POLS0008
INTRODUCTION TO QUANTITATIVE RESEARCH METHODS

SOURCING DATA

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

Today's lecture

- 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…

Data Sourcing

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.

- We term i.e., the forth point, as "**Open Data**" [LINK]

- **Open Data** 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 first hand 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 in collaborating with 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 occurring data 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.,

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

Also referred to as “Routinely collected data”

Examples of Sourcing Data

Example of primary data: Development Frontiers in Crime, Livelihoods and Urban Poverty in Nigeria (FCLP) [1]

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.

See source(s):

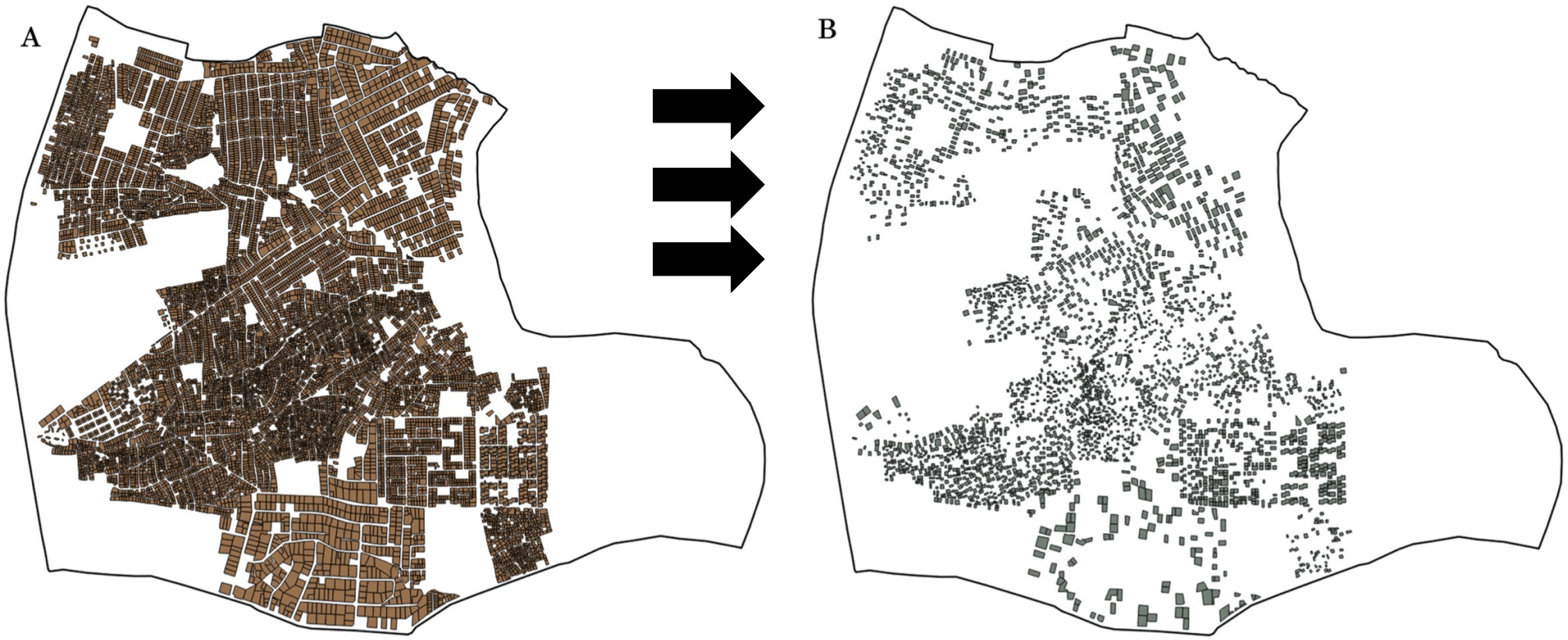
1. Musah et al, 2020
2. Umar et al, 2020
3. Umar et al, 2017

Themes:

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

Example of primary data: 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

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., 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 [ ]

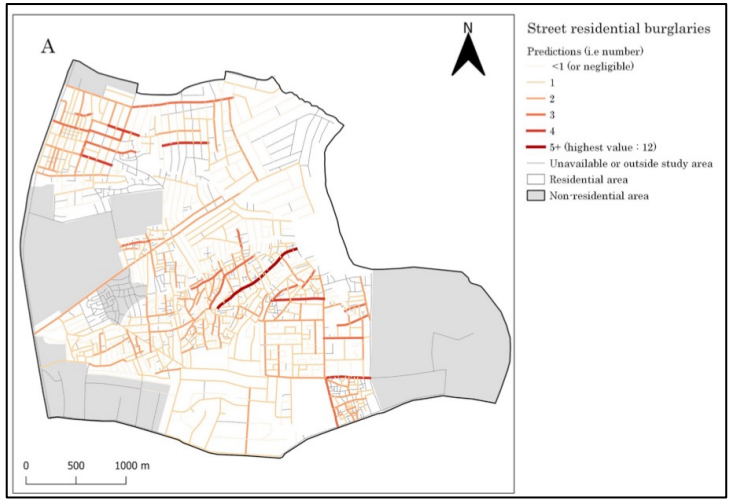
+

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

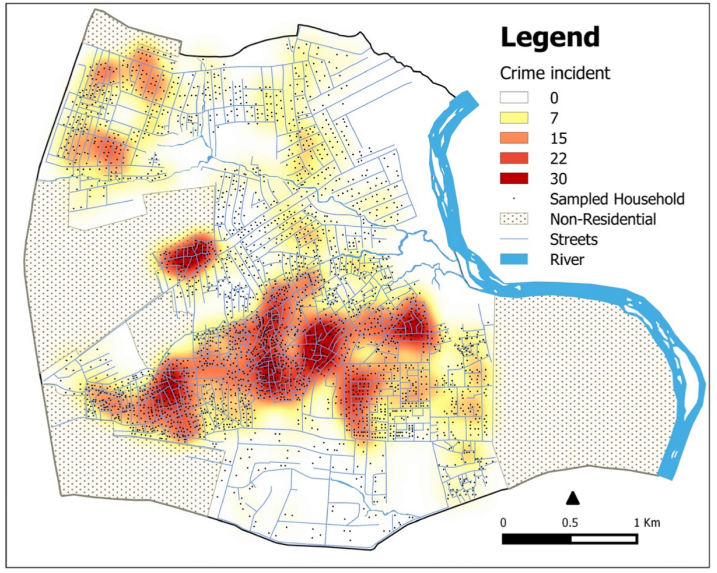
Victimisation (dependent) variables

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])

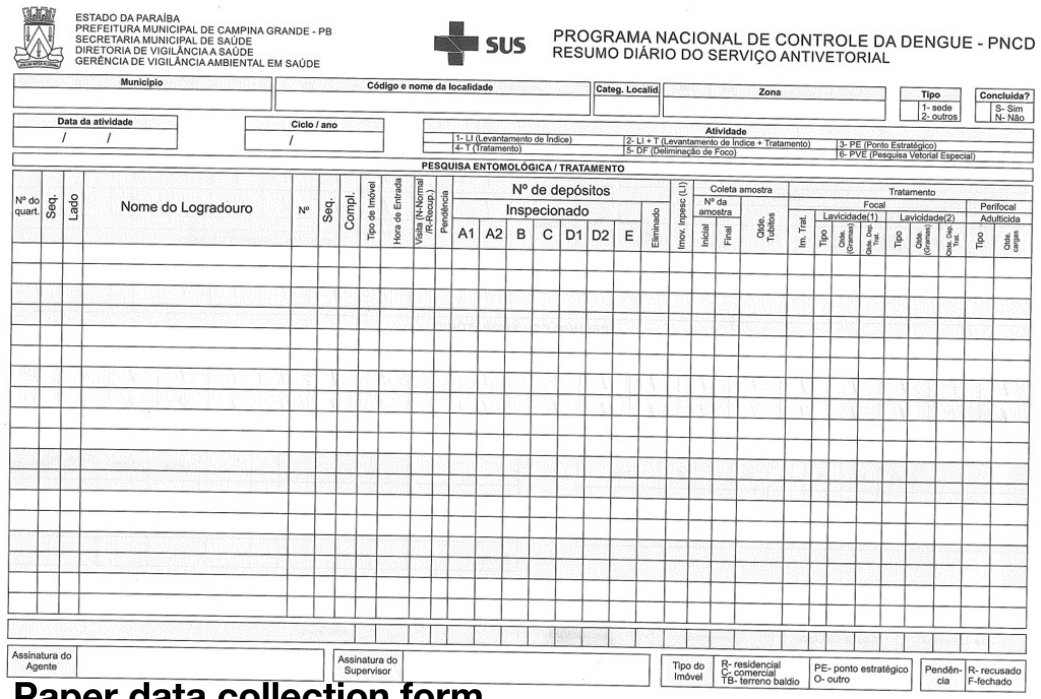
Example of primary data: Adoption of mobile phone technologies & IoTs for surveillance of mosquito populations [1]

- Currently work on a collaborative project with academics from UCL, Turkey & Centre of Environmental Surveillance in Recife and Campina Grande
- We are piloting mobile 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 [2]



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)

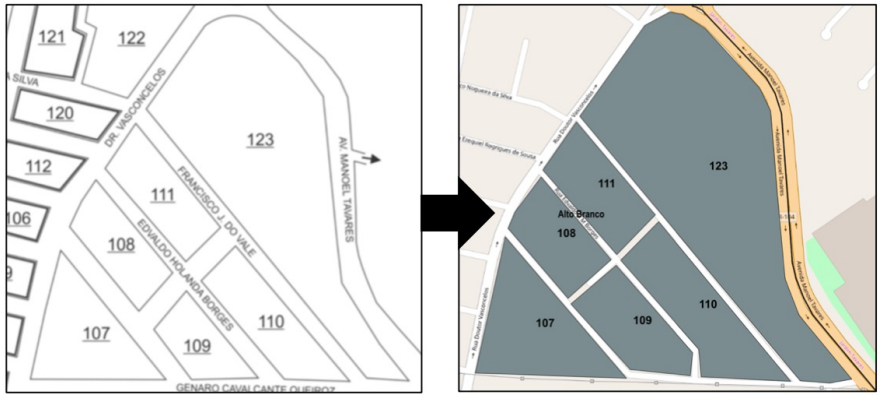
PESQUISA ENTOMOLÓGICA / TRATAMENTO

Nº do quart.	Seq.	Lado	Nome do Logradouro	Nº	Seq.	Compl.	Tipo de Imóvel	Hora da Entrada	Visita (N-Normal, R-Recup.)	Pendência	Nº de depósitos Inspecionado A1	A2	B	C	D1	D2	E	Eliminado	Coleta amostra Inov. Inspec (LI)	Nº da amostra Inicial	Final	Obds. Tubitos	Tratamento Focal Im. Trat.	Lavicidade(1) Tipo	Obds. (Gramas)	Data Dep. Tub.	Lavicidade(2) Tipo	Obds. (Gramas)	PNL. Dep. Tub.	Perifocal Adulticida Tipo	Obds. (gramas)

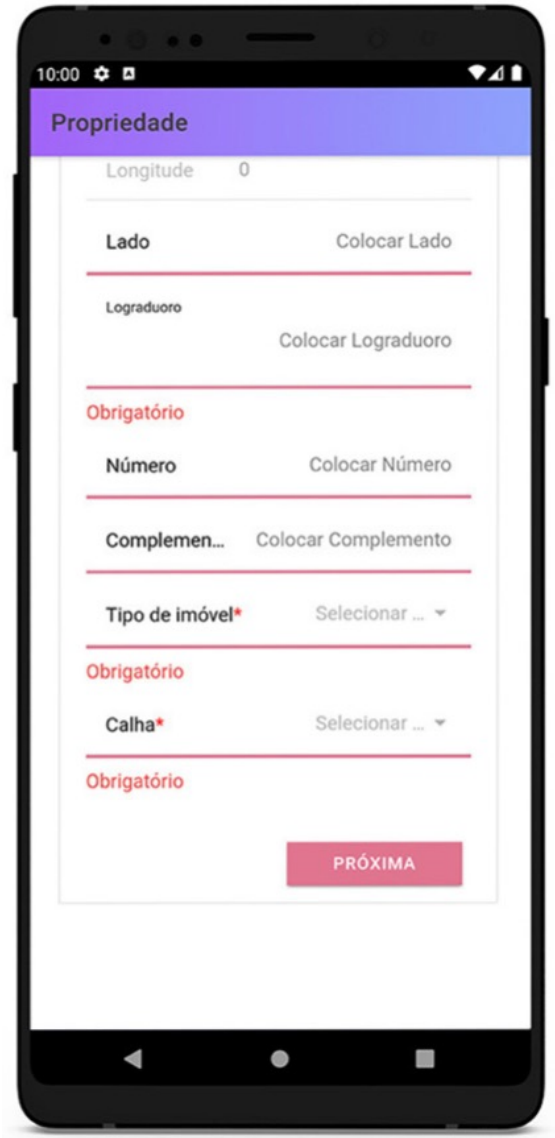
Assinatura do Agente | Assinatura do Supervisor

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

Paper data collection form

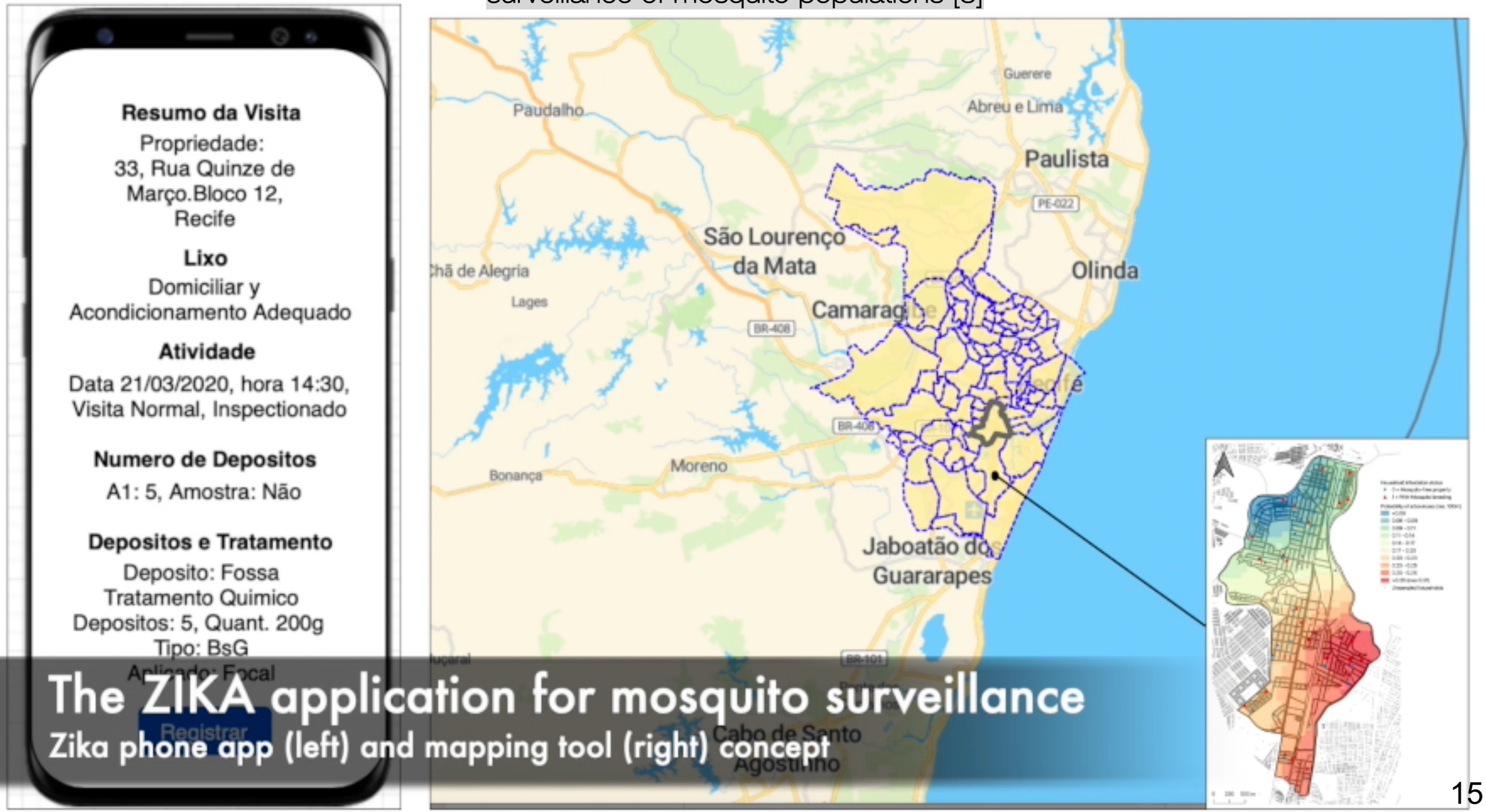


Scanned maps



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 [3]

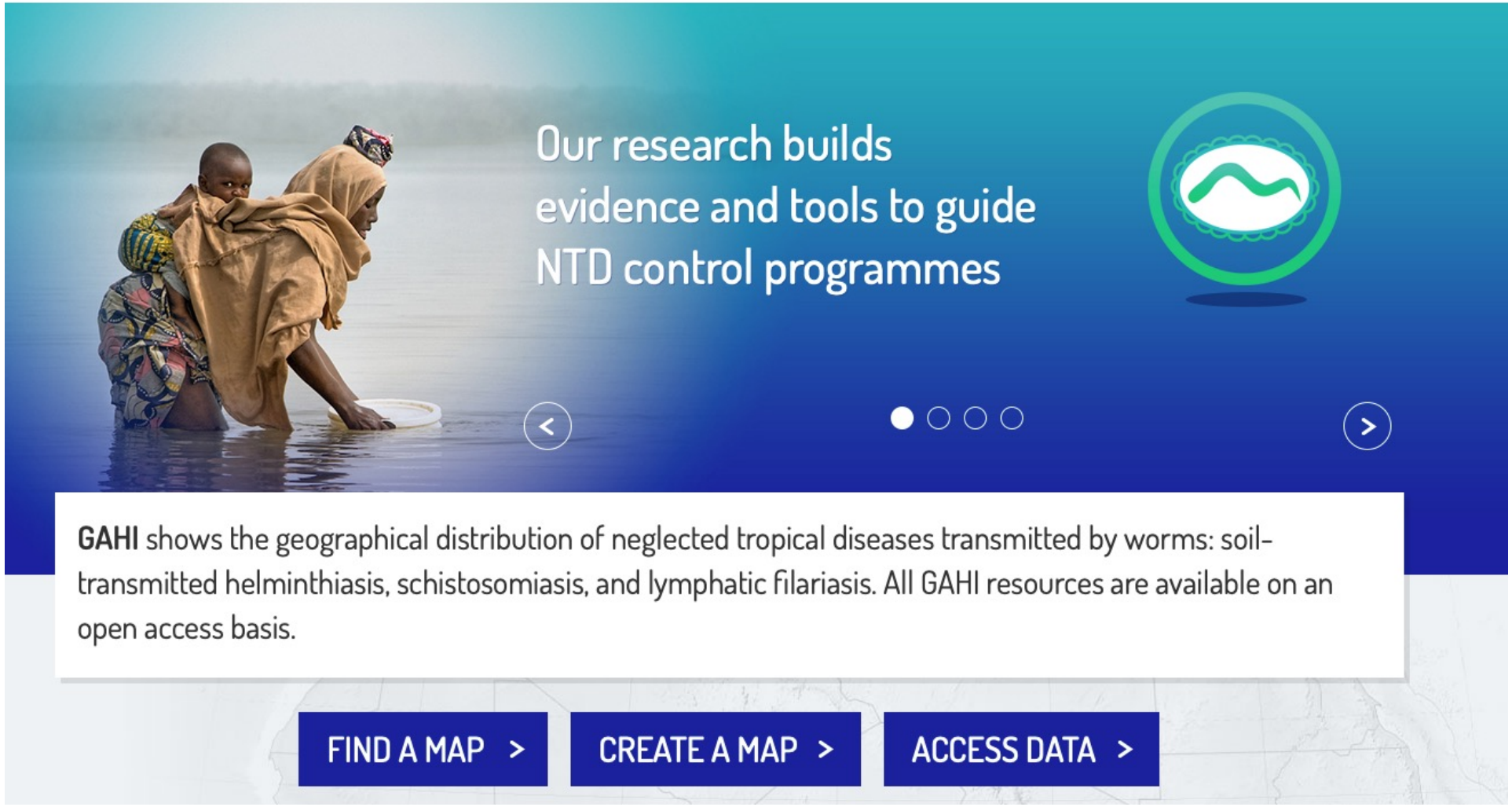


The ZIKA application for mosquito surveillance

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

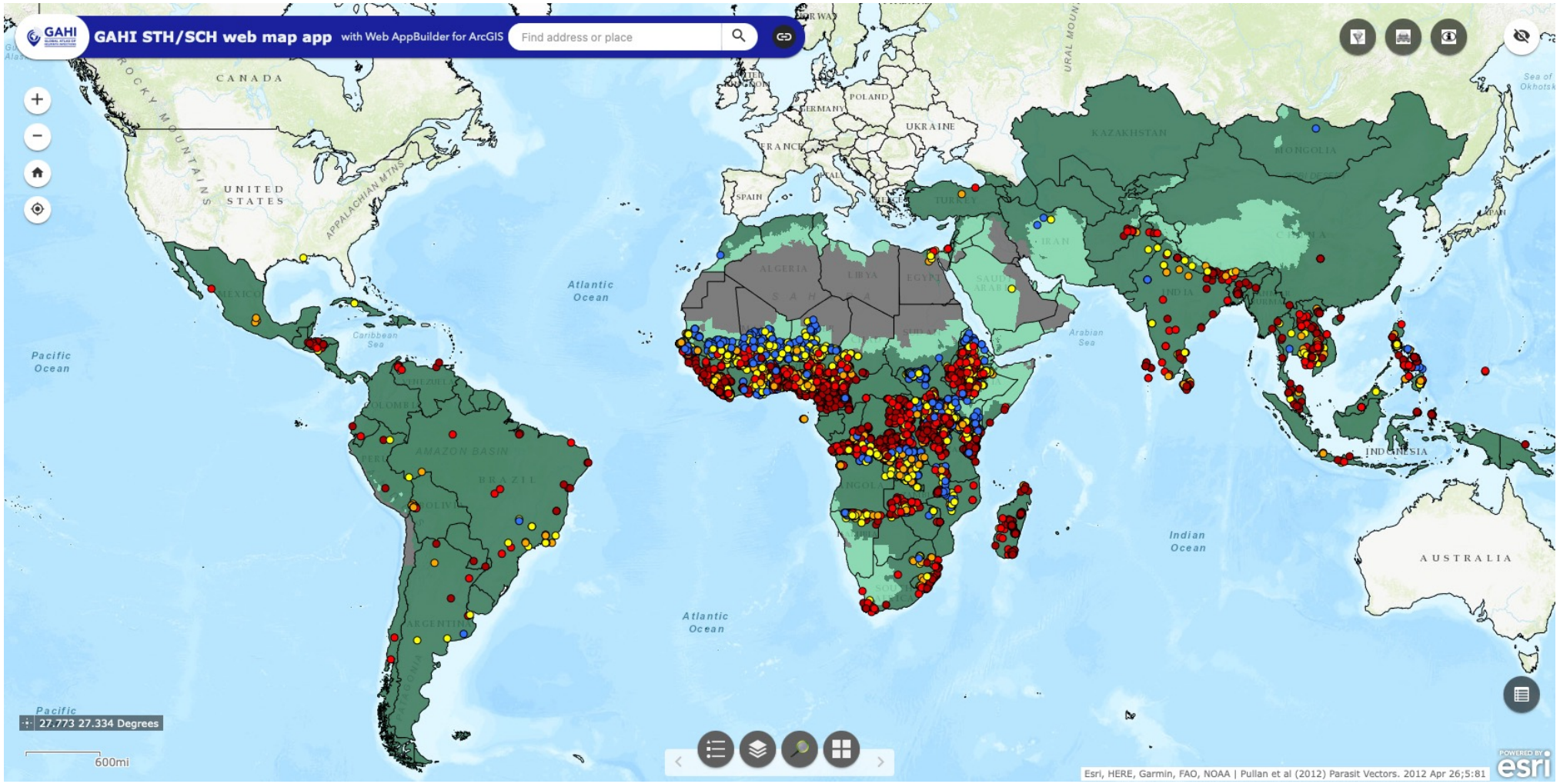
Example of secondary data: Global Atlas for Helminths Infections (GAHI) [1]

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



Website: http://www.thiswormyworld.org

Example of secondary data: Global Atlas for Helminths Infections (GAHI) [2]

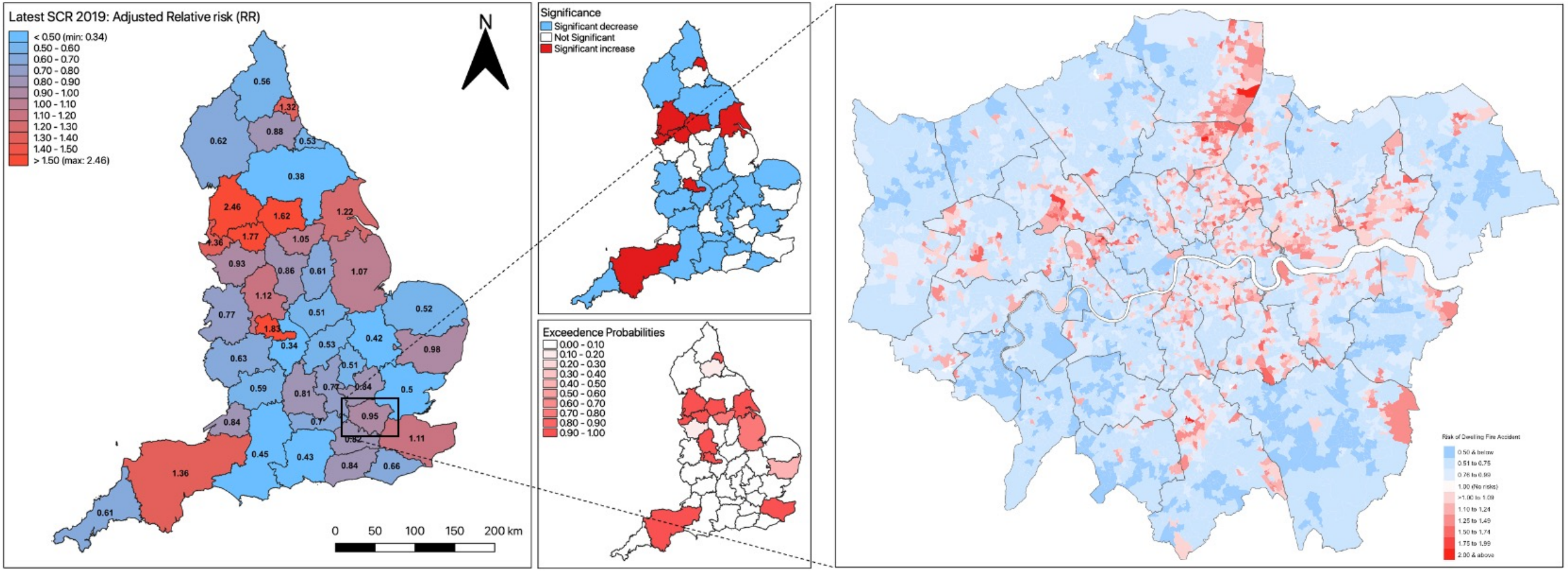


Downloadable data on georeferenced survey records collected primarily for mapping the burden of intestinal parasitic worm infections across the Global south.

Website: http://www.thiswormyworld.org

Examples of open secondary data for health research in Global South

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

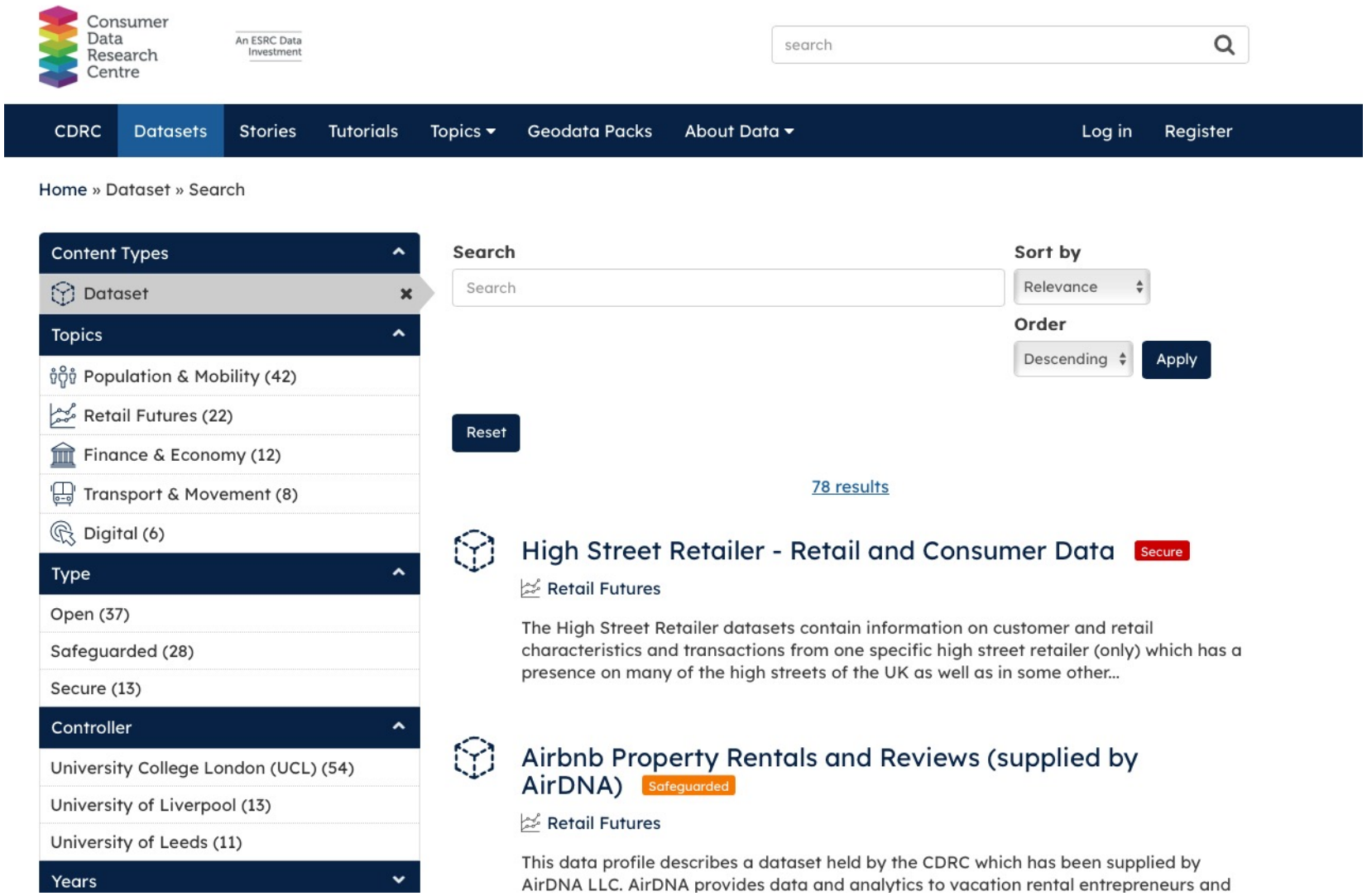


We used **external secondary data** from official statistics pertained to fire hazards and casualties for England and combined with **internal secondary data** from UCL's CDRC registry to account for socioeconomic deprivation

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]

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



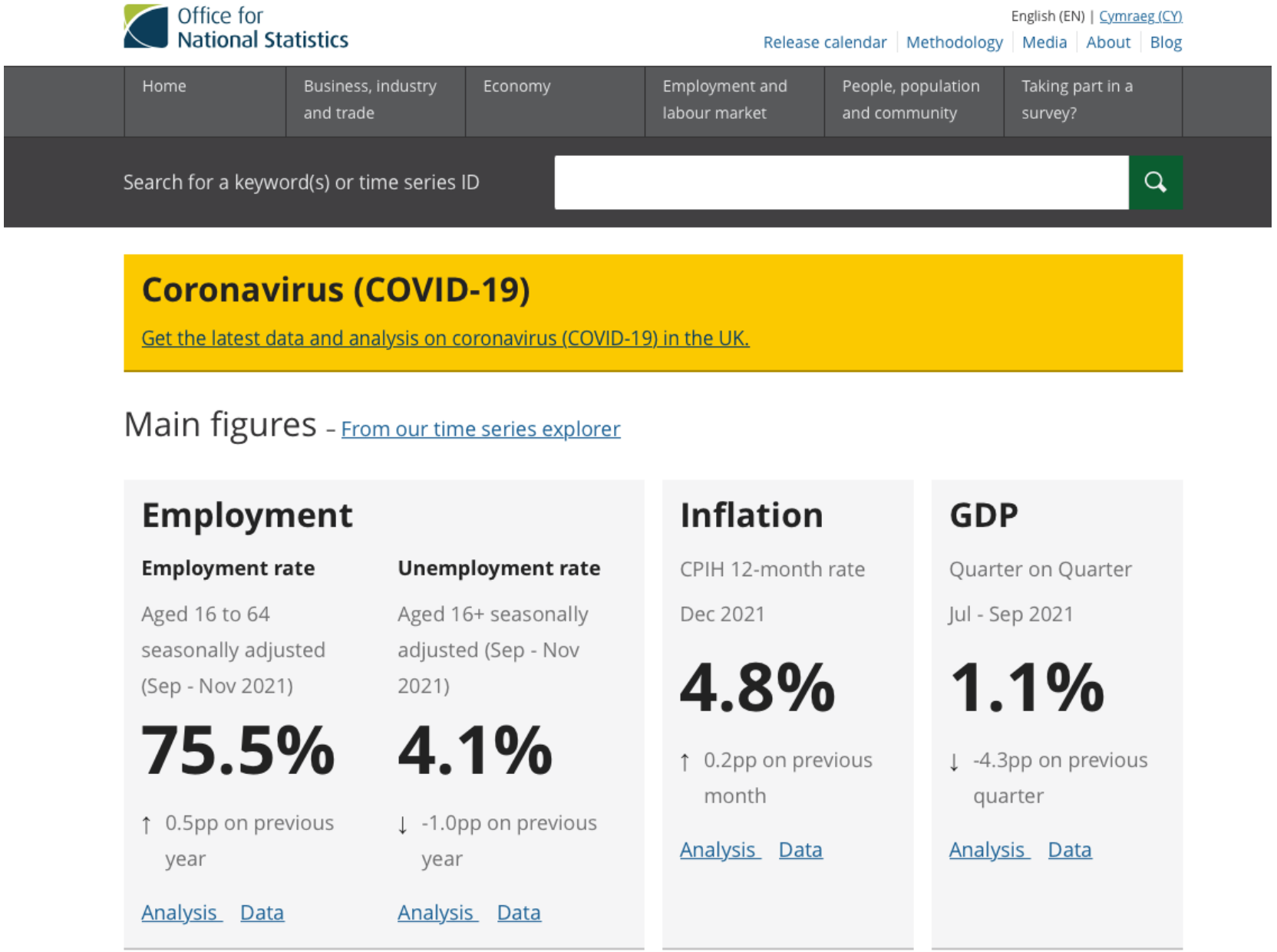
CDRC provides data for research to address societal and economic challenges in the UK.

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

GOV.UK

Ordnance Survey

LONDON DATASTORE

DATA.POLICE.UK

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 along as you do whatever in your power to minimise all kinds of systematic errors in the data collection process

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 DEFINITELY 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 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 full understanding of the 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) • Cross-sectional	• Unit of observation are at group- or aggregated level (e.g., geographic unit) • Data is collected on a single snap in time, and its useful for generating further hypothesis
Observational	• Case-control • Cohort (or longitudinal)	• Observe the effect of an independent variable(s) on an outcome that has already occurred. The time frame is a retrospective analysis • 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)

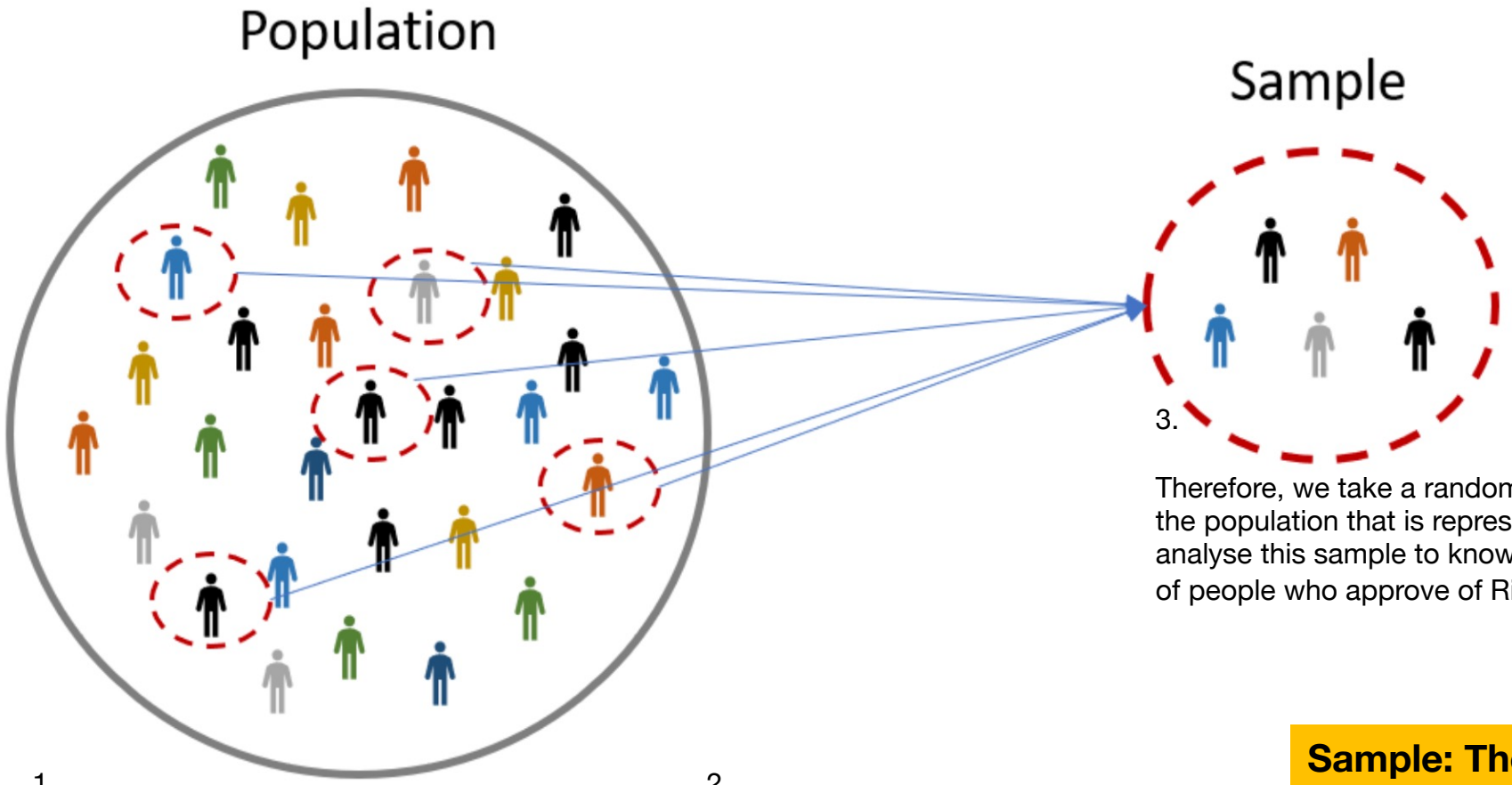
Pilot study is a special cases as it could be anyone of these study designs

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 Rishi Sunak example in Week 1?

Parameter: The proportion of people in the UK population who approve of Rishi Sunak (p)
Statistic: The sample proportion of people from the UK population who approve of Rishi Sunak ($\hat{p}$)



1.

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

2.

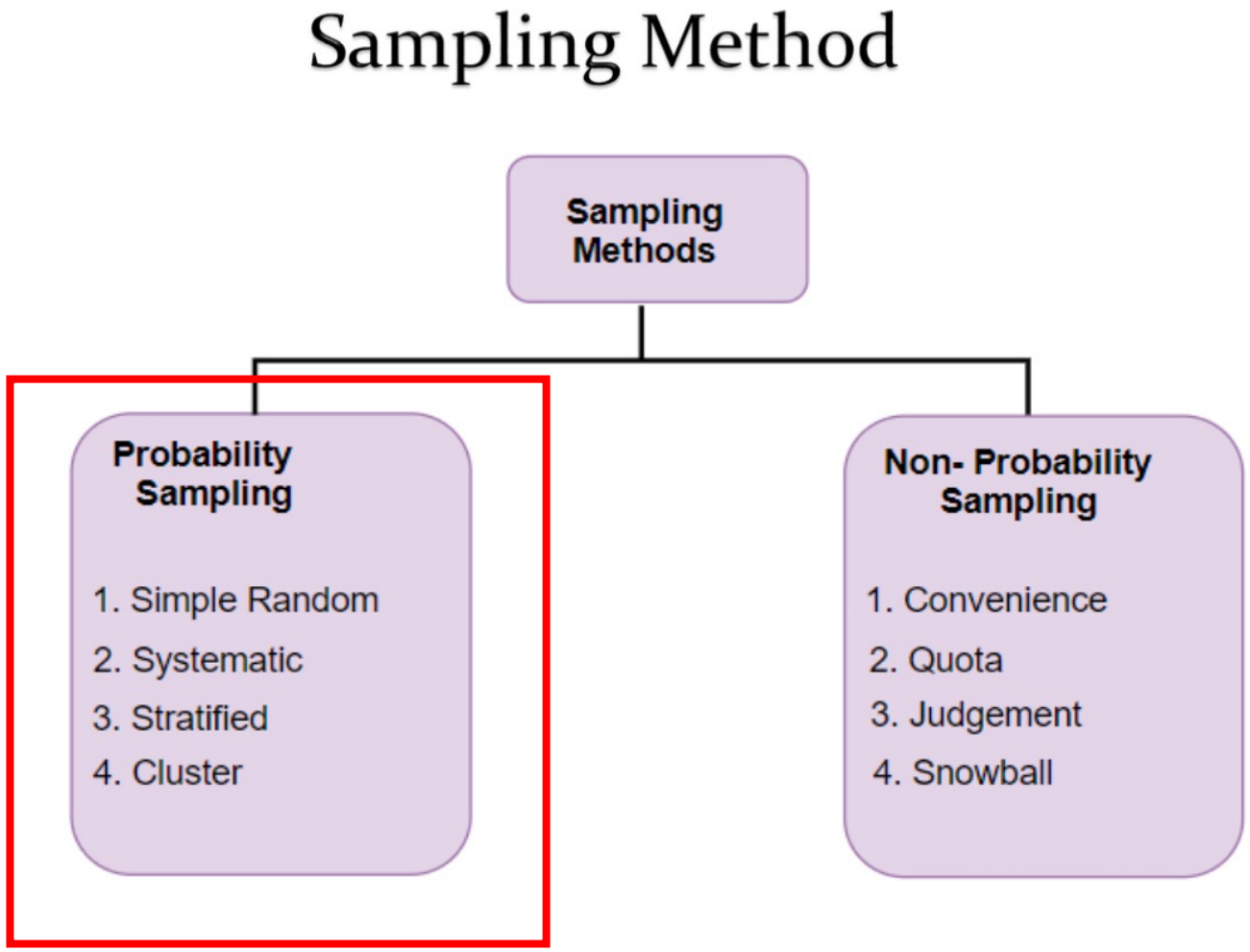
It is impossible, or it very tedious for us to collect data on everyone in the UK population

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 Rishi Sunak

Sample: The subset of subjects chosen for study from a population through data collection

Sampling Strategy [2]



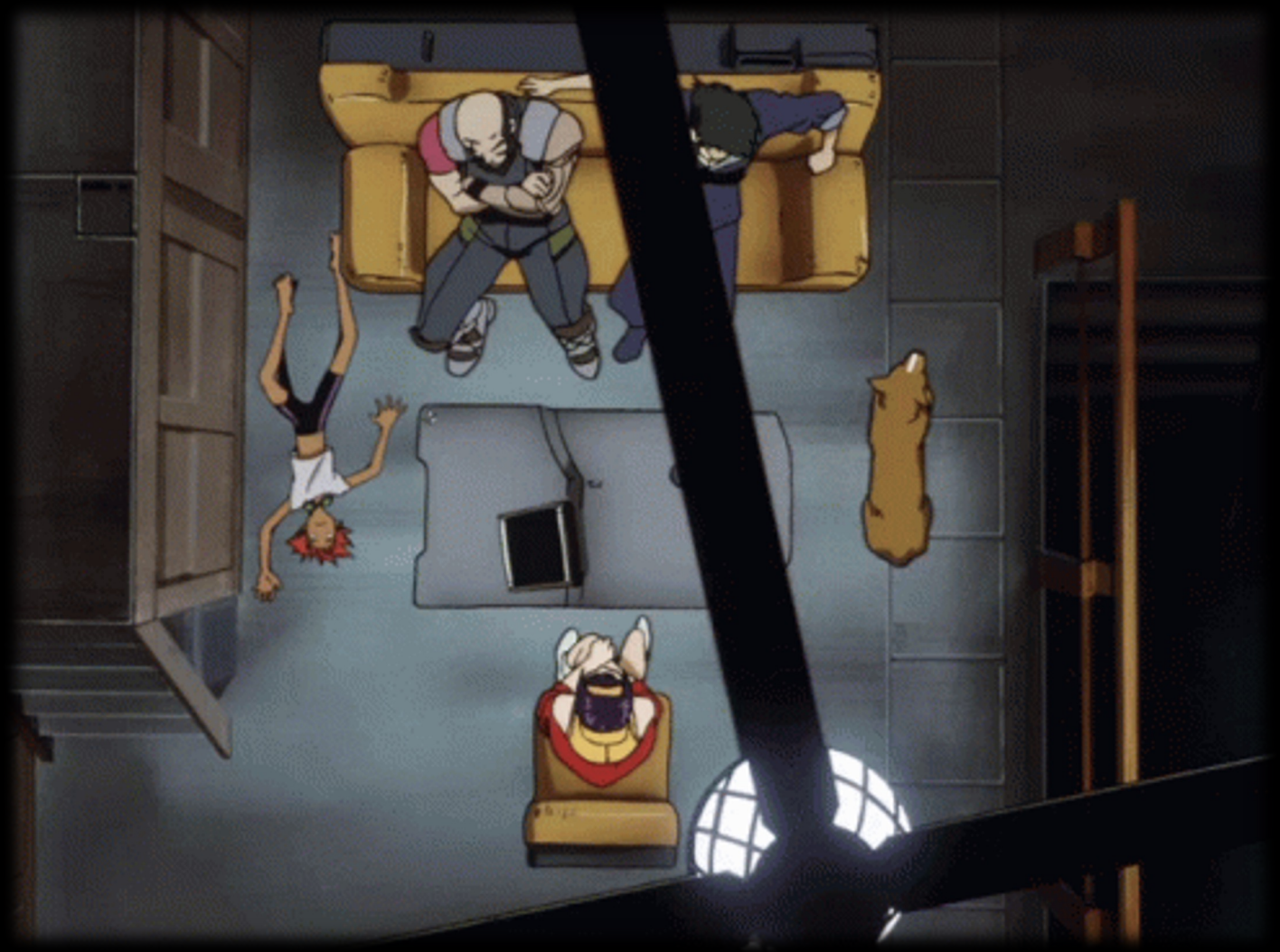
Notes:

- 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.

原作　矢立肇

Chillout time



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 –** 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)

Raw dataset

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

Show what's available

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%)

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). 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

Description of Data [4]

[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). 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]:

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

New 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.

Any questions?

Final words…

The nod of approval… you're ready for inferential statistics!

You’re gonna carry that weight…