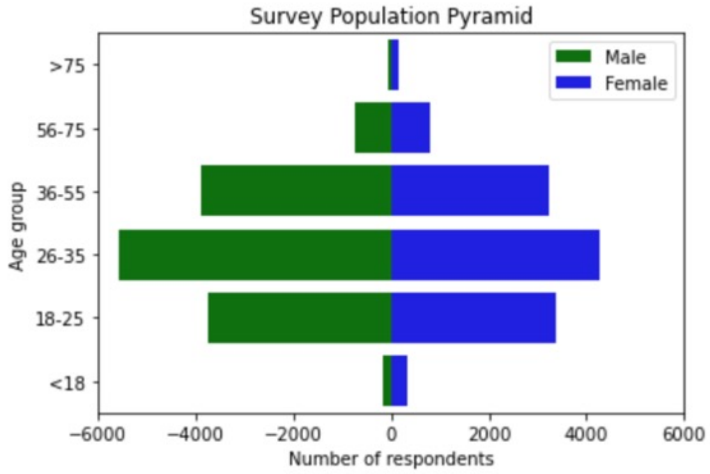
1. Determining optimal aggregation scale
   - handling the Modifiable Areal Unit Problem

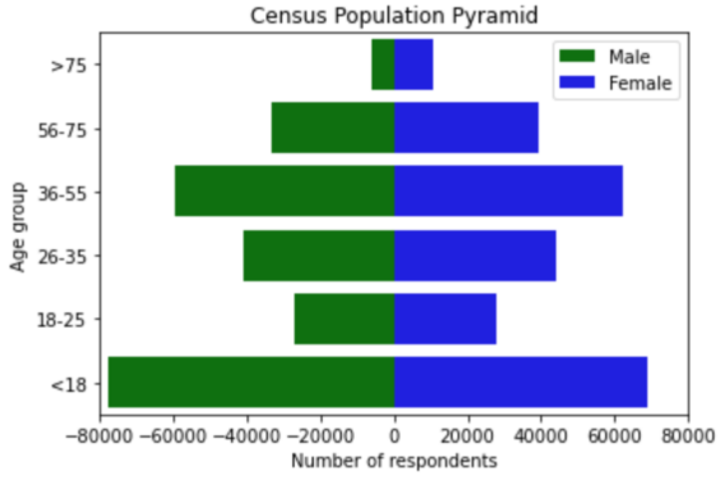
2. Multiscale GW Discriminant Analysis
   - Parameter specific, scale local classification

3. Methods for Under-sampling
   - resampling your sample

Methods for Under-sampling

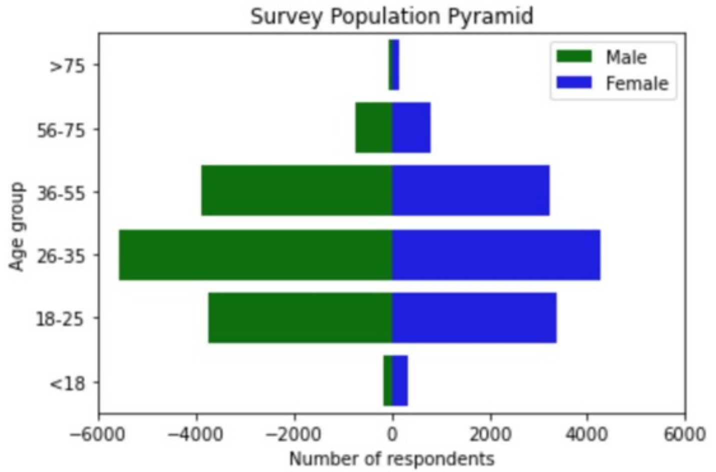
- Project survey of attitudes and behaviours
  - ~26,000 respondents, ~140 answers / variables
- But **bias** in respondent **demographics**
  - **difficult** to unpick trends from survey
  - and to construct **robust statistical** models
- Nick described Propensity Matching
  - for **Up-scaling** to link to Census data

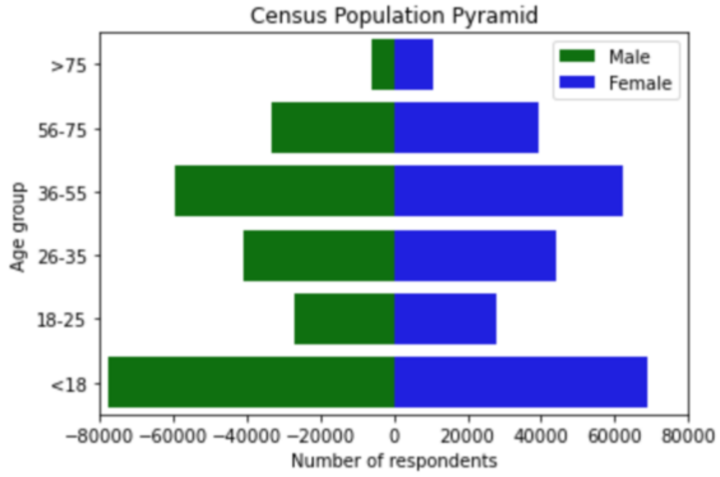




Methods for Under-sampling

- Project survey of attitudes and behaviours
  - ~26,000 respondents, ~140 answers / variables
- But **bias** in respondent **demographics**
  - **difficult** to unpick trends from survey
  - and to construct **robust statistical** models
- Nick described Propensity Matching
  - for **Up-scaling** to link to Census data
- Here I want to focus on **Down-scaling**
  - i.e. **resample** the survey
  - then analyse the survey data





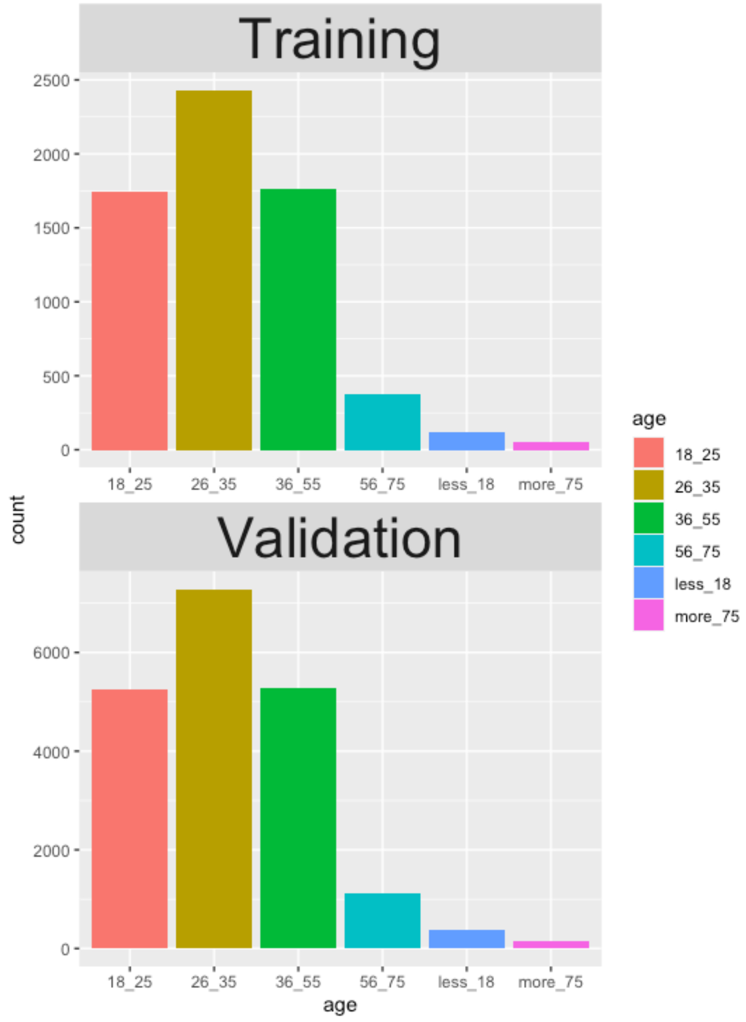
Methods for Under-sampling

- Methods exist for creating **data subsets** with **same distributions**
  - e.g. for **Training** and **Validation** splits

Methods for Under-sampling

- Methods exist for creating **data subsets** with **same distributions**
  - e.g. for **Training** and **Validation** splits

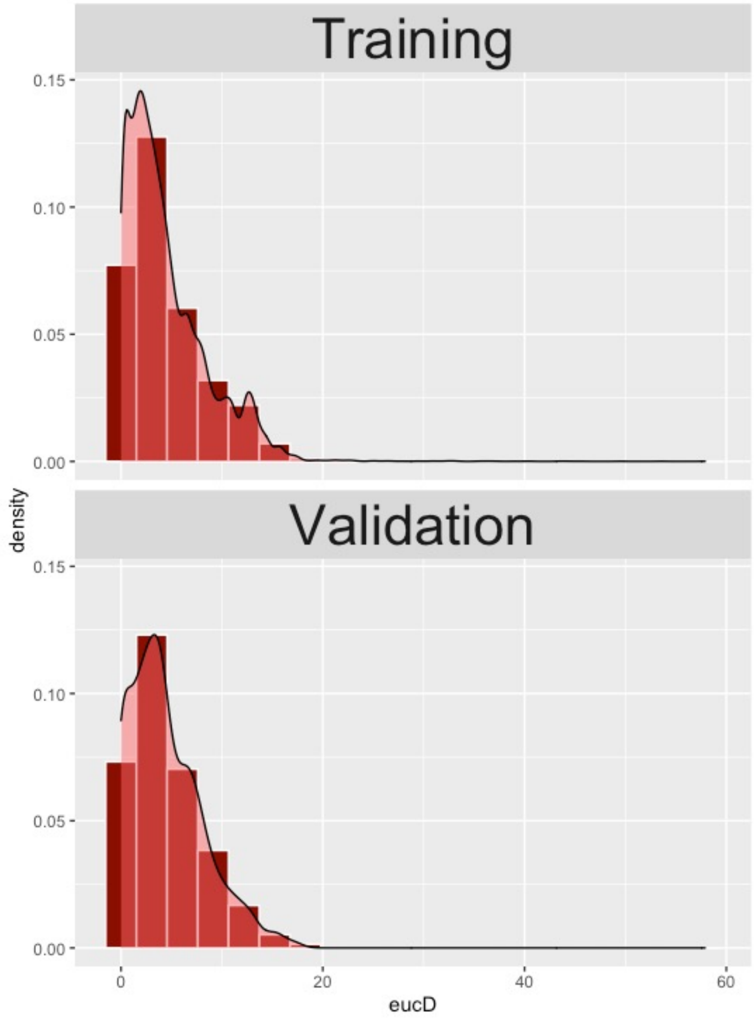
- These focus on a single **target** variable (y)
  - Example: Age (categorical)



Methods for Under-sampling

- Methods exist for creating **data subsets** with **same distributions**
  - e.g. for **Training** and **Validation** splits

- These focus on a single **target** variable (y)
  - Example: Age (categorical)
  - Example: Trip Distance (continuous)

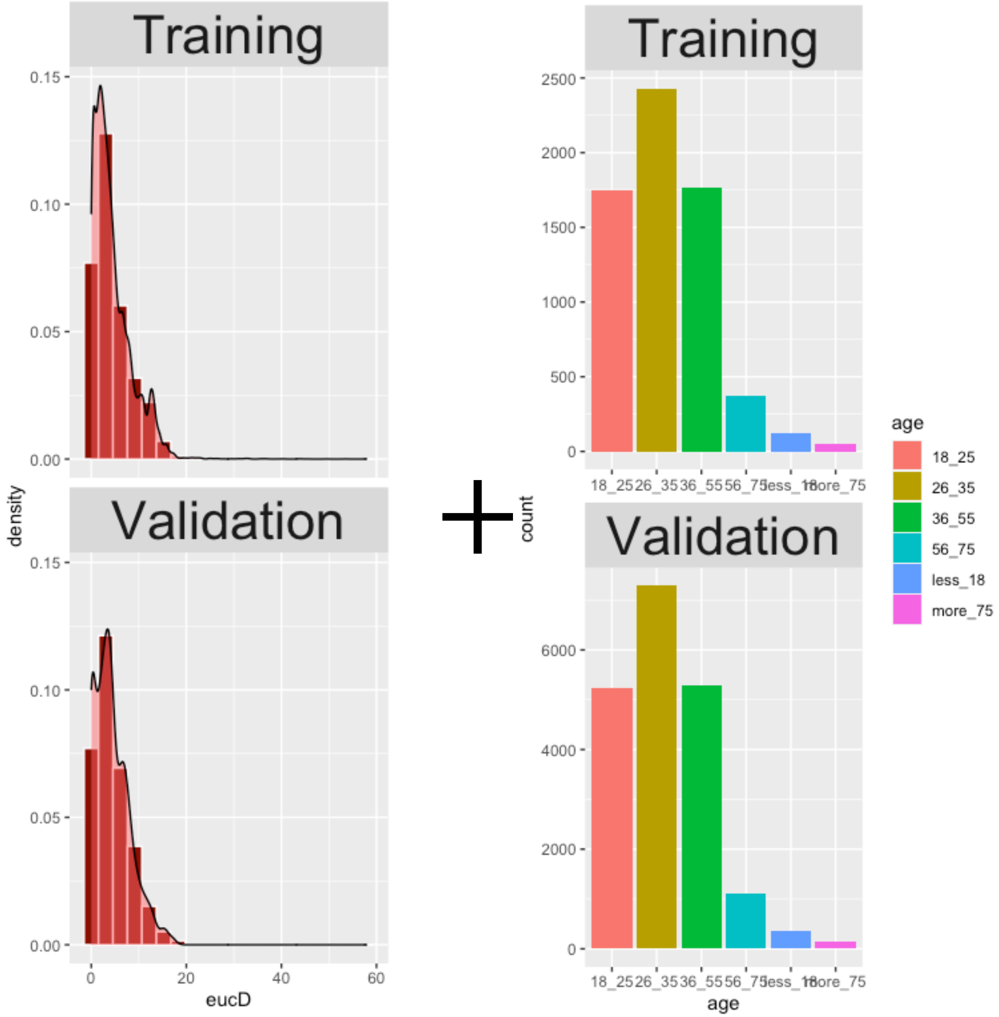


Methods for Under-sampling

- Methods exist for creating **data subsets** with **same distributions**
  - e.g. for **Training** and **Validation** splits

- These focus on a single **target** variable (y)
  - Example: Age (categorical)
  - Example: Trip Distance (continuous)

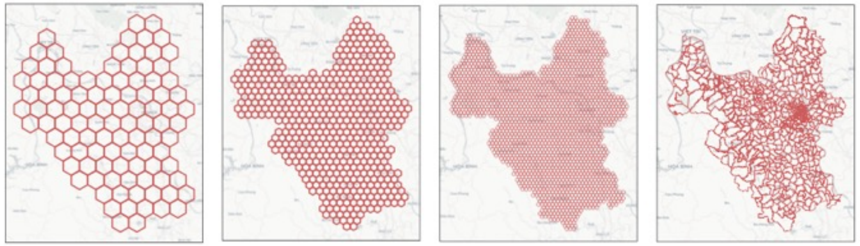
- BUT I want to focus on **multiple predictor variables** (the x's)



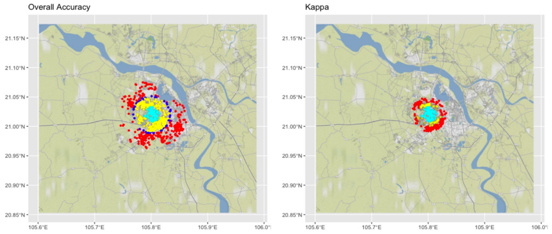
Summary

I have developed **novel methods** in 3 main areas

1. Determining optimal aggregation scale
   - handling the Modifiable Areal Unit Problem



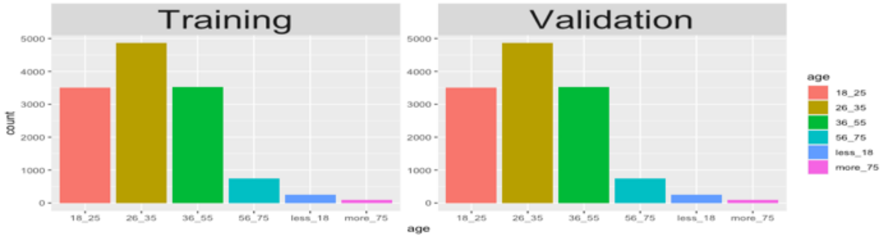
2. Multiscale GW Discriminant Analysis
   - Parameter specific, scale local classification



Percentage of data included in each local model

Variable	Overall	Kappa
Gender (red)	80%	40%
Purpose (blue)	50%	20%
Age (yellow)	40%	10%
Distance (cyan)	20%	10%

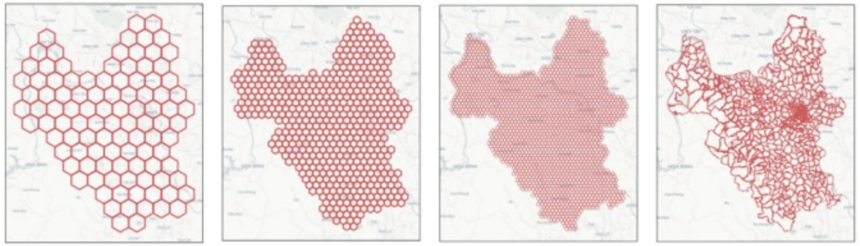
3. Methods for Under-sampling
   - resampling your sample

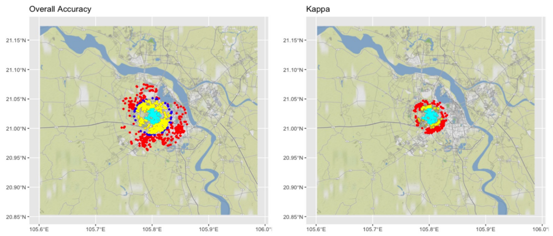


Summary

I have developed **novel methods** in 3 main areas

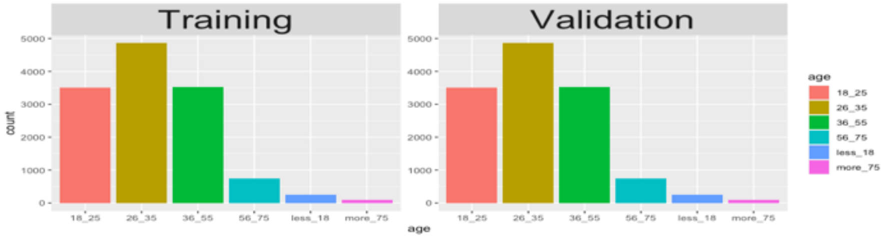
1. Determining optimal aggregation scale
   - handling the Modifiable Areal Unit Problem
   - Machine Learning, AI, Data Science → **ALL science**
2. Multiscale GW Discriminant Analysis
   - Parameter specific, scale local classification
   - Image vision, Remote Sensing → **ALL classification**
3. Methods for Under-sampling
   - resampling your sample
   - surveys more representative → **All Big Data**





Percentage of data included in each local model

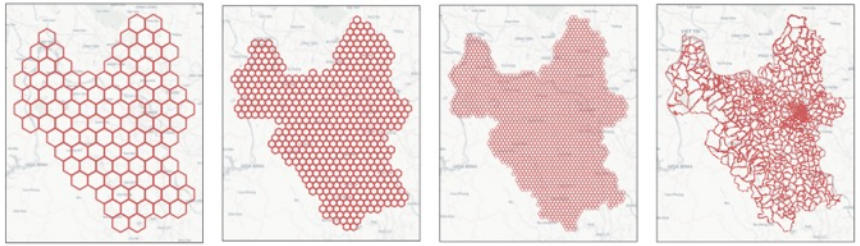
Variable	Overall	Kappa
Gender (red)	80%	40%
Purpose (blue)	50%	20%
Age (yellow)	40%	10%
Distance (cyan)	20%	10%



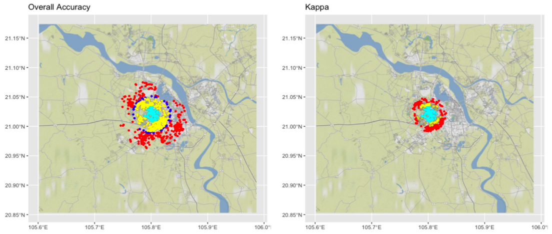
Summary

I have developed **novel methods** in 3 main areas

1. Determining optimal aggregation scale
   - handling the Modifiable Areal Unit Problem
   - Machine Learning, AI, Data Science → **ALL science**



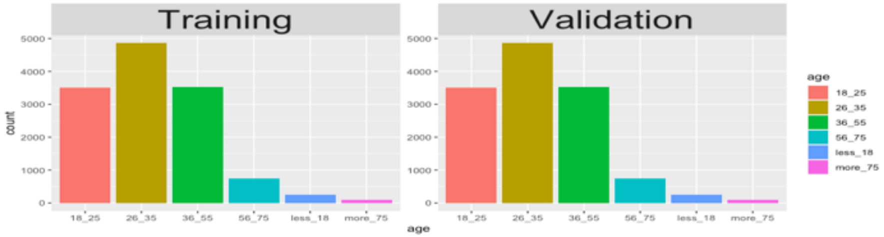
2. Multiscale GW Discriminant Analysis
   - Parameter specific, scale local classification
   - Image vision, Remote Sensing → **ALL classification**



Percentage of data included in each local model

Variable	Overall	Kappa
Gender (red)	80%	40%
Purpose (blue)	50%	20%
Age (yellow)	40%	10%
Distance (cyan)	20%	10%

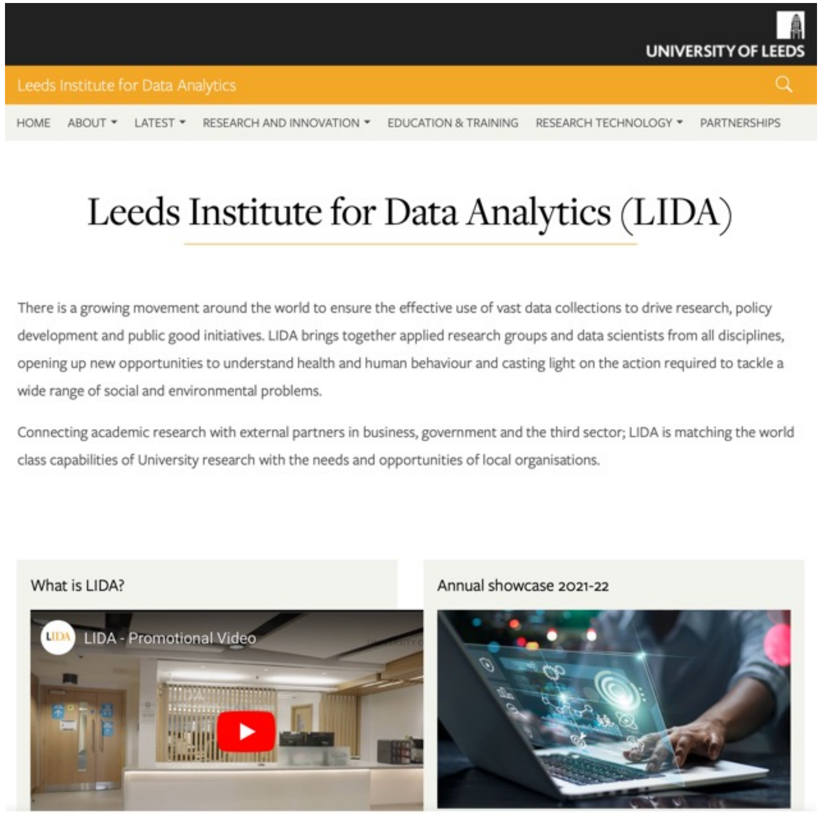
3. Methods for Under-sampling
   - resampling your sample
   - surveys more representative → **All Big Data**



• Relevance and Impact **beyond this project**

Future

- Working with VNU to establish a **Data Centre**
  - Based on LIDA
  - Link policy, industry and research
  - Provide **forum** for exchange of problems, expertise, ideas and data
  - Host a **new MSc** in **Spatial** Data Science
    - Leeds will contribute some materials / modules
- We have submitted a proposal for **extension funding** to deliver this
- We are looking for **collaborations** to take the next steps



Future

- We want to start the Data Centre through **this project**
  - Project survey data
  - Also build on previous work on SQTO (Dr Phe)
    - Quantifies **tangible** and **intangible** housing factors
    - Can detect **emergent** house price **bubbles**
- Centre to provide **hub** for data, organisations, & people!
  - And **methods**!
- Generate evidence to support **spatial planning**
  - Quantify **urban dynamics**
  - Underpin the concept of a **Smart City**