POLS0008
INTRODUCTION TO QUANTITATIVE RESEARCH METHODS

WEEK FOUR: SOURCING DATA

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

Last updated: 10:45am 04/02/2025

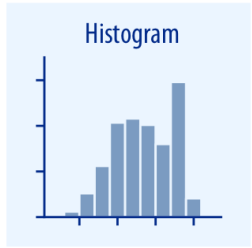
QUICK RECAP OF WEEK 3

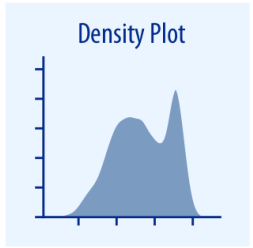
We covered some general techniques for data visualisation

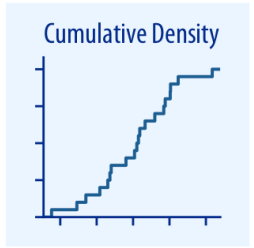
Visualisations representing densities & distributions of numerical data

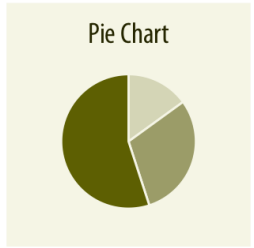
1. Plots for densities and distributions (numerical data) **2. Plots for proportions (qualitative or categorical data)**

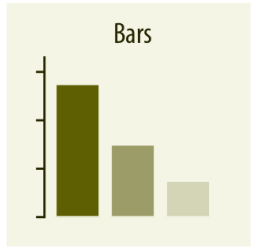
Single variable

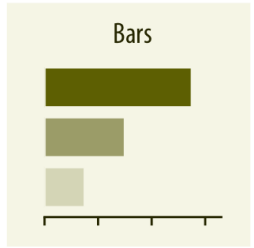


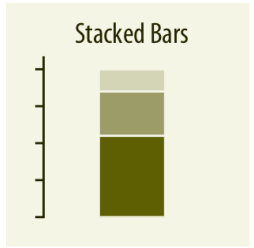




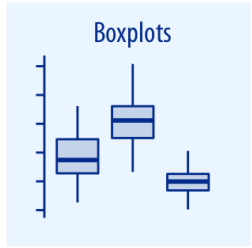


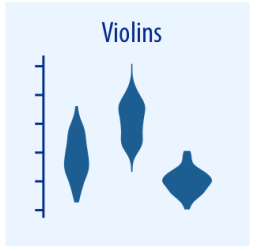


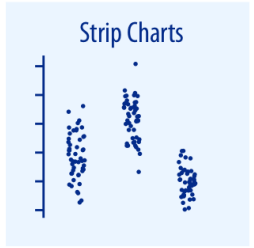


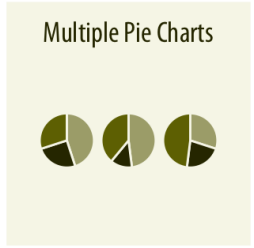


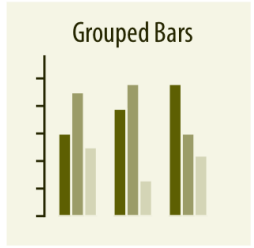
Multiple variables

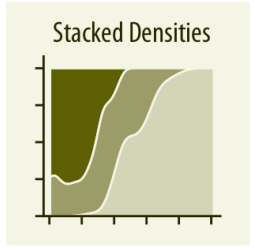


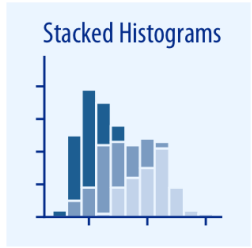


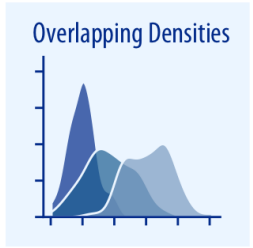


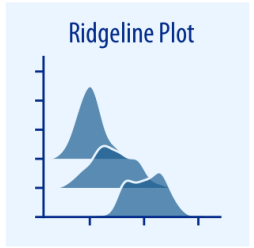












Source: Fundamentals of Data Visualization [LINK]

We covered the best practices for data visualisation

Everything on your graph should be labelled accordingly:

- **Title** – a clear short title letting the reader know what they are looking at should be present. Or alternatively, a figure legend for that image will do.

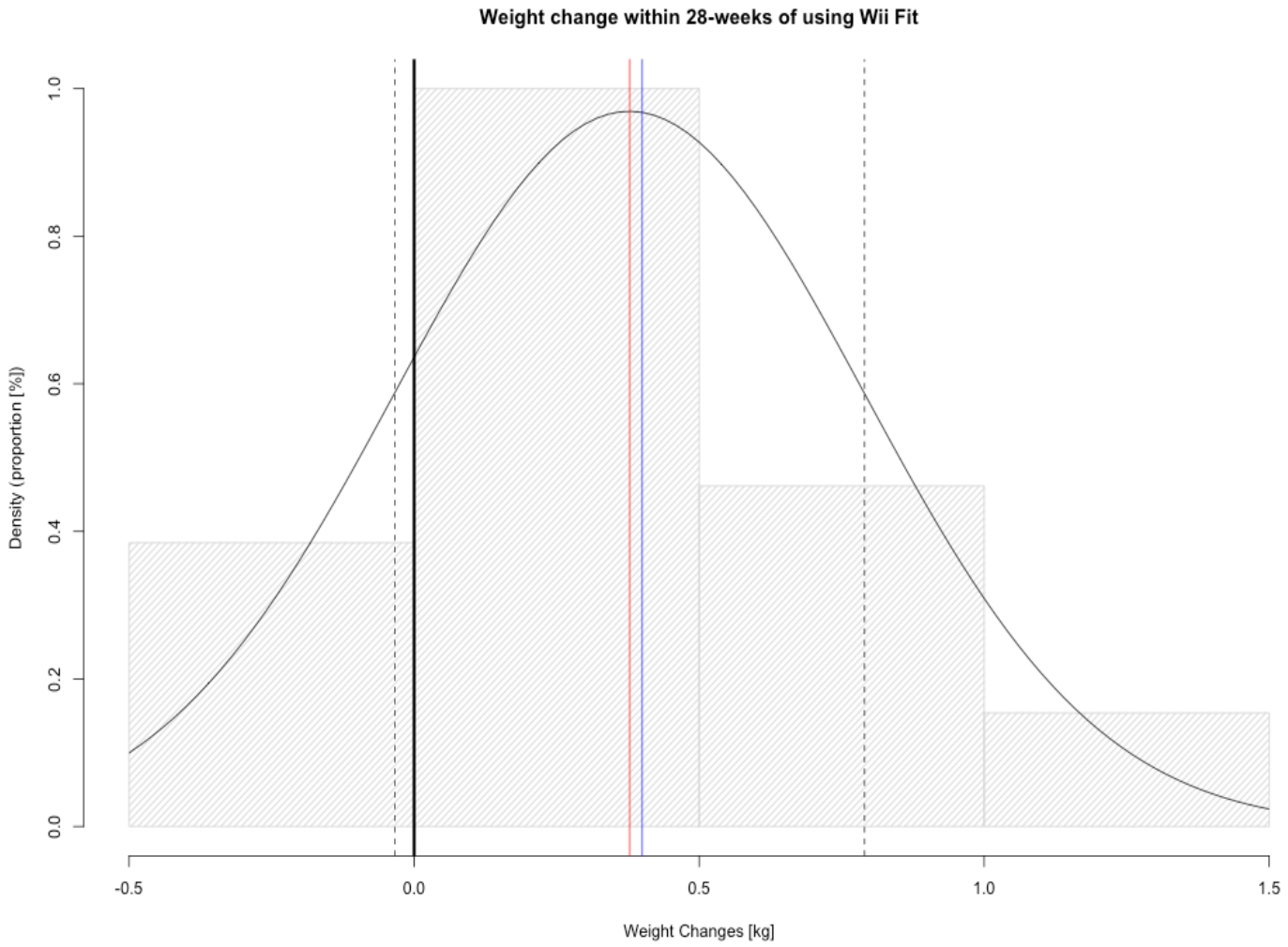
- **Axis labels/titles** – clear labels for the x and y axes must be present
  - ❖ Should include in the labelling the units of measurements [height (m), soil arsenic (mg/kg) etc.]
  - ❖ These labels should be short and descriptive

- **Legends** – for categories in categorical variables which keys/colour codes must be present and labelled accordingly
  - ❖ Male and Female, and not 0 and 1.

- **Captions** – If the graphics are **NOT** yours (i.e., its ripped from a source). Take the opportunity to apply a caption on the graph (on or beneath it) providing source attribution for the data.

- **Colour scheme** – Use of colour scheme matters
  - ❖ Sequential colours – for plotting quantitative variable that goes from low to high (vice versa)
  - ❖ Diverging – for contrasting the extremes (low, medium and high) of a quantitative variable
  - ❖ Qualitative – e.g., nominal categories. Use to distinguish between different categories in a categorical variable

We covered an example using Base-R code




```
# R Code:

# enter data
weightChanges <- c(-0.2,-0.4, 0, 0.1, 0.1, 0.3, 0.12, 0.4, 0.5, 0.8, 0.9,
0.5, 0.6, 0.7, 0.6, 1.3, -0.2, 0, 0.1, 0.1, 0.3, 0.4, 0.5, 0.5, 0.6, 1.2)

# next, extract mean and standard deviation from data
m<-mean(weightChanges)
std<-sd(weightChanges)

# plot histogram with normal curve
hist(weightChanges, density = 20, prob=TRUE, main="Weight change
within 28-weeks of using Wii Fit", xlab = "Weight Changes [kg]", ylab =
"Density (proportion [%])")

# adds the normal curve
curve(dnorm(x, mean=m, sd=std), add=TRUE)

# add red line for mean
abline(v = 0.378, col = "red")

# add blue line for median
abline(v = 0.400, col = "blue")

# add black dashed line for -sd
abline(v = -0.034, lty = "dashed", col = "black")

# add black dashed line for +sd
abline(v = 0.79, lty = "dashed", col = "black")

# add blue line for median
abline(v = 0, col = "black", lwd = 3)
```

Note: Area beneath the curve before weight change value at 0 kg (black solid line), tells me the predicted probability that the Wii Fit can reduce my weight.
Prob(Weight change < 0.0kg) = 0.1794 (17%)

```
# Calculate probability
pnorm(0, mean=m, sd=std)
```

We demonstrated further ways of visualising data in RStudio with ggplot( )

Creating impressive visualisation

- Base functions for creating graphics in R: **plot( )**

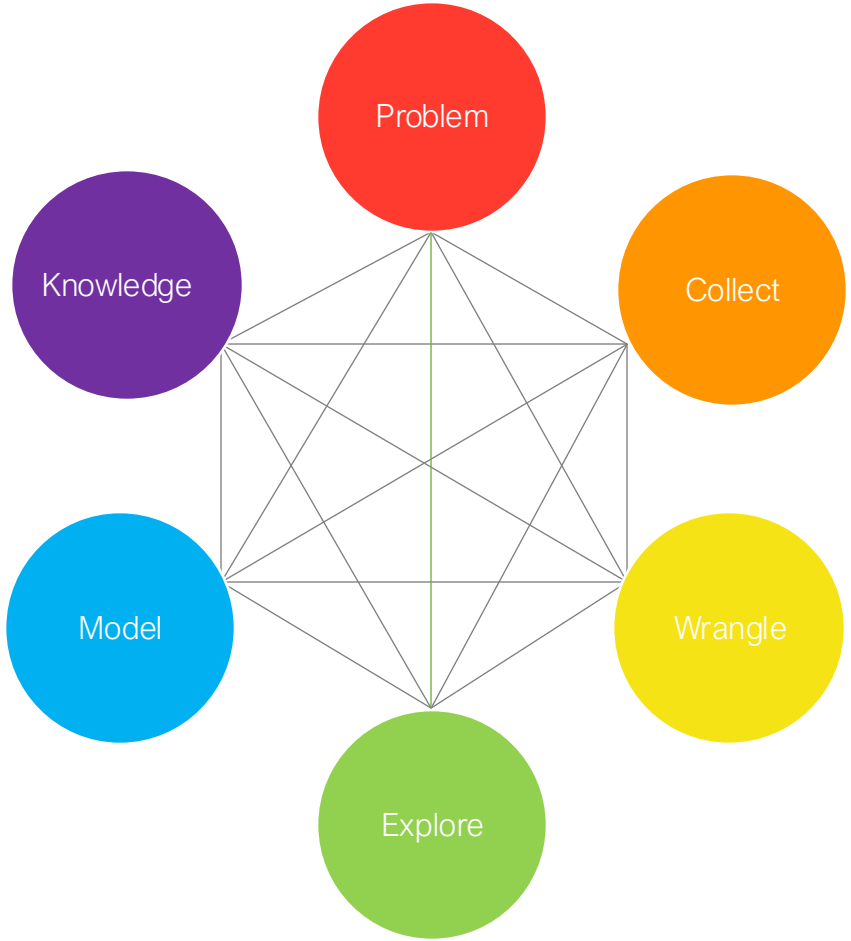
- R Packages for creating impressive plots: **ggplot2( )**
  - You will have to first install ‘ggplot2’ package first with the **install.package( )** function
  - After its installed, you will need to load the package into R with **library( )** function

- All this will be become clear in the practical session

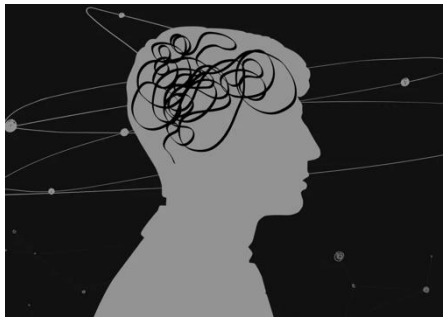
Data Sourcing

Contents

- What's data sourcing?
- The distinction of primary and secondary data sources
- Examples of sourcing data (based on my own research experience)
- Summarising the merits and limitations of using either primary or secondary data
- Brief notes on study design and sampling strategy (- make a mental of note of this, or the least, be cognisant of it)
- Description of data
- Final words…



Format of today's lesson goes…



Theory & Application

Student Feedback
5-10 minutes

Comfort Break
15 minutes

Description of Data
(Tips on assessment)

Final words…

Let’s begin teaching…

What is Data Sourcing?

- Data sourcing (or data collection) is referred to how one goes out to gather data – this can be done either 'ACTIVELY' or 'PASSIVELY'

- ACTIVE: these are ways of acquiring records (or data) actively through fieldwork, surveys, interviews etc.,

- PASSIVE: these are ways of acquiring records (or data) that are made available from some "official" (or "recognised") sources (e.g., NGOs, agencies, companies, government, educational or research institutions etc.,)

- Official (or recognised) sources can make their data freely available to everyone to use & republish, repurpose it as they wish without any patent, legal or copyright restrictions, or through a secured digital channel (e.g., Data Safehaven)

- On the fourth point, we often refer to this term as "**Open Data**" [LINK] for datasets that are freely available. While, the term "**Safe-Guarded Data**" is used for those acquired through secured channels.

- **Open & Safe-Guarded Data** are passive & falls under the broad type of data classification called "**Secondary Data**", whereas the second point on active data collection is called "**Primary Data**"

Types of data sources

Primary data

This type of data source refers to the firsthand data gathered by the user, researcher or enumerator through fieldwork, interviews, questionnaire surveys etc.,

Secondary data

This type of data source typically refers to the data that has already been collected through a primary source and made readily available for other researcher(s) to use for their study or investigation

Internal	External

Secondary data

This type of data source typically refers to the data that has already been collected through a primary source and made readily available for other researcher(s) to use for their study or investigation

Internal Secondary data

- If you are working within an organisation that has some relevant data of interest
- The organisation has collected the data and can provide access to such users working within it or collaborators
- Data scientists working within the UK Metropolitan Police Service seeking to quantify the burden of various crime outcomes: Burglary, sexual assault, vandalism, arson and so…
  - Data is collated by MPS and not by the data scientists
  - So MPS can release the data to such users, this is an example of **Internal Secondary Data**

Secondary data

This type of data source typically refers to the data that has already been collected through a primary source and made readily available for other researcher(s) to use for their study or investigation

External Secondary data

- This simply refers to acquiring open datasets from external sources
  - Open-source websites (freely accessible)
  - Paying for the data (requires a license)
  - Online data service which is free but requires users to register etc.,
  - Release for FOI (Freedom of Information)

Sometimes you can combine both internal and multiple external data sources to supplement each other!

The umbrella term ‘secondary data’ can also be referred to as “routinely collected data”

Examples of Sourcing Data

Example of primary data: **Development Frontiers in Crime, Livelihoods and Urban**

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

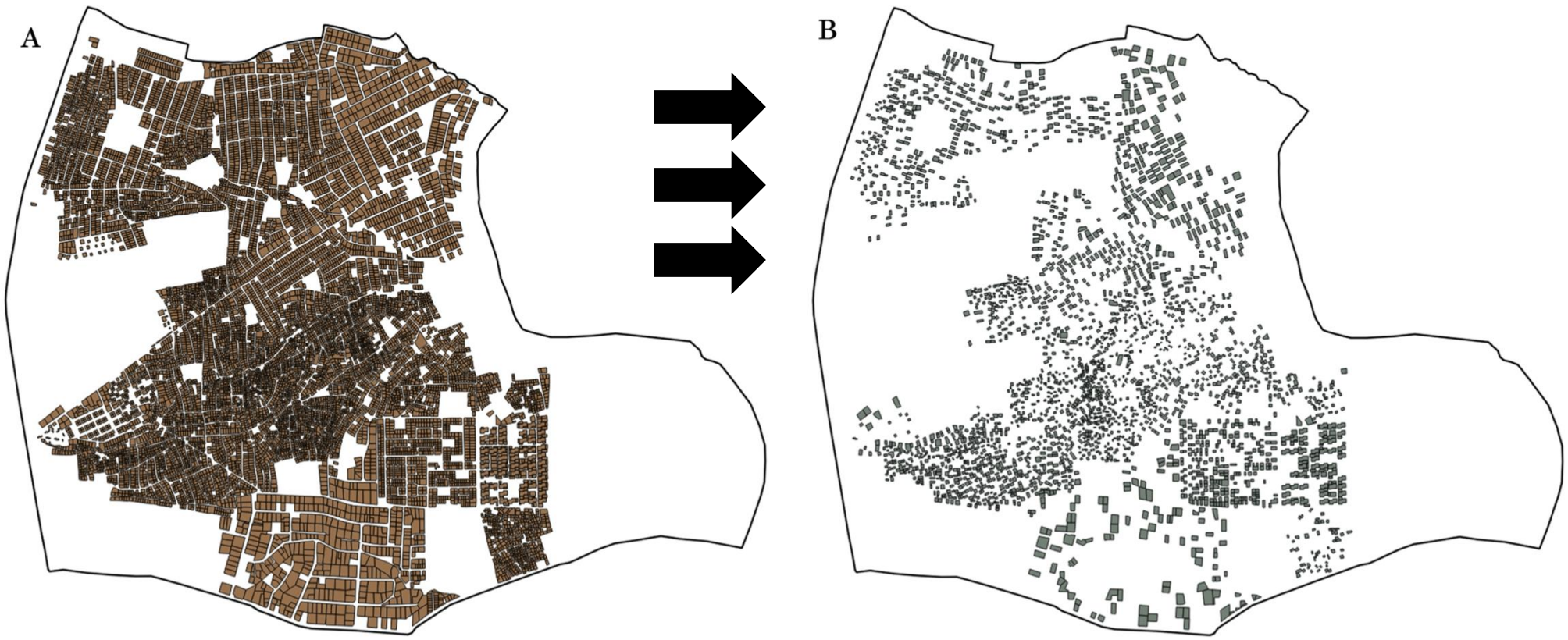
Conventional analyses of crime, based on European research models, are often poorly suited to assessing the specific dimensions of criminality in Africa. Development Frontiers in Crime, Livelihoods and Urban Poverty in Nigeria (FCLP) aims to provide an alternative framework for understanding the specific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

Example of primary data: **Development Frontiers in Crime, Livelihoods and Urban Poverty in Nigeria (FCLP)**

See source(s):

1. Musah et al, 2020
2. Umar et al, 2020
3. Umar et al, 2017



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

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)]

Houses mapped in B, we had to interview all those residents (i.e., 2,300 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	Traffic: Pedestrian	Traffic: Vehicular	Road: Type:	Road: Condition:	Surveillance: Morning	Surveillance: Afternoon	Surveillance: Evening	Surveillance: CCTV Camera:	Surveillance: Vigilante Post:	Surveillance: Street lights:	Surveillance: Street gates:	Signs of dis-order: Drainage:	Signs of dis-order: Encrochment:	Signs of dis-order: Garbage:	Signs of dis-order: Graffiti:
1			Occupied	Shops	Flat	Cement	High walls															
2			vacant	Kiosk	Storey	Mud	Burglary-proof															
3			Abandoned	In-trading	Attached	Temporary	Gate															
4			Construction	Out-trading	Detached		Garage															
							Outdoor sitting															
							Security Lights															
							Guards															
							Open Access															
							Danger Warning															
							CCTV Camera															
							Dogs															

Sociodemographic variables

Section A: Questions related to household Details

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_____

+

Perception and safety variables

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

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 [ ]

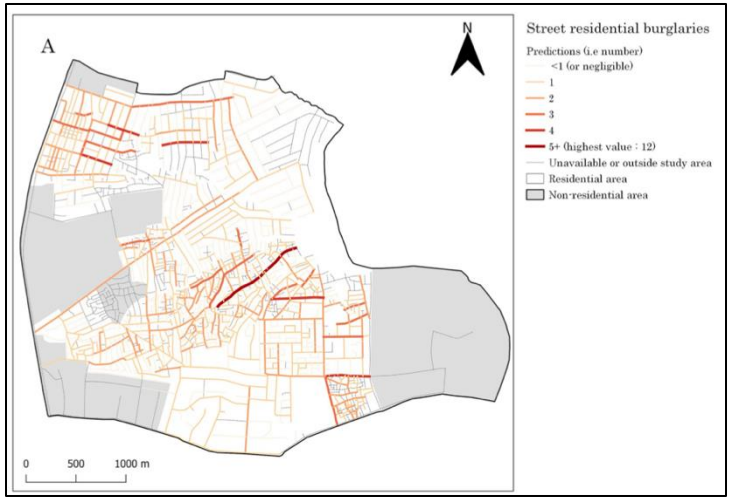
+

Victimisation (dependent) variables

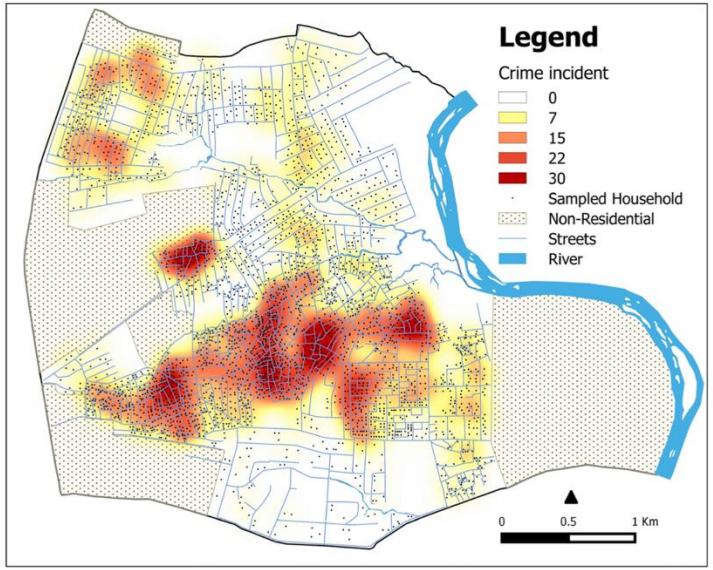
Section C: Questions related to incidents that had happened within your property

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])

Themes:

1. **Public heath**
2. **Epidemiology**
3. **Tropical Diseases**
4. **Human Geography**
5. **Quantitative Research**
6. **Global South**

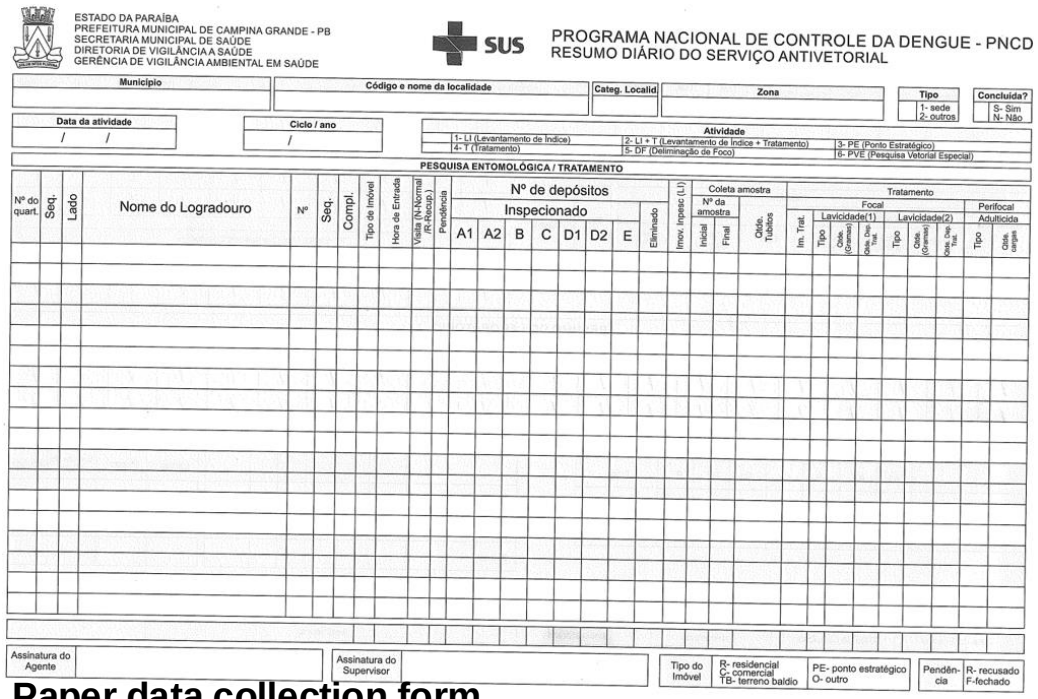
Example of primary data: **Adoption of mobile phone technologies & IoTs for surveillance of mosquito populations**

- Currently work on a collaborative project with academics from UCL, Turkey & Centre of Environmental Surveillance in Recife and Campina Grande
- We are piloting cell phone applications (since August 2019) to support agents in collecting new information on household-levels of mosquito infestation
- Early warning detection of dangerous mosquito-borne arboviruses (e.g., Zika, Dengue etc.)
- Mapping and prediction of hotspots and breeding sites
- Combating the social, environmental and climatic-related determinants of increased mosquito abundance

Environmental agents carry out on a bimester interval, i.e., visits to houses and residential premises to rid of potential breeding habitats and infestation

Mapping and prediction of hotspots and breeding sites

Example of primary data: **Adoption of mobile phone technologies & IoTs for surveillance of mosquito populations**


ESTADO DA PARAÍBA
PREFEITURA MUNICIPAL DE CAMPINA GRANDE - PB
SECRETARIA MUNICIPAL DE SAÚDE
DIRETORIA DE VIGILÂNCIA A SAÚDE
GERÊNCIA DE VIGILÂNCIA AMBIENTAL EM SAÚDE

SUS

PROGRAMA NACIONAL DE CONTROLE DA DENGUE - PNCD
RESUMO DIÁRIO DO SERVIÇO ANTIVETORIAL

Município | Código e nome da localidade | Categ. Localid | Zona | Tipo (1- sede, 2- outros) | Concluída? (S- Sim, N- Não)

Data da atividade | Ciclo / ano

Atividade
1- LI (Levantamento de Índice) | 2- LI + T (Levantamento de Índice + Tratamento) | 3- PE (Ponto Estratégico)
4- T (Tratamento) | 5- DF (Delimitação de Foco) | 6- PVE (Pesquisa Vetorial Especial)

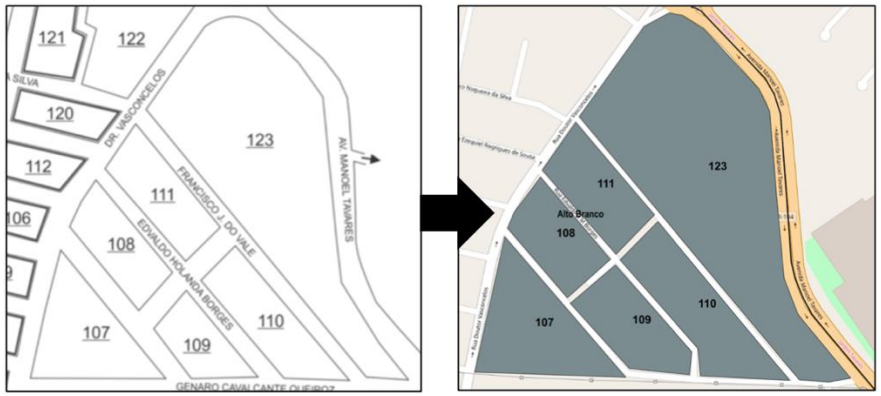
PESQUISA ENTOMOLÓGICA / TRATAMENTO

Nº do quart. | Seq. | Lado | Nome do Logradouro | Nº | Seq. | Compl. | Tipo de Imóvel | Hora da Entrada | Visita (N-Normal, R-Recuperada) | Pendência | Nº de depósitos Inspecionado (A1, A2, B, C, D1, D2, E) | Eliminado | Coleta amostra (Imov. Inspec. (LI)) | Nº da amostra (Inicial, Final) | Qtds. Tubitos | Tratamento (Focal: Im. Trat., Lavicidade(1): Tipo, Qtde. (Gramas), Qtde. Dep. Trat.; Lavicidade(2): Tipo, Qtde. (Gramas), Qtde. Dep. Trat.; Perifocal Adulticida: Tipo, Qtde. (ampl.))

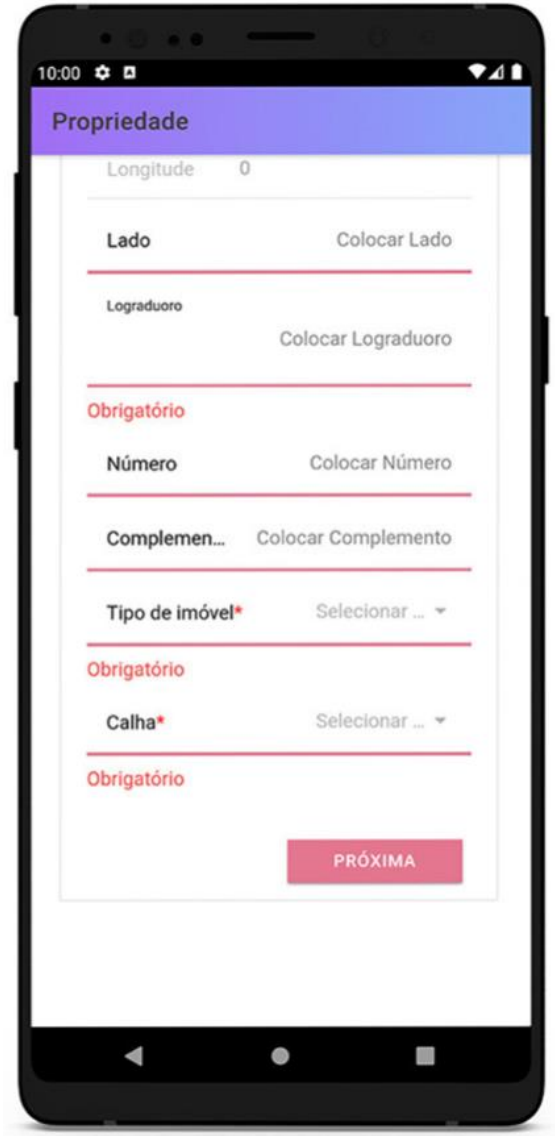
Assinatura do Agente | Assinatura do Supervisor

Tipo do Imóvel: R- residencial, C- comercial, TB- terreno baldio, PE- ponto estratégico, O- outro
Pendência: R- recusado, F-fechado

Paper data collection form

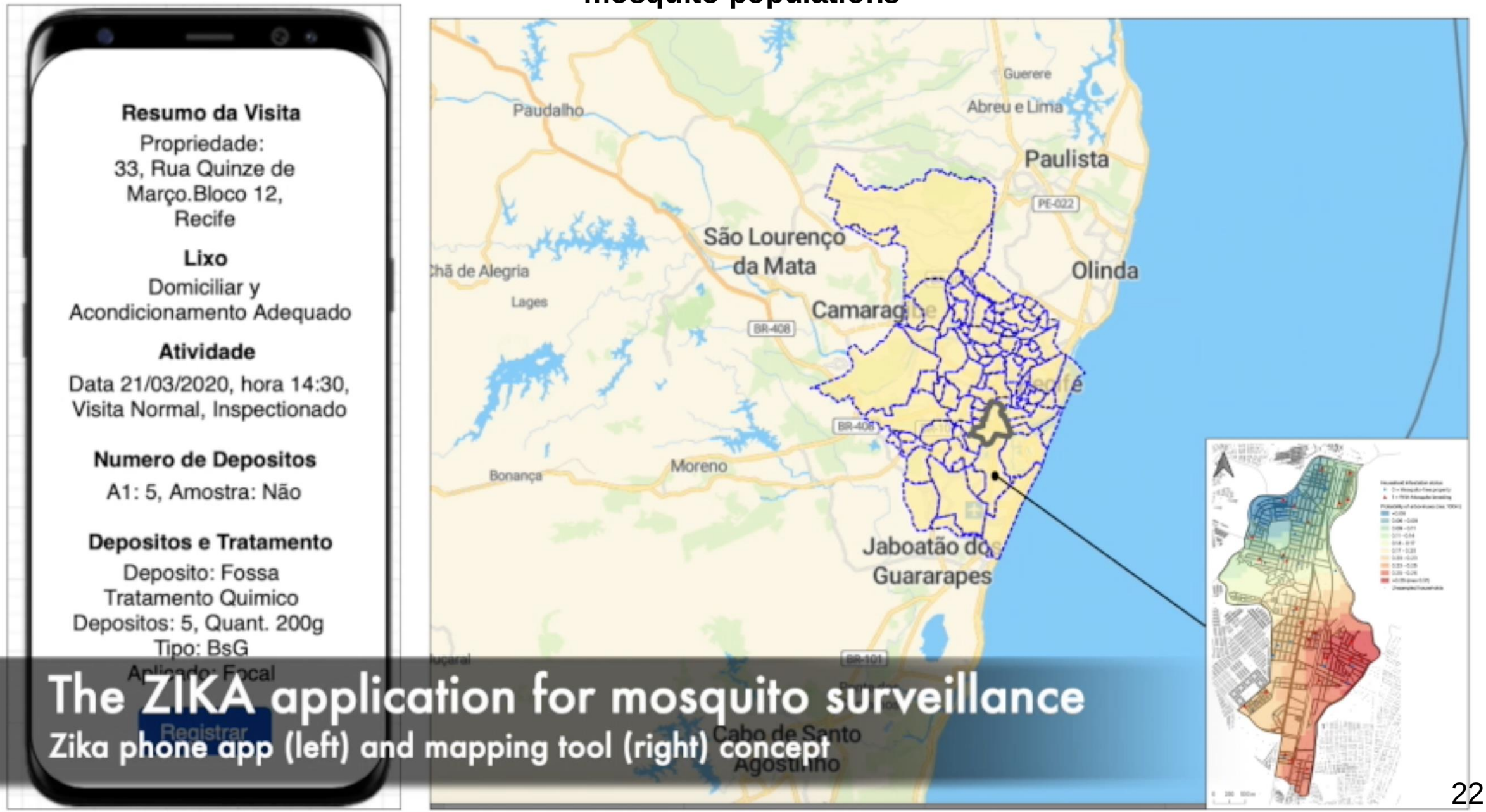


Scanned maps



See sources: Aldosery, A., Musah, A., Birjovanu, G., Moreno, G., Boscor, A., Dutra, L., . . . Kostkova, P. (2021). MEWAR: Development of a Cross-Platform Mobile Application and Web Dashboard System for Real-Time Mosquito Surveillance in Northeast Brazil. *Frontiers in Public Health, 9*. doi:10.3389/fpubh.2021.754072 **[LINK]**

Example of primary data: **Adoption of mobile phone technologies & IoTs for surveillance of mosquito populations**



The ZIKA application for mosquito surveillance

Zika phone app (left) and mapping tool (right) concept

Example of primary data: **Detection of open gutters and drainages for mosquito-borne infestation and sanitation in Accra**

Data collection procedure

We have the equipment for capture video footage, or picture images of the sewer layout in the study areas. These equipment are capable of spatially referencing the taken images

- Clogged with rubbish
- Stagnated ✓
- Clear
- Partially covered
- Covered

- Clogged with rubbish ✓
- Stagnated
- Clear
- Partially covered
- Covered

- ❖ 2 x GoPro cameras (with head straps)

We can develop our own bespoke Machine Learning algorithm and train it to classifying these images accordingly

Combined field survey

Primary outcome of interest

Measurable disease outcome
- Retrospective health survey (Malaria)

Secondary outcome(s) of interest(s)

Measurable disaster-related outcome
- Domestic Floods and vulnerability

Measurable data for urban landscape
- Open sewer Infrastructure and network

Measurable entomological outcome
- Mosquito abundance

Measurable environmental outcome
- Broader WASH index

This data is collected with the GoPro cameras and images with be trained and classified with a ML classification model

Clogged with rubbish
Stagnated ✓
Clear
Partially covered
Covered

+

We will visit properties to conduct the following survey:
- Household survey on sanitation
- Household survey on flood risk
- Infestation survey (interior and exterior) for mosquito breeding habitats

This data is collected with the tablet application tool i.e., Survey CTO with a trained enumerator.

Figure 1: Drainage survey with GoPro camera. All points represent locations of where pictures were taken to capture exposed drainages/gutters in Sabon Zongo (New Settlement), Accra, Ghana.

From the drainage survey, we manage to collect up to 2,033 sample points where these drainages/gutters were exposed.

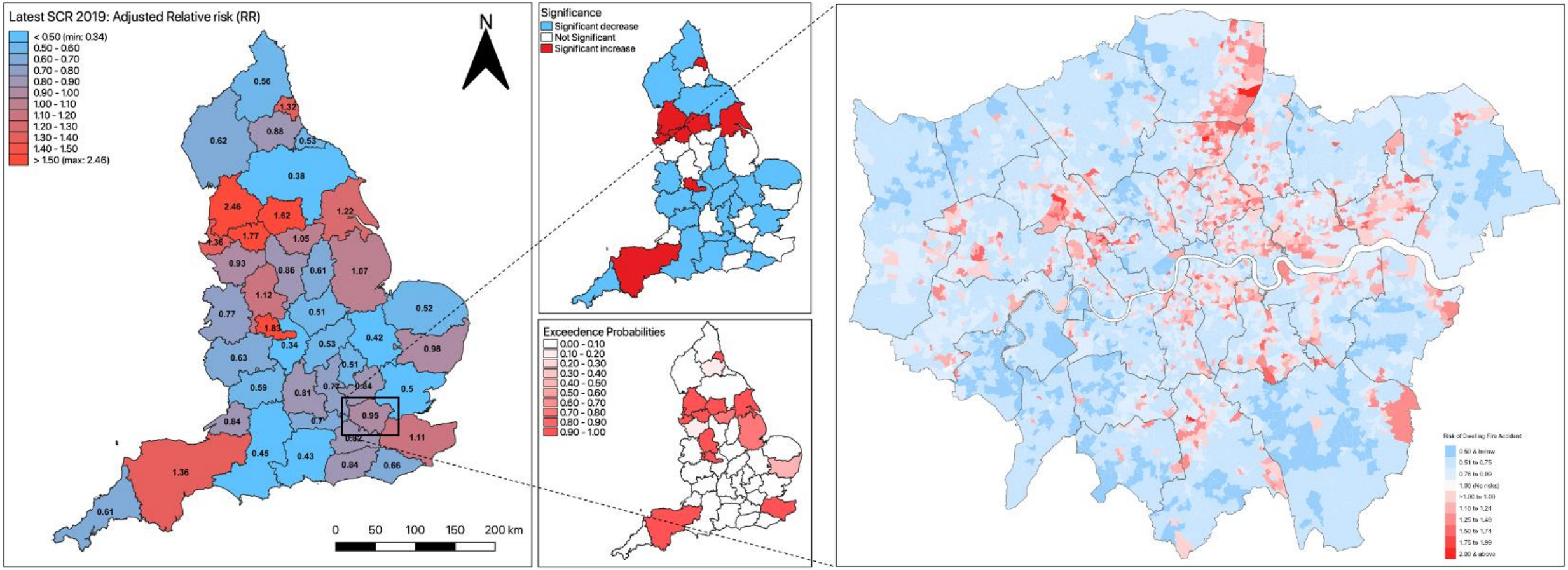
Figure 2: A - clean drain, B – stagnated drain; C – Drainage choked with rubbish; D – Run-off drain where the filth has overflown out of the drain and on to land. These images were drawn at random for the drainage survey and were implemented as exhibits in a focus group to test how the respondents perceive the health and sanitary dangers of these exposed drainages.

Format: three focus group sessions—one for 10 men only, one for 10 women only, and one mixed group for 10 youths (totalling 30 respondents)—were conducted to gauge their 'lived experiences' regarding the sanitation situation and gutters.

Qualitative data: 139 pages worth of information.

Examples of open primary data for health research in Global South

Example of secondary data: **UK GOV & Fire Hazards and Fire-related Casualties Data [LINK]**



We used **external secondary data** from official statistics pertained to fire hazards and casualties for England and combined with **internal secondary data** from UCL's CDRC registry to account for socioeconomic deprivation

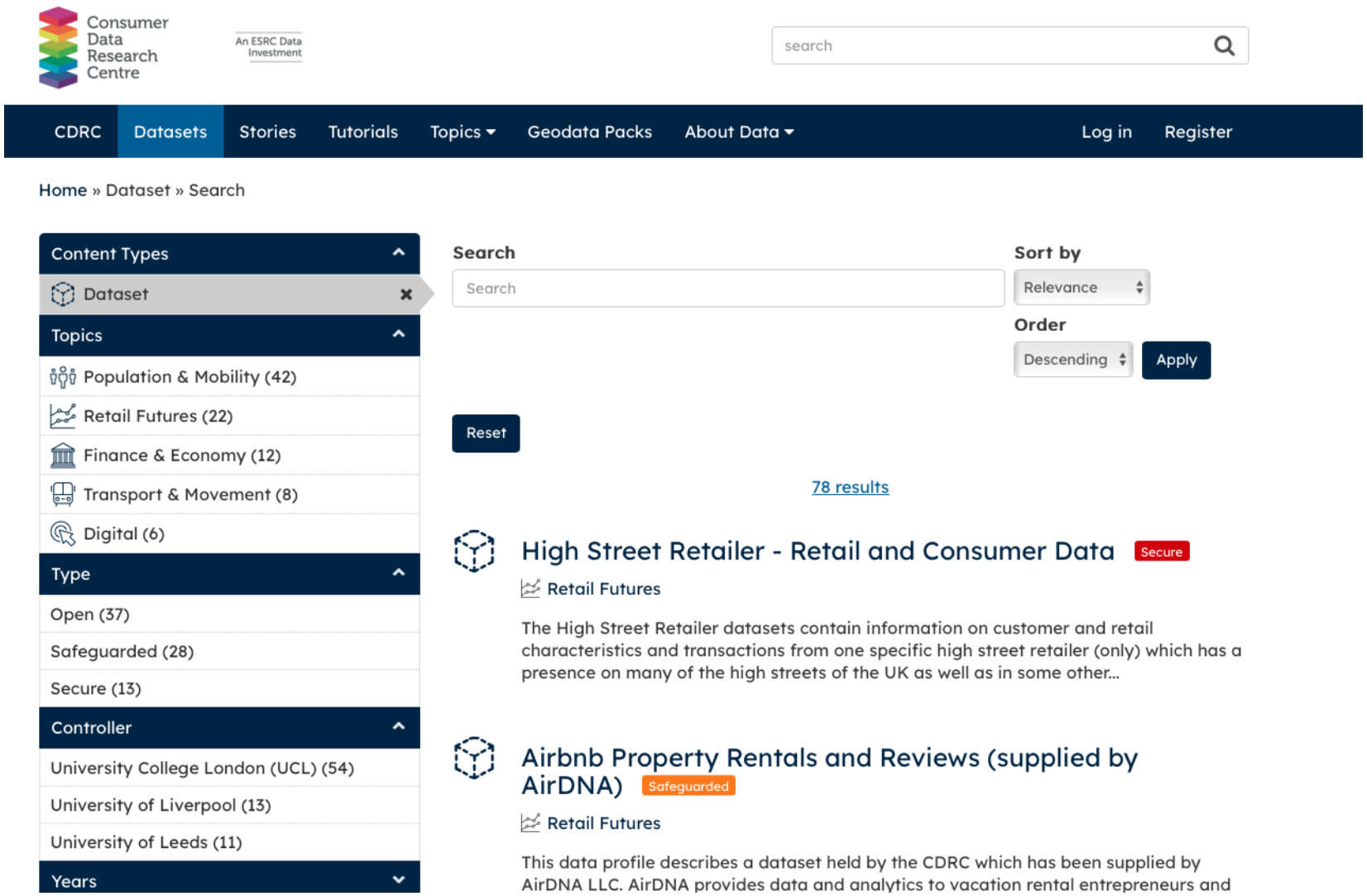
See source(s): Li, L., Musah, A., Thomas, M. G., & Kostkova, P. (2022). An ecological study exploring the geospatial associations between socioeconomic deprivation and fire-related dwelling casualties in the England (2010–2019). *Applied Geography, 144*, 102718. doi:10.1016/j.apgeog.2022.102718 [LINK]

Data sources: UK GOV Fire statistics incident level datasets [LINK]

Themes:

1. **Disaster risk reduction**
2. **Social Sciences**
3. **Human Geography**
4. **Quantitative Research**

Example of secondary data: **Consumer Data Research Centre (UCL)**



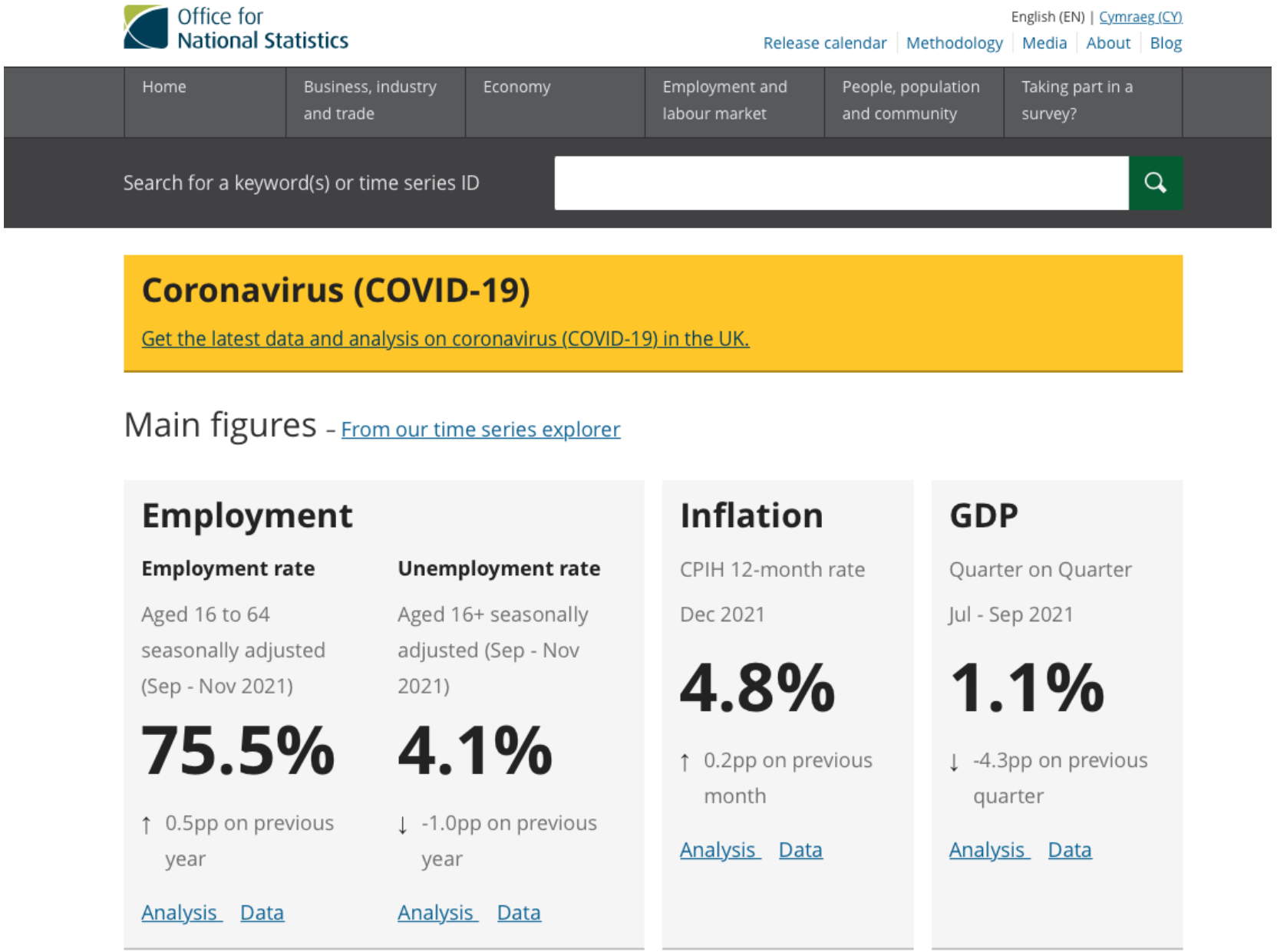
CDRC provides data for research to address societal and economic challenges in the UK.

Contains tonne of open, safeguarded and secure data.

Website: https://data.cdrc.ac.uk/search/type/dataset

Example of secondary data: **Office for National Statistics (ONS)**

LINK: https://www.ons.gov.uk



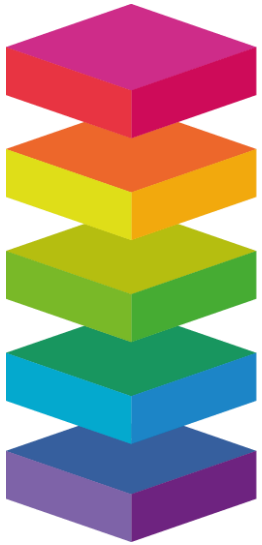
ABOUT ONS:

It is responsible for the collation census data relating to the UK population, as well as the publication of statistics related to the UK economy, it's population and wide range of societal matter as a whole.

UK government makes use of these statistics for their policy and decision making.

Actual raw UK data can be found from ONS and UK GOV.

Examples of open secondary data for social and political research in UK



Consumer Data Research Centre

Office for National Statistics

census 2021 [LINK]

GOV.UK

IMD: [LINK]

Ordnance Survey

LONDON DATASTORE [LINK]

DATA.POLICE.UK [LINK]

Merits & Limitations

Advantages and disadvantages of Primary data sources

Advantages

- Data collected is always up-to date
- Relevant and specific to user's research aims and objectives
- High-level, and greater of understanding about the nature and content of the dataset
- High-level of accuracy as long as you do whatever it is in your power to minimise all kinds of systematic errors in the data collection process (ensuring **Internal Validity**)

Disadvantages

- Depending on the study design – data collection is a very time consuming and expensive process
- If you are collecting personal and sensitive data, you must **apply** for ethnical approval before going out to get your data
- You will have to clean, manage and maintain your own data
- Possibility to falsify his/her data since one has his/her autonomy over the data

Advantages and disadvantages of Secondary data sources [2]

Advantages

- Ease of access and low cost, or even free if the data is from an open-source platform.
- Time-saving, especially if the data has already been processed and cleaned
- Unlike primary data; secondary data are often collected routinely hence allowing for longitudinal analysis
- Often secondary data are combined with different sources making it have a variety of variables.

Disadvantages

- You have no control over the quality
- The secondary data might not be specific to your needs
- The data can be biased in favour of the one whose gathered it. User will not know of this data artefact!
- You are not the owner and will never have full understanding of its nature & how it was collected (i.e., its essentially a “Blackbox”)

Brief notes on Study Design & Sampling

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 simply put it in plain English:

"… it's someone's blueprint for answering a research question"

Types of Research Design

Quantitative

This area allows the researcher to derive meaningful insight (or empirical evidence) about certain phenomena through analysis of numerical information

- Descriptive studies
- Observational Studies
- Experimental Studies

Purpose: Evidence of causality (or an association); Internal validity and External validity

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

Types of Study Design

Type	Study Design	Properties
Descriptive	• Ecological (or geographic)	• Unit of observation are at group- or aggregated level (e.g., geographic unit)
	• Cross-sectional	• Data is collected on a single snap in time, and its useful for generating further hypothesis
Observational	• Case-control	• Observe the effect of an independent variable(s) on an outcome that has already occurred. The time frame is a retrospective analysis
	• Cohort (or longitudinal)	• Observe the effect of an independent variable(s) on an outcome that has not happened yet. The time frame is a prospective analysis
Experimental	• Randomised control trials (RCT)*	• Observe the effect in an actual test group(s) (e.g., intervention, effectiveness of a teaching programme, clinical trials for a vaccine)

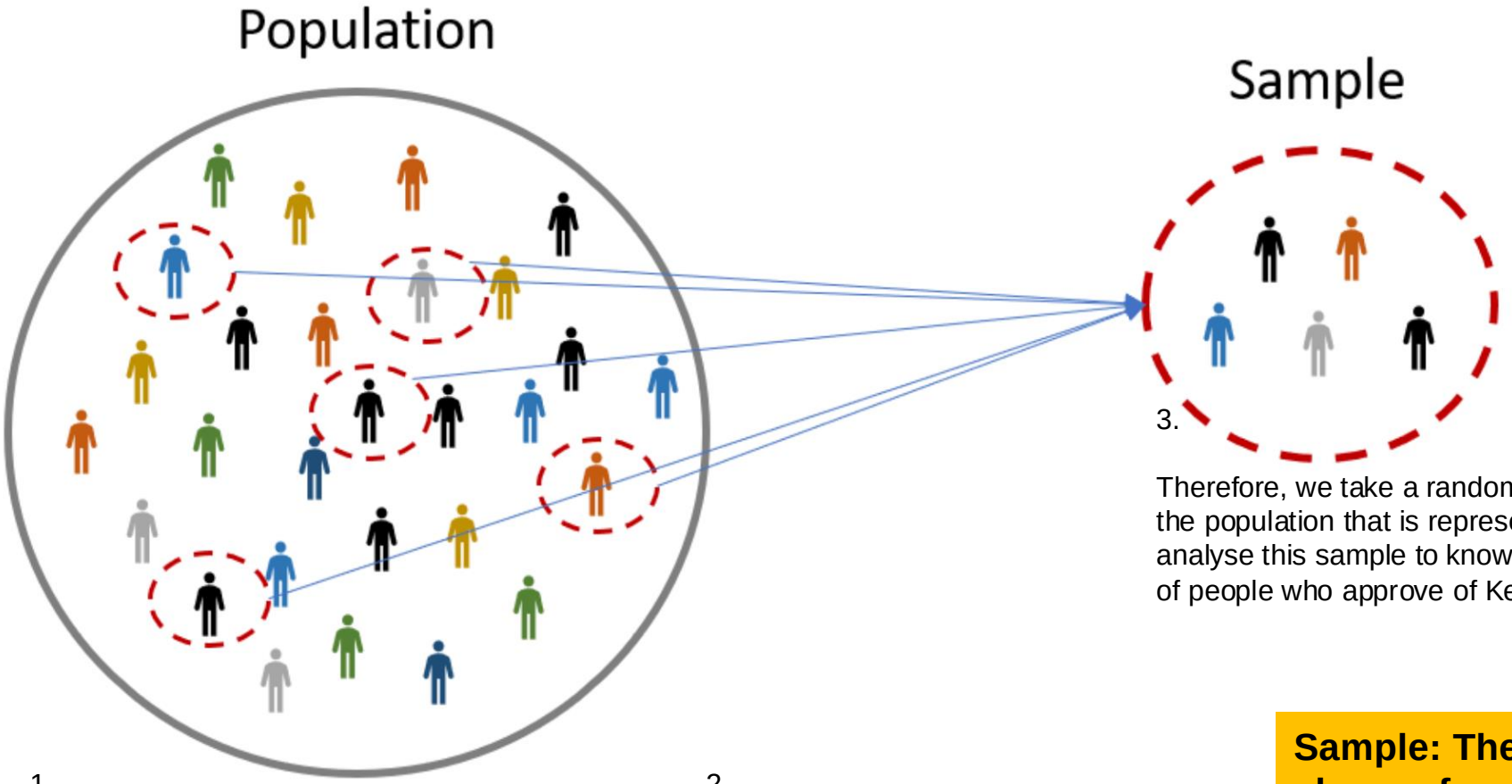
Pilot study is a special cases as it could be anyone of these study designs

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
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

Sampling Strategy [1]: Recall the Sir Keir Starmer's example in Week 1?

Parameter: The proportion of people in the UK population who voted for Keir Starmer [aka "my dad was a *toolmaker*"] (p)
Statistic: The sample proportion of people from the UK population voted for Keir Starmer [aka "my dad was a *toolmaker*"] ($\hat{p}$)



1.

Say, we want to know the percentage of people in the whole of the UK population who approve of Keir Starmer

2.

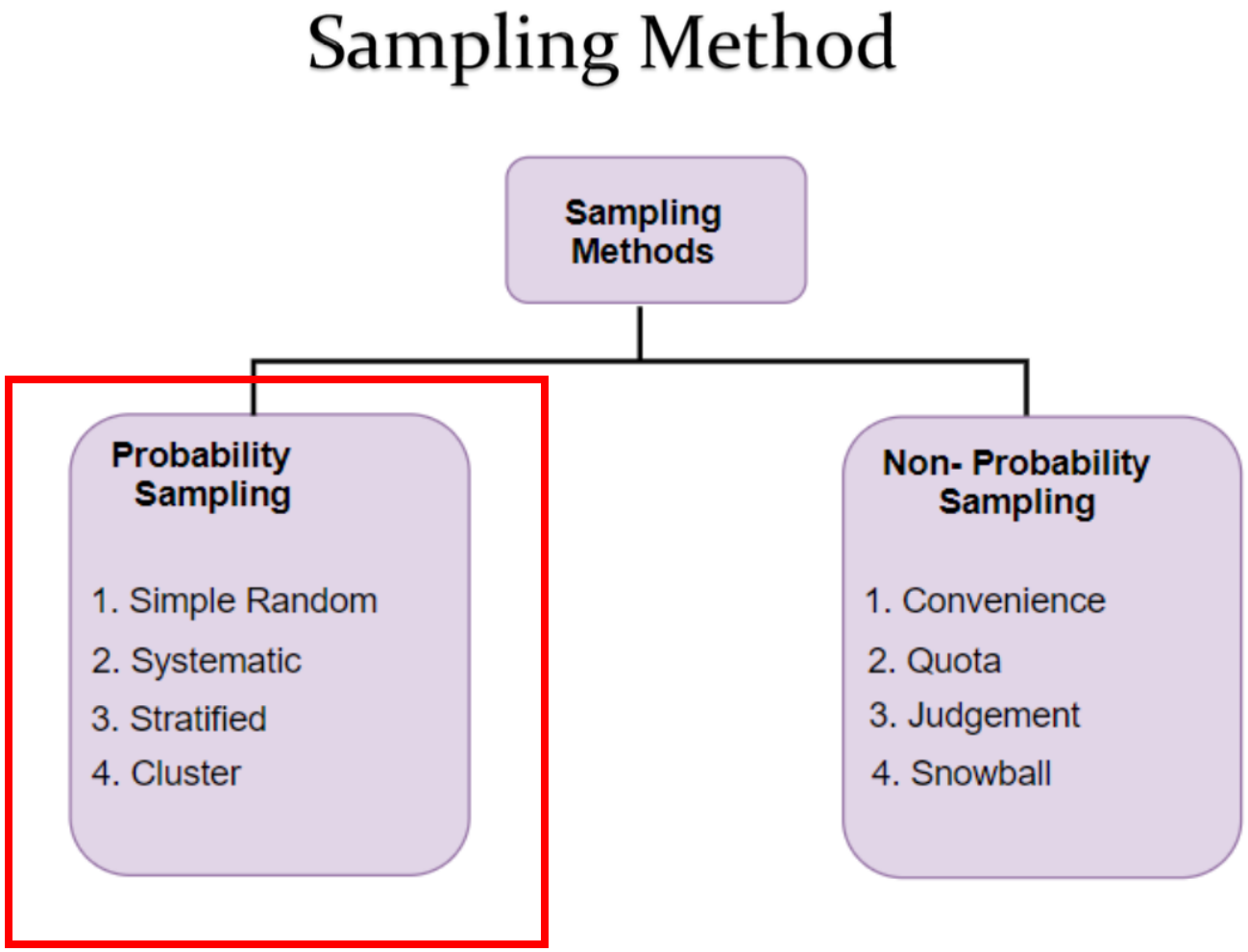
It is impossible, or it very tedious for us to collect data on everyone in the UK population

3.

Therefore, we take a random sample from the population that is representative of UK, analyse this sample to know the percentage of people who approve of Keir Starmer

Sample: The subset of subjects chosen for study from a population is done through random data sampling strategy

Sampling Strategy [2]



Notes:

- **Simple Random Sample** is a subset of population in which each member of the subset has an equal chance of being select at random.
- **Systematic Random sampling** is another probability sampling approach where you have a list of individuals from a target population and perform a selection by starting somewhere at random and then collecting data at a fixed interval from that starting point.
- **Stratified Random Sampling** is an approach that requires one to divide the population and group them by a common characteristic, and then start sampling individuals from each groups at random
- **Cluster Sampling** is an approach that requires one to select a cluster unit(s) at random and then survey everyone in the selected cluster unit (usually smaller populations such a villages, or small schools in remote areas etc.,)

Warning: You cannot use any statistical method on samples that were collected with non-probabilistic approach – that element of randomness has been removed and thus render the sample not representative of the population. If you intend to use the quantitative route do not use a non-probability sampling approach

POLS0008: Course Evaluation & Student Feedback (2024/25)

https://forms.gle/bpDBWw7CcBR7wvX96

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 **POLS0008: Introduction to Quantitative Research Methods**. We would like to receive your feedback regarding the sessions for week 1, 2, 3 (and 4). Your feedback is 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 Stephen.

原作　矢立肇

Breaktime

Description of data

Description of Data [1]

BEFORE YOU START THE ANALYSIS - you are required to provide some information about the dataset you are using regardless if its primary/secondary

- **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

- **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.,)

Description of Data [2]

[continue]

- **Describe the variables that's going to be used for the analysis:**
  - **Codebook (see next slide) –** 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)

A Video gamer's statistics

We have compiled the following information about the gaming habits of **Anwar Musah** (aka **The-PhD-Gamer**) across 3 console generations i.e., PlayStation 3, 4 and 5.

There are 161 game titles (~6,000 hours of game time) listed in the shared dataset (last updated 01/2024).

PSNProfiles (https://psnprofiles.com/)

Data Descriptor: [Downloadable Dataset]

This table is an example of codebook!

Variables Names	Descriptor
Number	[Numeric] Unique Identifier
GameTitle	[String] Name of the video game
Genre	[String] Type of genre (9 categories)
Platform	[String] Type of console (3 categories)
HourPlayed	[Numeric] Total number of hours invested in a game
CompletionRate	[Numeric] Percentage of completed content in a game
Status	[String] Current status of the game in terms of play is 'in-progress' or 'quit', or in 'hiatus' or 'done' as in completed (4 categories)
PlatinumTrophy	[Binary] 1 = 'Attained platinum trophy' and 0 = 'No platinum trophy'
DLCTrophies	[Binary] 1 = 'Yes' and 0 = 'No'. Are there any DLC trophies present (annoying feature as it effects the completion rate)?

We are going to use this data as a test dummy. We are just going to access RStudio server and then show how to import this dataset.

Then set you folks going for the practical material to prepare for the seminars on Thursday.

ID	Name	Age	Gender	Pathway	Gamer
0001	Brittany	18	Female	NA	No
0002	Idris	NA	Male	NA	NA
0003	Spike	19	Male	Political	Yes
0004	Lara	20	Female	Geography	Yes
0005	Nana-Kwesi	17	Male	Health	No
0006	Xiaoyu	23	Female	Geography	NA

Raw dataset

Characteristic	Breakdown on what's present in the dataset
Age (in years)	Mean 19.7 (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)

Description of Data [3]

[continue]:

Here, we make an explicit statement about missing data bias and complete case before proceeding with the analysis

- **Complete case analysis:** Here, you are making a declaration that you are sub-setting the dataset to cases who have complete information across **ALL** variables (no missing information), and that you are restricting the dataset & analysis to respondents without missing data.

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

Sample size = 6

Description of Data [4]

[continue]:

Code:
dataset_nomissing <- dataset[complete.cases(dataset),]

Here, we make an explicit statement about missing data bias and complete case before proceeding with the analysis

- **Complete case analysis:** Here, you are making a declaration that you are sub-setting the dataset to cases who have complete information across **ALL** variables (no missing information). You are restricting the dataset to respondents without missing data.

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

Sample size = 6

Delete 3 rows

Description of Data [5]

[continue]:

Code:
dataset_nomissing <- dataset[complete.cases(dataset),]

Here, we make an explicit statement about missing data bias and complete case before proceeding with the analysis

- **Complete case analysis:** Here, you are making a declaration that you are sub-setting the dataset to cases who have complete information across **ALL** variables (no missing information). You are restricting the dataset to respondents without missing data and use it for the analysis.

ID	Name	Age	Gender	Pathway	Gamer
0003	Spike	32	Male	Political	Yes
0004	Lara	29	Female	Geography	Yes
0005	Fiona	28	Female	Health	No

Sample size = 6
Delete 3 rows

Retained sample = 3
Drop-out rate = 50%

- **Missing data bias:** Here, you must acknowledge that due to the sample reducing from what was initially A to B, and restricting it to complete case sample, you have introduced some bias in your results.

Let’s make a mental summary of what we’ve learnt in the past 4 weeks

Any questions?

Final words…

You have my nod of approval… you're ready for inferential statistics!

You're gonna carry that weight…