GEOG0125

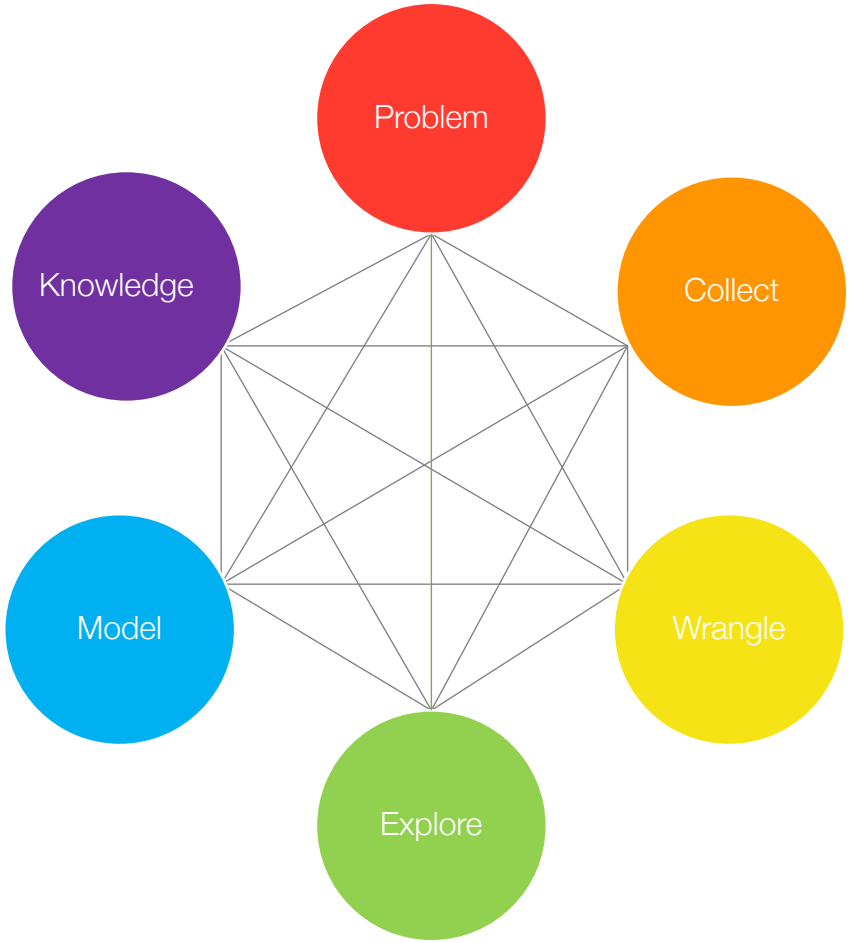
ADVANCED TOPICS IN SOCIAL AND GEOGRAPHIC DATA SCIENCE

RESEARCH METHODS, STUDY DESIGN & REVISION

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

Contents

- What is the meaning of Research Design
  - Types of Research Design
    - Quantitative
    - Qualitative
  - Types of Quantitative Study Methods
    - Descriptive studies
    - Observational studies
    - Randomized Control Trials
  - Study Designs
- Biases with focus on ecological designs
  - Ecological fallacies
  - Modifiable Areal Unit Problem (MAUP)
  - Residual confounding (with regression models)
  - Misclassification bias (with ML image classification)
- Best practices
- Summary of GEOG0125



What is a Research Design?

Definition:

Research Design typically refers to the investigator's strategy or plan for tackling a research question through collection of data, analysis and interpretation of such data, and finally a thorough discussion of that said data. In other words, or to simply put: "**… it's someone's blueprint for answering a research question**"

Types of Research Design:

- **Qualitative:** This area allows the investigator to gain meaningful insight (or empirical evidence) of certain phenomena through the study of non-numerical pieces of information

- **Quantitative:** This area allows the researcher to derive meaningful insight (or empirical evidence) about certain phenomena through analysis of numerical information by use of a **study design**.
  - ❖ Descriptive studies
  - ❖ Observational studies
  - ❖ Experimental studies

For **quantitative research**, the purpose is to establish **AT LEAST** one of three things: 1.) Evidence of causality (or an association); 2.) Internal validity; and/or 3.) External validity.

Types of Study Design

Type	Study Design	Properties
Descriptive	• Ecological (or geographic)*** • Cross-sectional	• Identify certain characteristics • Initial steps for searching the risk factors • Generating further hypothesis
Observational	• Case-control • Cohort (or longitudinal)	• Observe the effect of a risk factor(s)
Experimental	• Randomised control trials (RCT)*	• Observe the effect of diagnostic test, or other intervention in a group

Pilot study is a special cases as it could be anyone of these study designs

Ecological study

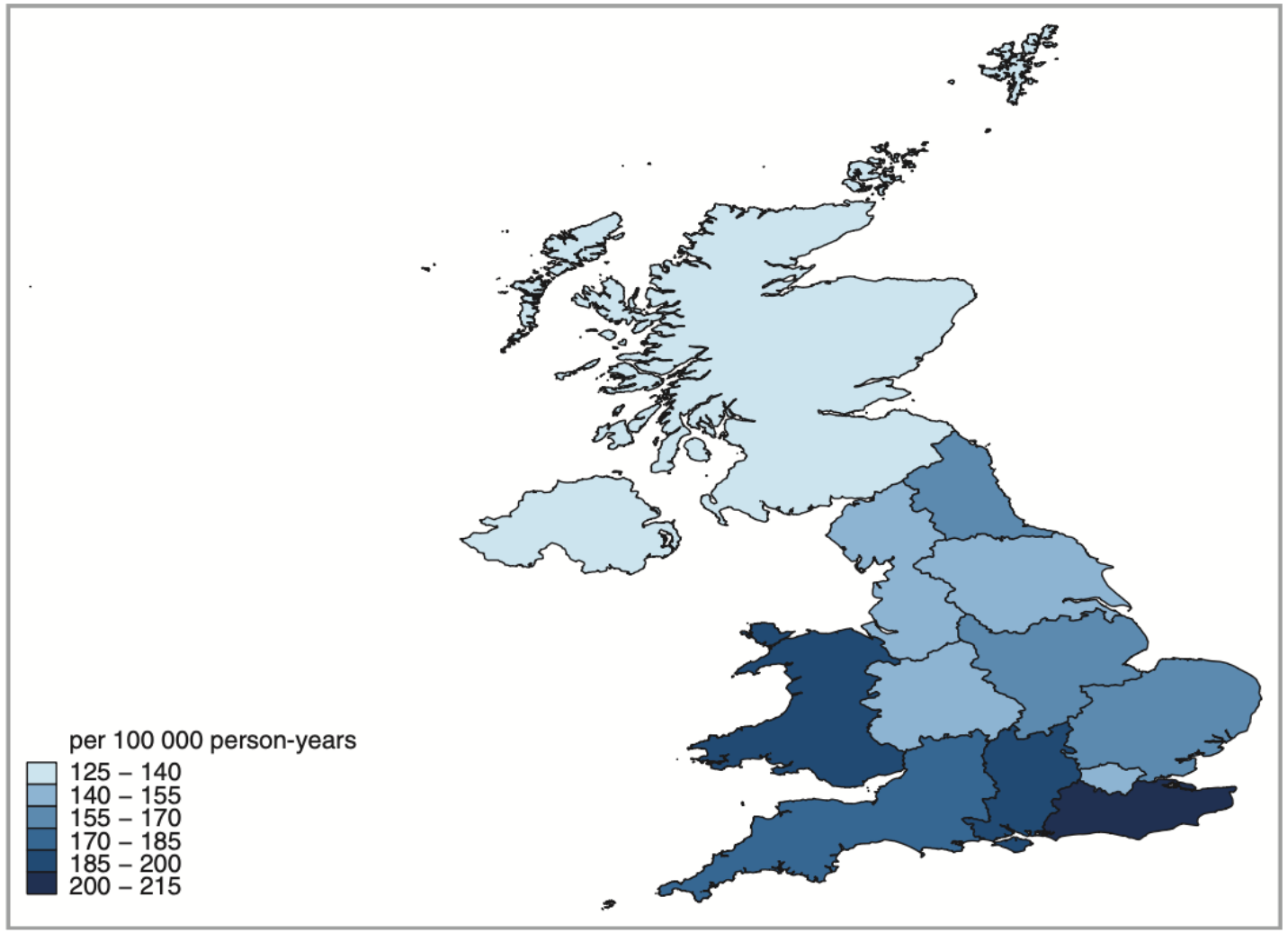
Ecological study design is a type **descriptive study**, where the population under investigation are analysed as groups, or aggregated units, rather than at an individual units. Aggregated units of data may include levels at a street-, postcode-, regional or even country-level

Key characteristics:

- An ecological study can be integrated within a **cross-sectional** or **longitudinal** framework
- There is a geographical component, which opens doors for **spatial analysis** and **mapping**
- Its descriptive analysis can be used to **generate more hypothesis** about the study population

Ecological study: Regional variation in skin cancer incidence in the UK

Musah et al., 2013



	Overall[b] IRR (95% CI)
Year	
2004	1
2005	1·04 (0·99–1·08)
2006	1·09 (1·04–1·13)
2007	1·15 (1·10–1·20)
2008	1·16 (1·12–1·20)
2009	1·15 (1·10–1·19)
2010	1·16 (1·12–1·21)
Annual increase, %	2·5 (1·9–3·0)
P for trend	< 0·001
Socioeconomic deprivation[c]	
5th (most deprived)	1
4th	1·07 (1·02–1·12)
3rd	1·21 (1·16–1·26)
2nd	1·37 (1·31–1·43)
1st (least deprived)	1·50 (1·44–1·56)
Unknown	1·17 (1·09–1·25)
P for trend	< 0·001
Regions	
London	1
Scotland	0·87 (0·83–0·91)
Northern Ireland	0·92 (0·85–0·98)
West Midlands	0·92 (0·88–0·97)
North West	0·96 (0·91–1·01)
Yorkshire & Humber	1·01 (0·95–1·08)
East Midlands	1·02 (0·96–1·09)
North East	1·05 (0·98–1·13)
East of England	1·04 (0·98–1·09)
Wales	1·23 (1·16–1·29)
South Central	1·21 (1·15–1·27)
South West	1·15 (1·09–1·21)
South East Coast	1·28 (1·22–1·34)

Figure 1: Shows incidence age-sex standardised rates of skin cancer across SHA in the United Kingdom. The table on left panel show year-on-year, regional, and areal-levels of deprivation in association with skin cancer

Cross-sectional study

Cross-sectional survey is a type of descriptive study that analyses data from a population at a specific point of time. Alternatively called ‘transverse’ or ‘prevalence’ study.

Key characteristics:

- Use of routinely collected data to perform large study
- It is the cheapest option among other study designs
- Questions can be asked to collect information retrospectively at a specific time point
- Can be an analysis of individual units or aggregated units

Development Frontiers in Crime, Livelihoods and Urban Poverty in Nigeria (FCLP) [1]

See source(s):
1. Musah et al, 2020
2. Umar et al, 2020
3. Umar et al, 2017

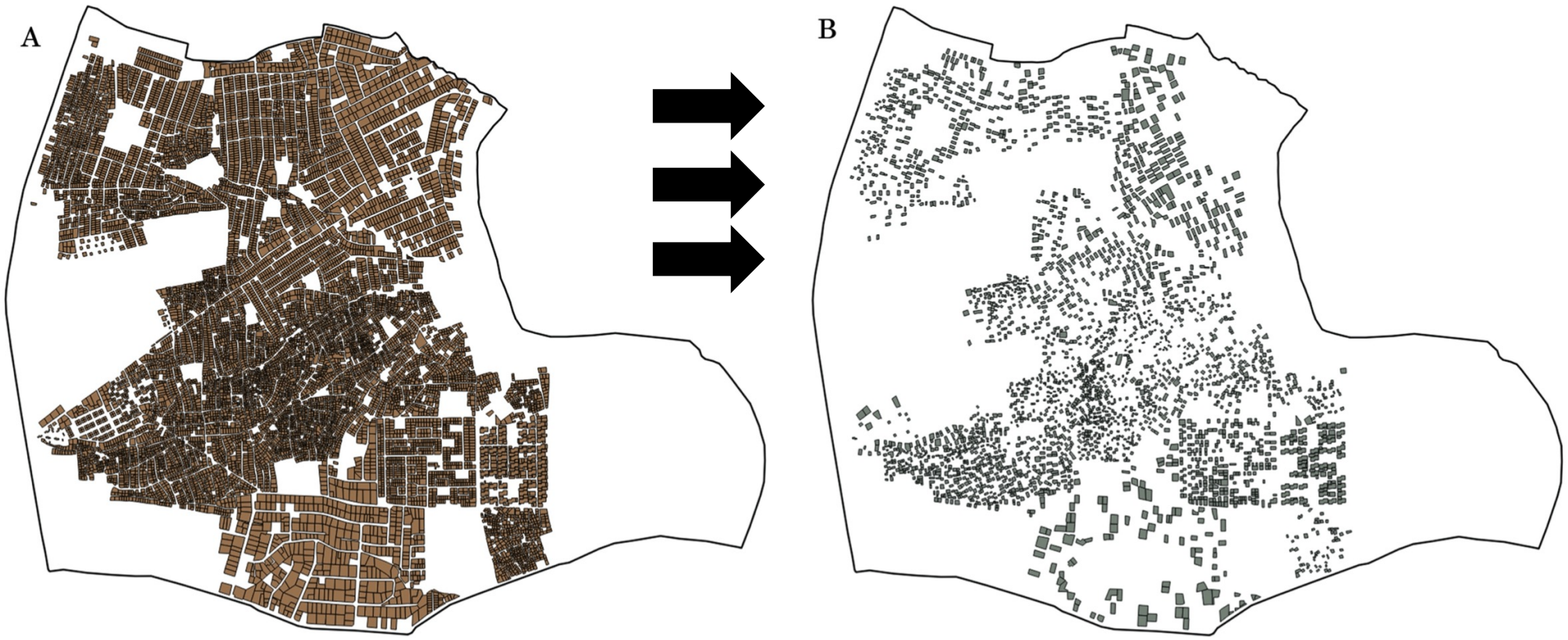
Themes:
1. Criminology
2. Social Science
3. Qualitative Research
4. Quantitative Research
5. Global South

Conventional analyses of crime, based on European research models, are often poorly suited to assessing the specific dimensions of criminality in Africa. Development Frontiers in Crime, Livelihoods and Urban Poverty in Nigeria (FCLP) aims to provide an alternative framework for understanding the specific drivers of criminality in a West African urban context.

Employing a mixed-methods approach combining statistical modelling, geo-visualisation and ethnography, the project situates insecurity and crime against a broader backdrop of rapid urban growth, seasonal migration, youth unemployment and informality. The study provides researchers both in Nigeria and internationally with a richer and more nuanced evidence base on the particular dynamics of crime in African cities.

Development Frontiers in Crime, Livelihoods and Urban Poverty in Nigeria (FCLP) [2]

See source(s):
1. Musah et al, 2020
2. Umar et al, 2020
3. Umar et al, 2017



Themes:
1. Criminology
2. Social Science
3. Qualitative Research
4. Quantitative Research
5. Global South

Data collected:

- Block Inventory Survey for the collection of environmental data
- Household victimisation survey on indicators for crime
- Perception of risk and neighbourhood safety
- Demographic survey
- **Survey CTO Collect (open source – need a server set-up)**

Primary data and sampling strategy: 13,000 households, and the target sample was 2,300 for the victimisation survey (in Nigeria); we therefore used Systematic sampling to select at random 2,300 households [the applied criteria: k = 13,000/2,300 = 5.6 ~ 6th property (starting from the left-side of road)]. Survey CTO application was used in the data collection of victimisation data.

Houses mapped in B, we had to interview all those residents (i.e., 2300 households) to collect a range of information.

Physical characteristics variables

BLOCK ENVIRONMENTAL INVENTORY FORM Block Ref:_____

s/n	Ref No	Land use	Plot	Other uses	Building Type	Building Material	Access Control & Target Hardening	Road Type:	Road Condition:	Traffic Pedestrian	Traffic Vehicular	Surveillance Morning	Surveillance Afternoon	Surveillance Evening	Surveillance CCTV Camera:	Surveillance Vigilante Post:	Surveillance Street lights:	Surveillance Street gates:	Signs of dis-order Graffiti:	Signs of dis-order Garbage:	Signs of dis-order Encrochment:	Signs of dis-order Drainage:
1		Occupied																				
2		vacant																				
3		Abandoned																				
4		Construction																				
		Shops																				
		Kiosk																				
		In-trading																				
		Out-trading																				
		Flat																				
		Storey																				
		Attached																				
		Detached																				
		Cement																				
		Mud																				
		Temporary																				
		High walls																				
		Burglary-proof																				
		Gate																				
		Garage																				
		Outdoor sitting																				
		Security Lights																				
		Guards																				
		Open Access																				
		Danger Warning																				
		CCTV Camera																				
		Dogs																				

Section A: Questions related to household Details

Sociodemographic variables

1. Are you the household head? Yes [ ] No [ ]
   If No, please indicate your relationship to the household head_____
2. **a)** Sex: Male [ ] Female [ ]
   b) Age: [ ]
   c) Ethnicity:_____
3. Occupation: Civil Service [ ] Private Organisation [ ] Craftsman [ ] Trader [ ] Farmer [ ] Student [ ] Retiree [ ] Unable to work[ ] Unemployed [ ] Others, please specify_____
4. Employment Level: Executive [ ] Managerial [ ] Expert [ ] Intermediate [ ] Trainee [ ] Large business proprietor [ ] Small business proprietor [ ] Others, please specify_____

+

Section B: Questions related to your relationship with neighbours and your perception of safety within this street

Perception and safety variables

Note: - Properties in a street are those on both street block faces between two road intersections
- Neighbours are those people who live in the same street with you

1. How safe do you feel living on this street? Extremely safe [ ] Very safe [ ] Moderately safe [ ] Slightly safe [ ] Not safe at all [ ]
2. How worried are you about being a target of property crime while you are away from home?
   Not worried at all [ ] Slightly worried [ ] Moderately worried [ ]
   Very worried [ ] Extremely worried [ ]
3. How many of your neighbours do you know? All of them [ ] Most of them [ ]
   Half of them [ ] A few of them [ ] None of them [ ]

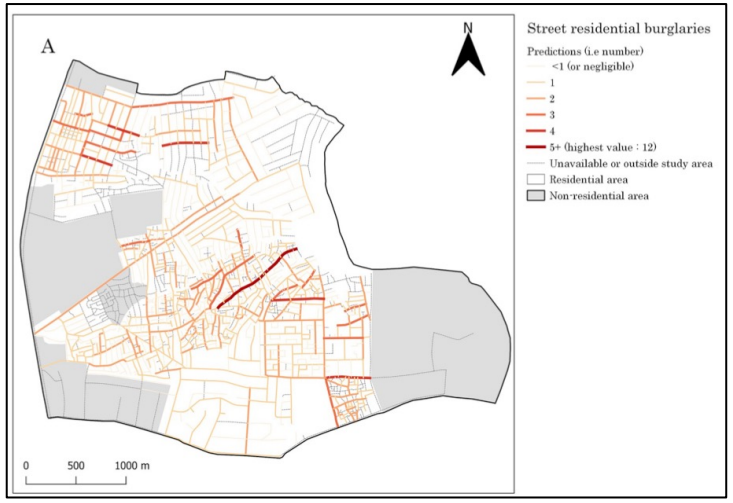
+

Section C: Questions related to incidents that had happened within your property

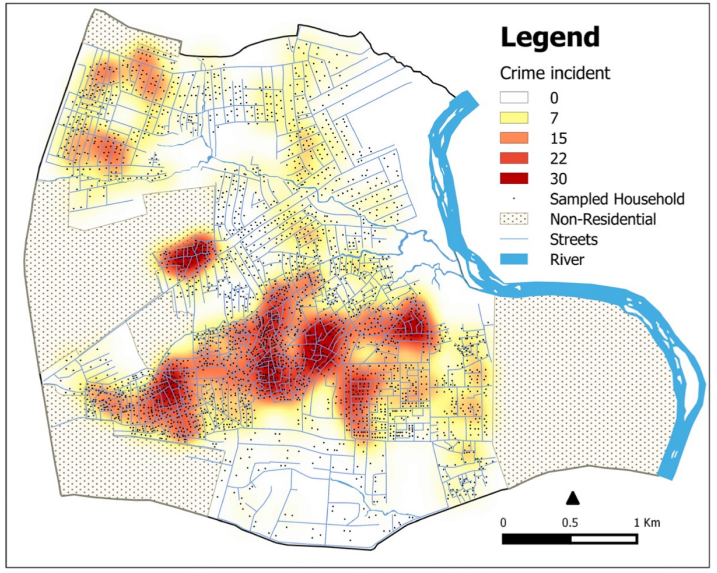
Victimisation (dependent) variables

In the **LAST 1 YEAR**, have any of the following incidents **HAPPENED** within your Property?

1. Burglary (Breaking-in) - Yes [ ] No [ ]
   If yes, how many times? [ ]
2. Stealing of valuables (Not breaking-in) - Yes [ ] No [ ]
   If yes, how many times? [ ]
3. Deliberate damaging of your property Yes [ ] No [ ]
   If yes, how many times? [ ]
4. Theft **from** Automobile Yes [ ] No [ ]
   If yes, how many times? [ ]



Research 1: From the 2,300 household sample, we used the "crime pattern theory" to assess the risk of burglaries & victimisation at a street-level (see source: [LINK])



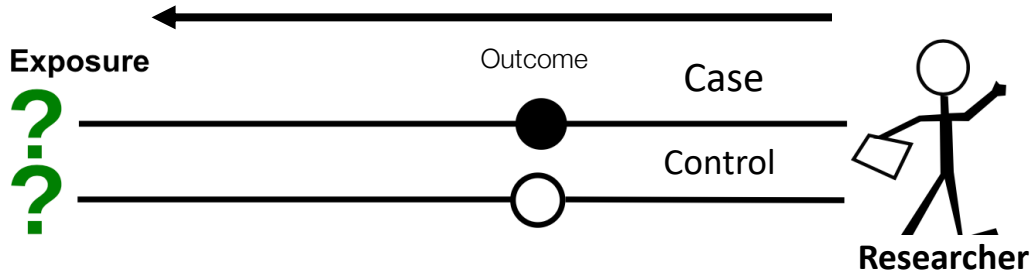
Research 2: From the 2,300 households were sampled, we used the "laws of crime concentration" to assess the concentration of reported victimisation in this city (see source: [LINK])

Case-control study

Case-control study is an **observational study** which has two groups under analysis – a set of cases (with outcome), and a set of controls (with outcome that's absent). The two groups are compared 'retrospectively' on the basis on a common exposure/independent variable.

Key characteristics:

- It always deals with individual-level or units of data
- The outcome is measured exclusively as a binary variable ("yes", "no"), ("win", "lose"), ("alive", "dead) etc. Thus, this limits the selection of models to a **simple logistic** or **conditional logistic regression**
- Cases are most often matched to a control based on similar characteristics (same age group, gender and postcode). The matching can be done 1:1, 1:2 or 1:5 etc.
- It is also referred to as '**retrospective**' study, as odds of outcome are predicted based on exposures determined retrospectively

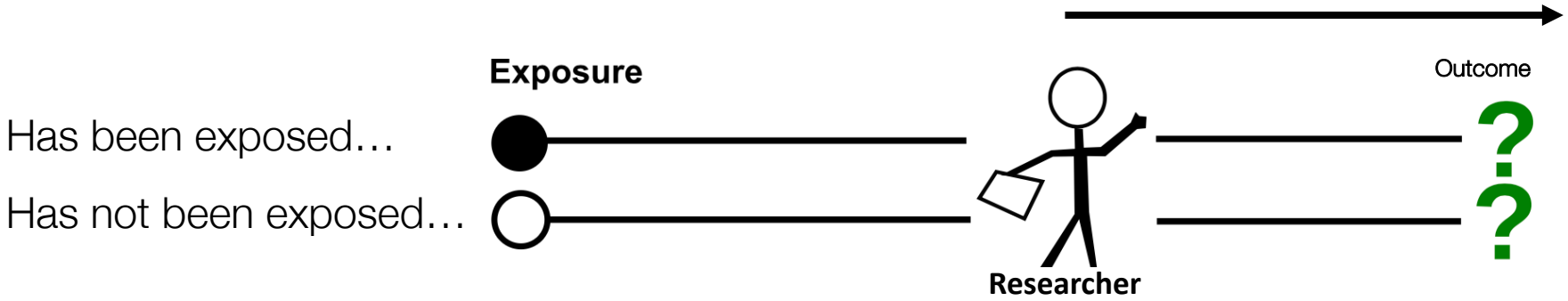


Cohort study

Cohort study is a type of **observational study** which focuses on the following the study participants **prospectively** at different points in time – answer the question about 'change' (e.g. pre and post, time series).

Key characteristics:

- This type of study design allows to tracking the trajectory or changes in one's exposure (or risk factor) status over his/her life course
- You can literally observe the onset of an outcome
- There is that time-dependencies that is taken into consideration when tracking a participant over the course of the study
- Due to the time-dependent nature of study design – again, investigators must think carefully about the selection of statistical methods (i.e., survival analysis or time series are best for this).
- This is referred to as '**Prospective**' or '**Longitudinal**' study



Pilot Study

A Pilot study is the most unique type of study design which is versatile, its basically a small scale preliminary study. It is useful to test your questions on a small pilot responders, and correct the issues before launching it with the full sample

Key characteristics:

- Highly flexible i.e., a pilot can be within a quantitative or qualitative framework
- Highly flexible i.e., pilot can be performed with a cross-sectional or longitudinal framework
- It is the most cost-effective and never time-consuming to evaluate feasibility
- It is the best way to generate a hypothesis about a small population before going big

Cross comparison of quantitative-based study designs

	Ecological study	Cross-sectional	Case-control	Longitudinal	Pilot
Resolution of study design	This solely deals with aggregated units of data for analysis	Individual-level at a particular point in time	Individual-level where the outcome has already been observed among cases. There must be a control population	Individual-level at several points in time	Could be either aggregated or individual-level (1 mark)
Unit of time analysis	Has the flexibility of being a cross-sectional (i.e. single point in time) or longitudinal (i.e. several different time points).	At a single point in time	Past exposure	Several points in time, or before/after	Has the flexibility of being either ecological, cross-sectional or longitudinal study but dealing with smaller sample size as pilot before to doing a much bigger study.
Its cost effectiveness	Cheapest as it relies on routinely collected data most of the time	Less expensive as it conducted as a single time point and requires less resources	Quite expensive to interview participants to provide past experiences	Most expensive as it requires two or more follow-up of subjects enrolled in the study so more resource and time are required.	It's a cheap way for assessing whether to do a bigger (e.g. population-based) study
Common biases (or limitation)	Ecological fallacy	Results are only representative at time of study	Recall Bias	People dropping out of study can introduce lost-to-follow-up bias	Safest options as it's a pilot
Strength	Weakest	Strong	Strong	Strongest	Safest option
Retrospective or prospective?	Both	Present or retrospective	Always retrospective	Always prospective (even it's using historical data)	Both

Biases (within an Ecological Study context)

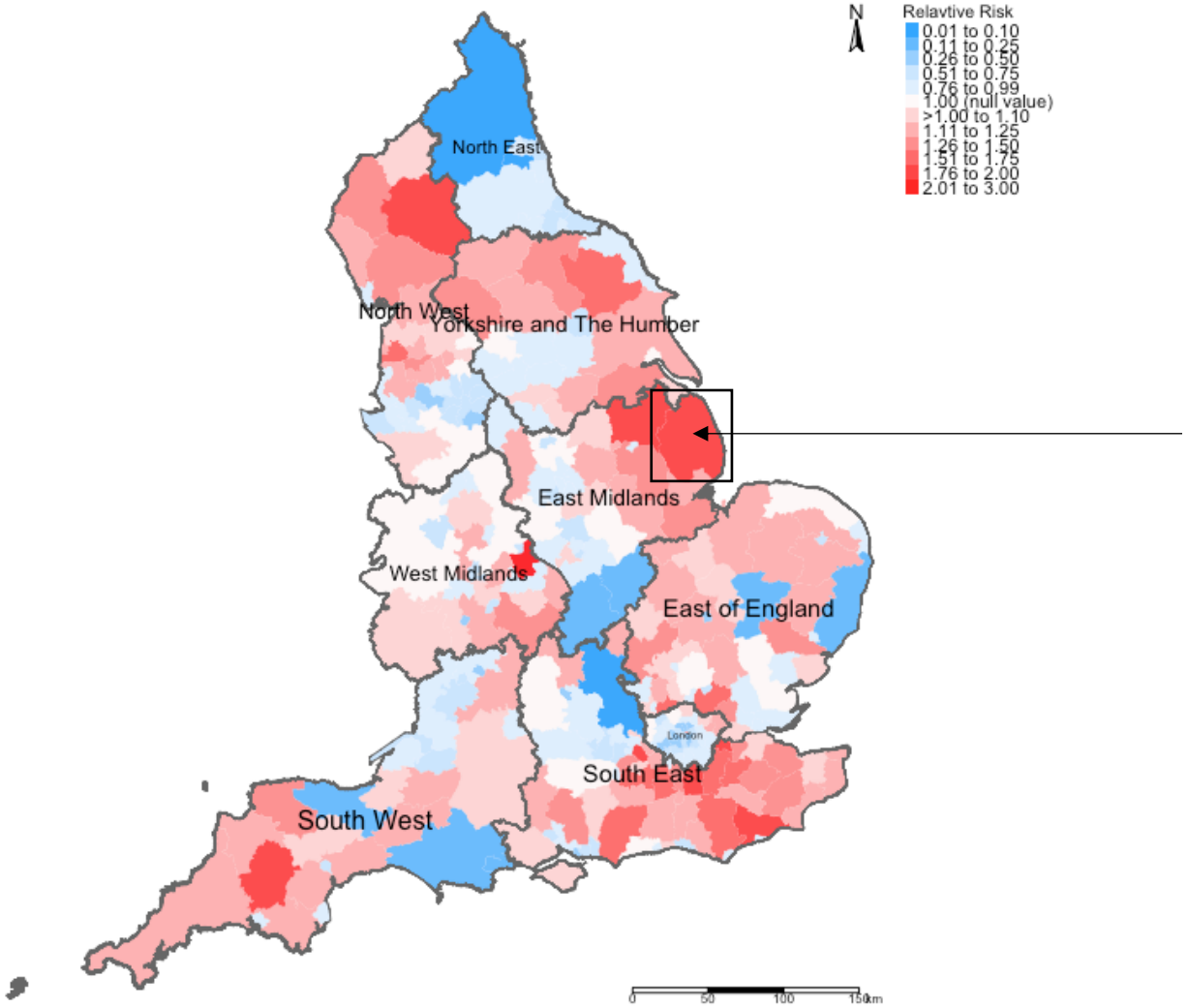
Ecological Fallacy

Ecological study design always carries this bias baggage known as **Ecological Fallacy**, a bias which must always be acknowledged. This is logical error that occurs when statistical inference made on groups are attributed to an individual.

Important notes to keep in mind:

- Ecological studies assume what is true for a population is also true for the individual members of that population which is **false**.
- Think of inferences made here are a **gross generalization** or **stereotyping**
- To remove this type of bias completely from one's research – it is best to perform research by using **individual units of data** instead of aggregated data.
- Because of this bias – **ecological studies are weakest!**

Example: Risks of Road-related casualties in England 2015-2020



Relative Risks (RR)

Essentially, this result is saying that everywhere in this local authority has risks that are 2/3-fold for road accidents.

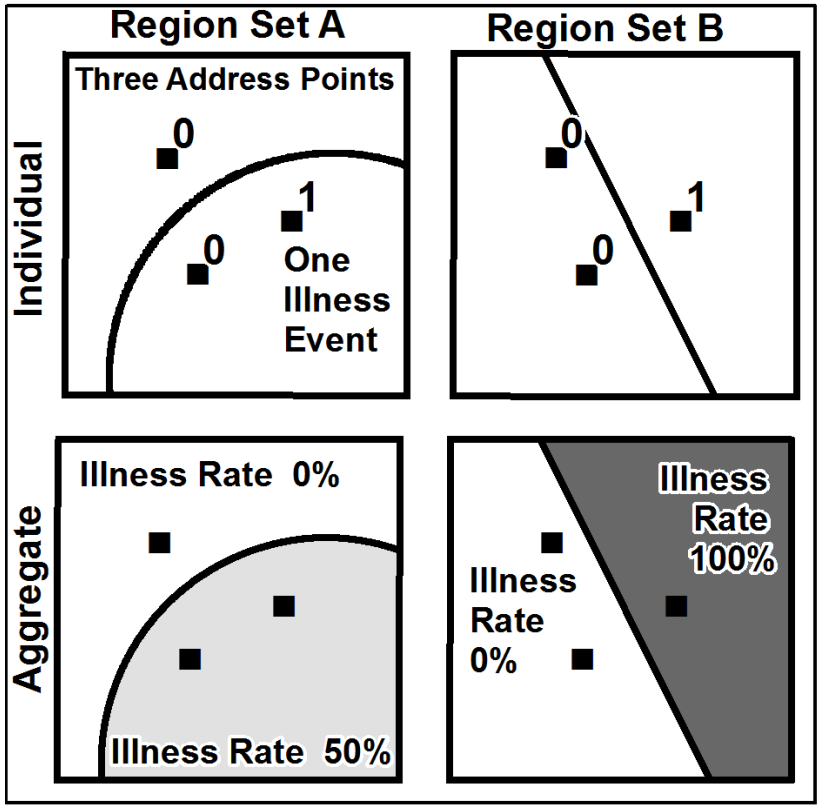
Do you see the problem and bias that must be acknowledged?

Modifiable areal unit bias (MAUB)

MAUB is a source of statistical bias in an ecological study with data on geographic units. This is concerned with how points or estimated are aggregated within areas whose boundaries change (Stan Openshaw [1984])

Important notes to keep in mind:

- This is an aggregation problem (gridded, point or small areas) that arises with changing geographic boundaries.
- This bias can severely distort you results.
- Today, provide contemporary estimates for boundaries A. Boundaries A are modified and changes to Boundaries B. You must perform a new analysis to provides updated results as the changes in boundaries change everything!
- You must acknowledge this bias whenever you have tempered with a shapefile (i.e., dissolved, clipped etc.,)



Residual Confounding (any regression analysis)

A confounding variable is a something that distorts the relationship between your primary independent variable of interest and outcome.

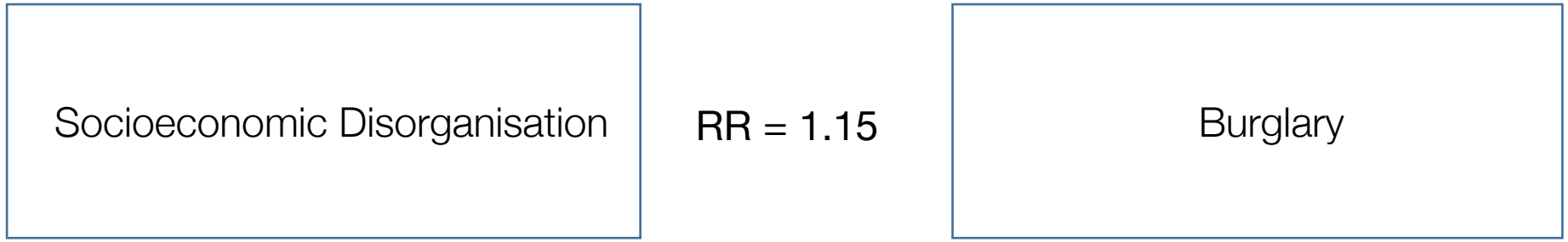
Important notes to keep in mind:

- Confounders are central to the interpretation of any epidemiological study
- Confounding factors, if not controlled for, will cause a distortion in the estimate of the impact of the independent variable(s) being study – this is important to know!
- There are 'generally' three conditions for determining whether a variable is a confounder or not: 1.) **A confounder is associated the primary independent variable of interest**; 2.) but **it is also a risk factor of the outcome** as well; and 3.) this confounder does not exist between the causal pathway for the **independent variable of interest** to **outcome**.

Notes: In your research, you will always have an independent which is what you are primarily interested, and you want to know whether its associated to an outcome.

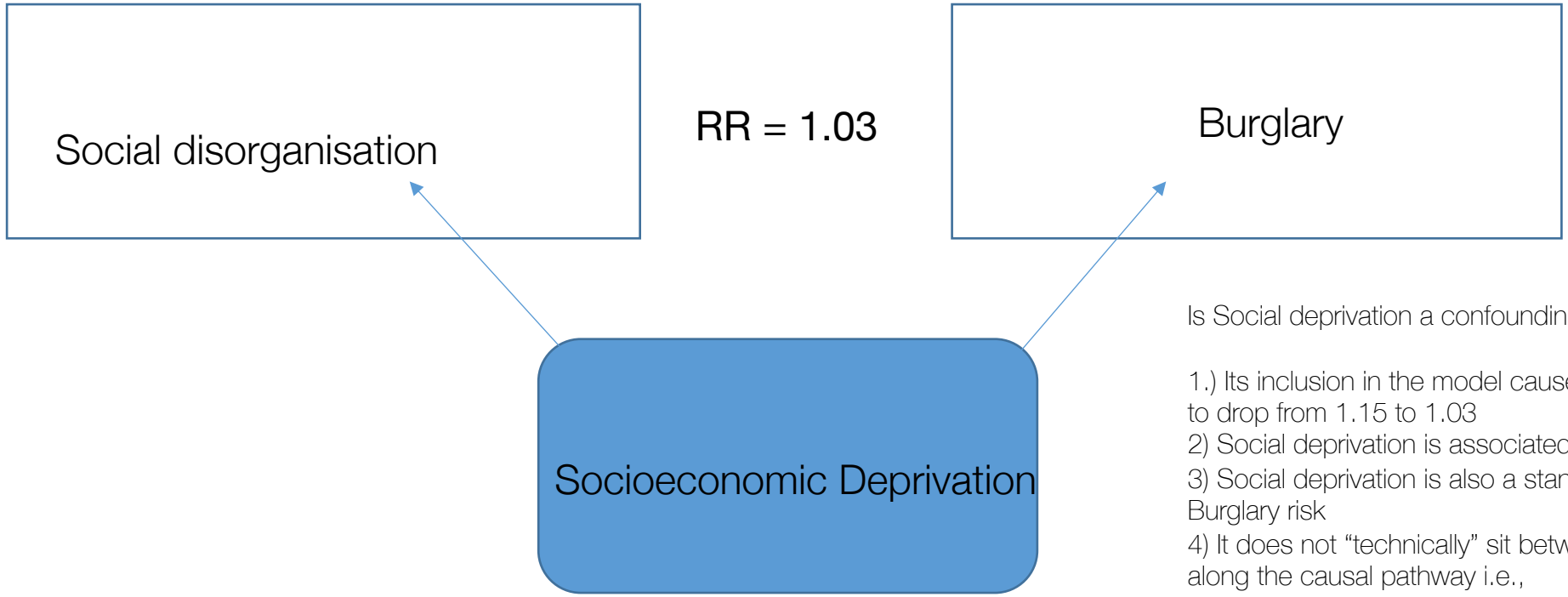
However, the can be other variables which must be accounted for in the analysis which can confound your association. Your lack of including other important variables in the analysis will lead to a bias known as residual confounding which you must acknowledge!

Suppose you want to perform an ecological study crimes and wanted to understand how social disorganisation as a primary independent variable was associated with burglary risk. You run a univariable spatial risk model with one single independent variable.



Notes: There are ‘generally’ three conditions for determining whether a variable is a confounder or not: 1.) A confounder is associated the primary independent variable of interest; 2.) but it is also a risk factor of the outcome as well; and 3.) this confounder does not exist between the causal pathway for the independent variable of interest to outcome.

Suppose you want to perform an ecological study crimes and wanted to understand how social disorganisation as a primary independent variable was associated with burglary risk. You run a univariable spatial risk model with one single independent variable.



Is Social deprivation a confounding variable?

1.) Its inclusion in the model caused a distortion in the RR to drop from 1.15 to 1.03
2) Social deprivation is associated with disorganisation
3) Social deprivation is also a stand alone risk factor for Burglary risk
4) It does not "technically" sit between the two variables or along the causal pathway i.e.,

Deprivation → Disorganisation → Burglary

Yes, Socioeconomic Deprivation is a confounding variable

Notes: There are 'generally' three conditions for determining whether a variable is a confounder or not: 1.) A confounder is associated the primary independent variable of interest; 2.) but it is also a risk factor of the outcome as well; and 3.) this confounder does not exist between the causal pathway for the independent variable of interest to outcome.

Bias that can occur in machine learning analysis

Bias	Information
Sample (or selection) bias	Sample bias occurs when a dataset does not reflect the realities of the environment in which a model will run. An example is the development of facial recognition systems that's solely trained on images on one particular race (e.g., only black men). These models have considerably lower levels of accuracy with women and people of different ethnicities. Another name for this bias is selection bias
Measurement bias	This type of bias occurs when the data collected for training is different from that collected in the real world (i.e., ground truth), or when faulty measurements result in data distortion.
Racial bias	Racial bias occurs when data skews in favour of particular demographics (e.g., this can be seen in facial recognition and automatic speech recognition technology which fails to recognize people of colour as accurately as it does Caucasians)
Association bias	This bias occurs when the data for a machine learning model reinforces and/or multiplies a cultural bias. Your dataset may have a collection of jobs in which all men are doctors and all women are nurses. This does not mean that women cannot be doctors, and men cannot be nurses. However, as far as your machine learning model is concerned, female doctors and male nurses do not exist. Association bias is best known for creating gender bias.

Best practices for writing a methodology

Writing the methodology section [1]

- **Describing the where, what and when:**
  - ❖ **What:** Here, you are mentioning the name of the data source and its use for a bigger study
  - ❖ **When:** The year(s) (or date(s)) at which it was collated
  - ❖ **Where:** The location of focus for which the data was collated from and the description of the study area (a map of the study area helps a lot!).

- **Specific details about how the data was collated:**
  - ❖ Whether its through questionnaire survey, interview etc.,
  - ❖ Research framework i.e., ecological, pilot, cross-sectional, or longitudinal etc.,
  - ❖ Sampling strategy
  - ❖ Who and what the target population was i.e., target sample size and group of focus (e.g., Adults only i.e., 18 years and above etc.,)

Writing the methodology section [2]

[continue]

- **Describe the variables that's going to be used for the analysis:**
  - **Codebook –** provide a list of all the intended variables that is going to be analysed. You must mention specific details about variable – information such as the name, variable type (i.e., numeric or categorical) etc.,
  - **Initial sample characteristic table –** this is a breakdown on the number of observation documented for each variable (- i.e., what's present and missing).
    - Example 1, suppose the total is 100 and 89 respondents provided the ages. In this table, the mean age is calculated from that 89. You must report that the mean is based on 89 point, and 11 points are missing data.
    - Example 2, suppose the total is 100, where 51 were women and 42 were men. You must report the numbers and percentage for each category, and report the numbers (& proportion) of missing data in the gender variable. This means including a third category: men (42), women (51), and unknown/missing (7)

ID	Name	Age	Gender	Pathway	Gamer
0001	Brittany	34	Female	NA	No
0002	Idris	NA	Male	NA	NA
0003	Spike	32	Male	Political	Yes
0004	Lara	29	Female	Geography	Yes
0005	Fiona	28	Female	Health	No
0006	Xiaoyu	29	Female	Geography	NA

Raw dataset

Characteristic	Breakdown on what's present in the dataset
Age (in years)	Mean 30 (n = 5, 1 missing)
Gender	
Women	4 (66%)
Men	2 (33%)
Missing/Unknown	0 (0%)
Pathway	
Political	1 (16.5%)
Health	1 (16.5%)
Geography	2 (33%)
Missing/Unknown	2 (33%)
Gamer Status	
Yes	2 (33%)
No	2 (33%)
Missing/Unknown	2 (33%)

Show what's available

To show missing numbers by category

table(dataset$pathway, useNA = "always")

To exclude the missing entries

mean(dataset$age, na.rm = TRUE)
summary(dataset$age, na.rm = TRUE)

Writing the methodology section [3]

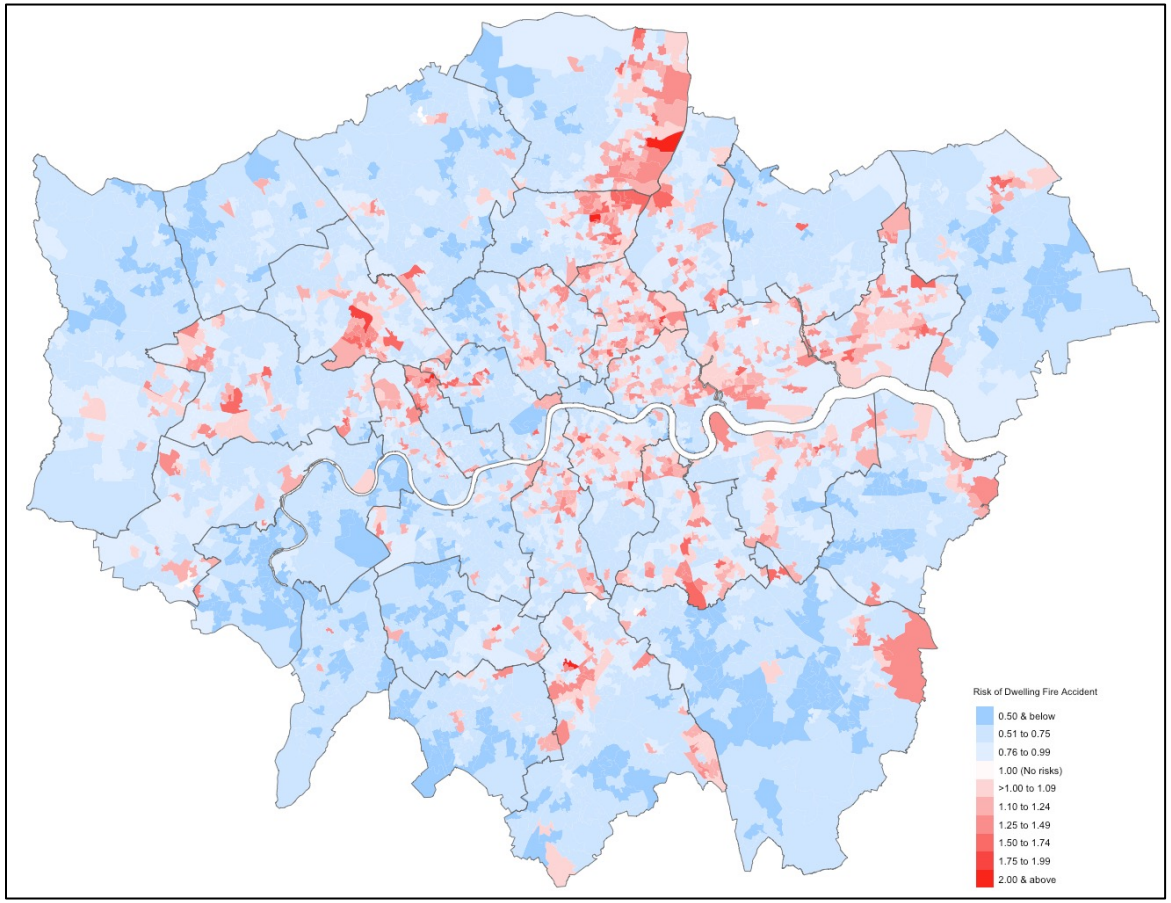
[continue]

- **Describe the statistical methodology to be used for the analysis:**
  - ❖ **Description of approaches and what parameters you intend to estimate**
  - ❖ **Structure of any statistical analysis:** 1.) Descriptive and/explorative analysis; and 2.) Evidence based analysis (i.e., Machine learning or statistical models)
  - ❖ **Model formulation and description of the parameters and they'll be interpreted and how is significance determined.**
    - **Ask yourself –** is this going to be from a Frequentist or Bayesian approach?

READ A LOT OF PAPERS TO SEE HOW THE METHODS SECTION ARE ARTICULATED!

Summary of GEOG0125

Bayesian Methodology



- We have introduce the basics of Bayesian inference
- We learnt how to implement the following for evidence-based research:
  - GLMs
  - GAMs
  - Hierarchical models
  - Spatial iCAR models
- We have learned how to use Stan with RStudio for Bayesian inference

Machine Learning and GeoAI

Image Classification

Object Detection

- We have introduce the subfield of Machine Learning for complex image classification in geography.
- We have learnt how to implement the following approaches:
  - ❖ Deep Learning Methods
  - ❖ Convolutional Neural Networks
  - ❖ GeoAI
- We have learned how to use PyTorch & Python for implementing these complex machine learning algorithms

GEOG0125: Course Evaluation & Student Feedback (Weeks 1-3 and 7-9)

https://forms.gle/3AHiH4Q2L7NmXZCe9

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 **GEOG0125: Advanced Topics in Social & Geographic Data Science**. 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 minutes, and your responses are completely anonymous.

Thank you,

Anwar and Stephen.

GEOG0125: Course Evaluation & Student Feedback (Week 4 to 6)

https://forms.gle/rddFzY8UABaZhyDK9

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 **GEOG0125: Advanced Topics in Social & Geographic Data Science**. 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 minutes, and your responses are completely anonymous.

Thank you,

Anwar and Stephen.

Any questions?

Revision and Assessment

GEOG0125 Assessment

1. Be concise, accurate and purposeful in what you write. Justify all your decisions, either with your data analysis or literature including model choice, parameter choice.
2. A excellent introduction contains the following: rationale and justification for pursuing a research question and scientific reason for selecting a particular study area.
3. Tables and Figures should be clear (readable), purposeful, captioned, and linked to the data report.
4. The two parts of the assessment should be seen as two extended abstracts with its own introduction, method, results, conclusion.
5. For the Bayesian analysis subsection, select a dataset that is of areal units and apply a Bayesian Spatial risk model. Perform a risk assessment and generate mapped output for the relative risks, significance (95% credibility limits) and exceedance probabilities.
6. Provide an overall interpretation of the global and area-specific estimates from the Bayesian model.
7. For the Machine learning subsection, select a dataset and the type of problem you want to answer (e.g., image classification). Remember to explain the model you will test and the pipeline you will go through. Always report the “out of sample test” set results (e.g., in a table) as generalisation is key in ML.
8. Look at the coursework assessment criteria and please do not include actual “code” in the report.
9. Have fun!!!

Machine learning workflows and pitfalls

Contents

1. **Preprocess and Explore** to visualize, describe and understand data, standardise to stabilize training and make features comparable.
2. **Resample dataset** into a train, val and test set. The aim of machine learning model is to optimize out-of-sample accuracy.
3. **Train multiple models** with different methods and features. A common research problem is to test a different method or feature.
4. **Fit** models with training data, **tune** parameters with validation data and **evaluate** on test data. Check for overfitting.
5. **Sanity check** the outputs of the final model
6. **Report with a table** on the different models and predictive plots.
7. **What do the result mean**? What are its implications? What are the limitations?

Lones, M. A. (2021). How to avoid machine learning pitfalls: a guide for academic researchers. arXiv preprint arXiv:2108.02497. https://arxiv.org/pdf/2108.02497.pdf

“You’re gonna carry that weight…”