Administrative and Open Data

Lecture Week 4

Zachary Dickson and Dan de Kadt

London School of Economics

GV330

Overview

1. Administrative data
2. “Open Data”

What do we mean by...?

- Administrative data: Created when people interact with government or public services

What do we mean by...?

- Administrative data: Created when people interact with government or public services
  - E.g., tax records, vital record (births and deaths)

What do we mean by...?

- Administrative data: Created when people interact with government or public services
  - E.g., tax records, vital record (births and deaths)
- "Open data": Published by governments on online portals; publicly available and accessible
  - Urban data: open data about cities (e.g., 311 data)
  - Open administrative data

What do we mean by...?

- Administrative data: Created when people interact with government or public services
    - E.g., tax records, vital record (births and deaths)
- "Open data": Published by governments on online portals; publicly available and accessible
    - Urban data: open data about cities (e.g., 311 data)
    - Open administrative data
- Can be about individuals, businesses, government agencies, etc.

What do we mean by...?

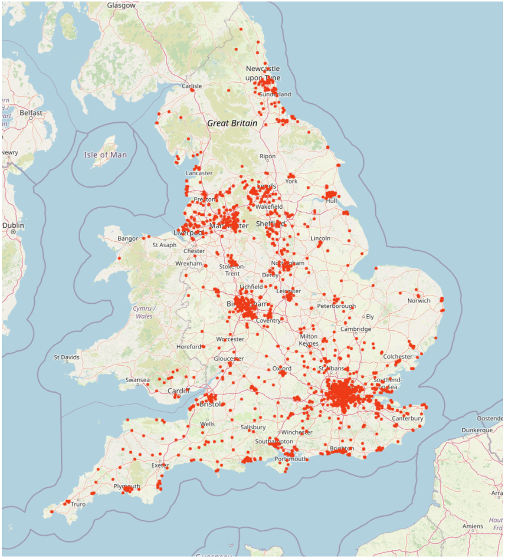
- Administrative data: Created when people interact with government or public services
  - E.g., tax records, vital record (births and deaths)
- "Open data": Published by governments on online portals; publicly available and accessible
  - Urban data: open data about cities (e.g., 311 data)
  - Open administrative data
- Can be about individuals, businesses, government agencies, etc.
- Can sometimes be linked to other data

Some examples of administrative data

- Voter files
- Tax records
- Health records
- Criminal justice records
- Property transactions

Figure: UK Property Price Inequality (2022)

Figure: GP Practice Closures in the UK (2013-2023)



Overview

1. Administrative data

2. "Open Data"

Administrative data example: “voter files”

Administrative data example: “voter files”

- Social scientists increasingly rely on “voter files” where available
- Lists of registered voters indicate whether or not someone voted in a given election (not whom they voted for)

Administrative data example: “voter files”

- Social scientists increasingly rely on “voter files” where available
- Lists of registered voters indicate whether or not someone voted in a given election (not whom they voted for)
- Typically include voter name, address, DOB, gender
    - May also include party affiliation (US); can be supplemented with other data (e.g., commercial data, imputed race/ethnicity)

Administrative data example: "voter files"

- Social scientists increasingly rely on "voter files" where available
- Lists of registered voters indicate whether or not someone voted in a given election (not whom they voted for)
- Typically include voter name, address, DOB, gender
  - May also include party affiliation (US); can be supplemented with other data (e.g., commercial data, imputed race/ethnicity)
- Used by campaigns, pollsters, journalists, academic researchers

Administrative data example: “voter files”

- Social scientists increasingly rely on “voter files” where available
- Lists of registered voters indicate whether or not someone voted in a given election (not whom they voted for)
- Typically include voter name, address, DOB, gender
  - May also include party affiliation (US); can be supplemented with other data (e.g., commercial data, imputed race/ethnicity)
- Used by campaigns, pollsters, journalists, academic researchers
- But, data quality varies (e.g., due to highly decentralized election administration)

Administrative data example: “voter files”

- Social scientists increasingly rely on “voter files” where available
- Lists of registered voters indicate whether or not someone voted in a given election (not whom they voted for)
- Typically include voter name, address, DOB, gender
  - May also include party affiliation (US); can be supplemented with other data (e.g., commercial data, imputed race/ethnicity)
- Used by campaigns, pollsters, journalists, academic researchers
- But, data quality varies (e.g., due to highly decentralized election administration)
  - Commercial vendors sell clean, augmented files

Example: validation (Ansolabehere & Hersh 2012)

- Survey respondents regularly misreport their voting history
  - Misremember
  - Lie
    - Social desirability: "People who are under the most pressure to vote are the ones most likely to misrepresent their behavior when they fail to do so." (Bernstein, Chaha, and Montjoy 2001)

Example: validation (Ansolabehere & Hersh 2012)

- Survey respondents regularly misreport their voting history
  - Misremember
  - Lie
    - Social desirability: “People who are under the most pressure to vote are the ones most likely to misrepresent their behavior when they fail to do so.” (Bernstein, Chaha, and Montjoy 2001)
- Ansolabehere & Hersh (2012) conduct 50 state vote validation
  - Link voter file from private vendor to survey data

Example: validation (Ansolabehere & Hersh 2012)

- Survey respondents regularly misreport their voting history
  - Misremember
  - Lie
    - Social desirability: "People who are under the most pressure to vote are the ones most likely to misrepresent their behavior when they fail to do so." (Bernstein, Chaha, and Montjoy 2001)
- Ansolabehere & Hersh (2012) conduct 50 state vote validation
  - Link voter file from private vendor to survey data
    - One approach is fuzzy string matching

Example: validation (Ansolabehere & Hersh 2012)

- Survey respondents regularly misreport their voting history
  - Misremember
  - Lie
    - Social desirability: "People who are under the most pressure to vote are the ones most likely to misrepresent their behavior when they fail to do so." (Bernstein, Chaha, and Montjoy 2001)
- Ansolabehere & Hersh (2012) conduct 50 state vote validation
  - Link voter file from private vendor to survey data
    - One approach is fuzzy string matching
    - in `R`, see packages `stringdist`, `tidystringdist`, `fuzzyjoin`, `inexact`, `refinr`, `fuzzywuzzyR`

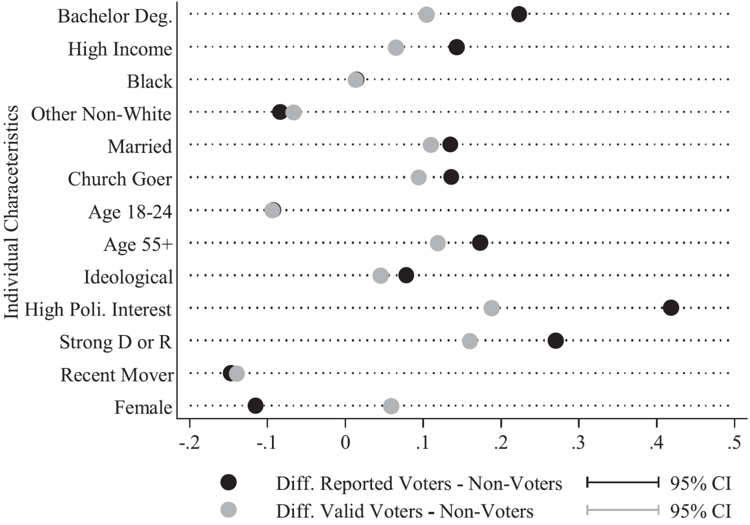
Example: validation (Ansolabehere & Hersh 2012)

- Survey respondents regularly misreport their voting history
  - Misremember
  - Lie
    - Social desirability: "People who are under the most pressure to vote are the ones most likely to misrepresent their behavior when they fail to do so." (Bernstein, Chaha, and Montjoy 2001)
- Ansolabehere & Hersh (2012) conduct 50 state vote validation
  - Link voter file from private vendor to survey data
    - One approach is fuzzy string matching
    - in `R`, see packages `stringdist`, `tidystringdist`, `fuzzyjoin`, `inexact`, `refinr`, `fuzzywuzzyR`
  - Compare self-reports (on survey) to actual voting (in voter file)

Ansolabehere, Stephen, and Eitan Hersh. "Validation: What big data reveal about survey misreporting and the real electorate."

Political Analysis 20, no. 4 (2012): 437-459.

Correlates of reported & validated turnout (Ansolabehere & Hersh 2012)



Side note on race / ethnicity imputation

Predicting individual race from voter registration records (Imai & Khanna 2016)

Side note on race / ethnicity imputation

Predicting individual race from voter registration records (Imai & Khanna 2016)

- Uses Bayes' Rule to estimate $Pr(R_i = r | S_i = s, G_i = g)$, or the conditional probability that voter i belongs to racial group r given his/her surname s and geolocation g.
  - Requires $Pr(R_i = r | S_i = s)$, the racial composition of frequently occurring surnames, $Pr(R_i = r | G_i = g)$, the racial composition of each geolocation (e.g., Census blocks and voting precincts), and $Pr(G_i = g)$, the population proportion of each geolocation.

$\longrightarrow$ Gives a probabilistic prediction of individual race / ethnicity.

Side note on race / ethnicity imputation

Predicting individual race from voter registration records (Imai & Khanna 2016)

- Uses Bayes' Rule to estimate $Pr(R_i = r | S_i = s, G_i = g)$, or the conditional probability that voter i belongs to racial group r given his/her surname s and geolocation g.
  - Requires $Pr(R_i = r | S_i = s)$, the racial composition of frequently occurring surnames, $Pr(R_i = r | G_i = g)$, the racial composition of each geolocation (e.g., Census blocks and voting precincts), and $Pr(G_i = g)$, the population proportion of each geolocation.

  $\longrightarrow$ Gives a probabilistic prediction of individual race / ethnicity.

- `R` package, `wru: Who Are You? Bayesian Prediction of Racial Category Using Surname and Geolocation`

Side note on race / ethnicity imputation

Predicting individual race from voter registration records (Imai & Khanna 2016)

- Uses Bayes' Rule to estimate $Pr(R_i = r | S_i = s, G_i = g)$, or the conditional probability that voter i belongs to racial group r given his/her surname s and geolocation g.
  - Requires $Pr(R_i = r | S_i = s)$, the racial composition of frequently occurring surnames, $Pr(R_i = r | G_i = g)$, the racial composition of each geolocation (e.g., Census blocks and voting precincts), and $Pr(G_i = g)$, the population proportion of each geolocation.

  $\longrightarrow$ Gives a probabilistic prediction of individual race / ethnicity.

- `R` package, `wru: Who Are You? Bayesian Prediction of Racial Category Using Surname and Geolocation`

Imai, K., & Khanna, K. (2016). Improving ecological inference by predicting individual ethnicity from voter registration records. *Political Analysis*, 24(2), 263-272.

Partisan segregation (Brown & Enos 2021)

Partisan segregation (Brown & Enos 2021)

- Partisan geographic clustering prevalent in many countries, with consequences for representation and polarization

Partisan segregation (Brown & Enos 2021)

- Partisan geographic clustering prevalent in many countries, with consequences for representation and polarization
- Use individual-level administrative data to measure partisan residential segregation (isolation or exposure)

Partisan segregation (Brown & Enos 2021)

- Partisan geographic clustering prevalent in many countries, with consequences for representation and polarization
- Use individual-level administrative data to measure partisan residential segregation (isolation or exposure)
- **Data**: exact residential address of every registered voter (more than 180 million individuals) in the U.S., and their partisanship

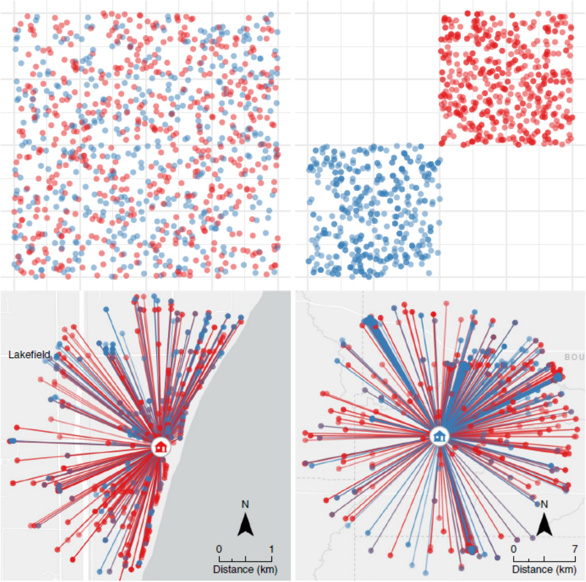
Partisan segregation (Brown & Enos 2021)

- Partisan geographic clustering prevalent in many countries, with consequences for representation and polarization
- Use individual-level administrative data to measure partisan residential segregation (isolation or exposure)
- **Data**: exact residential address of every registered voter (more than 180 million individuals) in the U.S., and their partisanship
- **Approach**: For each individual, measure distance to their k = 1,000 nearest neighbours; create a weighted (by inverse distance) average of exposure to out-partisans

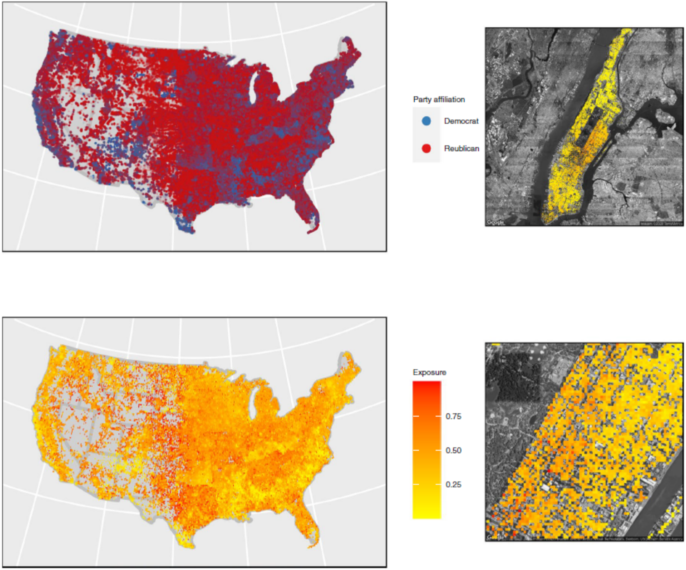
Partisan segregation (Brown & Enos 2021)

- Partisan geographic clustering prevalent in many countries, with consequences for representation and polarization
- Use individual-level administrative data to measure partisan residential segregation (isolation or exposure)
- **Data**: exact residential address of every registered voter (more than 180 million individuals) in the U.S., and their partisanship
- **Approach**: For each individual, measure distance to their k = 1,000 nearest neighbours; create a weighted (by inverse distance) average of exposure to out-partisans
- **Finding**: Partisans are extremely isolated from one another; this persists within cities and neighborhoods and is not explained by racial/ethnic segregation

Spatial & aspatial segregation (Brown & Enos 2021)



Measuring spatial exposure (Brown & Enos 2021)



Overview

UK Open Data: `https://data.gov.uk/`

Find open data

Find data published by central government, local authorities and public bodies to help you build products and services

Search data.gov.uk

Business and economy
Small businesses, industry, imports, exports and trade

Crime and justice
Courts, police, prison, offenders, borders and immigration

Defence
Armed forces, health and safety, search and rescue

Education
Students, training, qualifications and the National Curriculum

Environment
Weather, flooding, rivers, air quality, geology and agriculture

Government
Staff numbers and pay, local councillors and department business plans

Government spending
Includes all payments by government departments over £25,000

Health
Includes smoking, drugs, alcohol, medicine performance and hospitals

Mapping
Addresses, boundaries, land ownership, aerial photographs, seabed and land terrain

Society
Employment, benefits, household finances, poverty and population

Towns and cities
Includes housing, urban planning, leisure, waste and energy, consumption

Transport
Airports, roads, freight, electric vehicles, parking, buses and footpaths

Digital service performance
Cost, usage, completion rate, digital take-up, satisfaction

Government reference data
Trusted data that is referenced and shared across government departments

Benefits & limitations of open data

Benefits & limitations of open data

Benefits

- Transparency

Benefits & limitations of open data

Benefits

- Transparency
- Public service improvement

Benefits & limitations of open data

Benefits

- Transparency
- Public service improvement
- Innovation and economic value

Benefits & limitations of open data

Benefits

- Transparency
- Public service improvement
- Innovation and economic value
- Efficiency

Benefits & limitations of open data

Benefits

- Transparency
- Public service improvement
- Innovation and economic value
- Efficiency

Limitations

- Privacy

Benefits & limitations of open data

Benefits

- Transparency
- Public service improvement
- Innovation and economic value
- Efficiency

Limitations

- Privacy
- Completeness

Benefits & limitations of open data

Benefits

- Transparency
- Public service improvement
- Innovation and economic value
- Efficiency

Limitations

- Privacy
- Completeness
- Accessibility

Benefits & limitations of open data

Benefits

- Transparency
- Public service improvement
- Innovation and economic value
- Efficiency

Limitations

- Privacy
- Completeness
- Accessibility
- Accuracy

Non-democracies & data (Carlitz & McLellan 2021)

- Authoritarian regimes manipulate official data to legitimize their authority
- Systematically missing or biased data may jeopardize research integrity and lead to false inferences

Non-democracies & data (Carlitz & McLellan 2021)

- Authoritarian regimes manipulate official data to legitimize their authority
- Systematically missing or biased data may jeopardize research integrity and lead to false inferences
- Official statistics may...
  - Exaggerate development progress (Sandefur & Glassman 2015)
  - Reflect politicized population counts (Akinyoade, Appiah and Asa 2017; Elemo 2018)
  - Reflect politicization of macroeconomic indicators (Martinez 2019; Rawski 2001; Tsai 2008; Wallace 2016)

Non-democracies & data (Carlitz & McLellan 2021)

- Authoritarian regimes manipulate official data to legitimize their authority
- Systematically missing or biased data may jeopardize research integrity and lead to false inferences
- Official statistics may...
    - Exaggerate development progress (Sandefur & Glassman 2015)
    - Reflect politicized population counts (Akinyoade, Appiah and Asa 2017; Elemo 2018)
    - Reflect politicization of macroeconomic indicators (Martinez 2019; Rawski 2001; Tsai 2008; Wallace 2016)
- Why release data?

Non-democracies & data (Carlitz & McLellan 2021)

- Authoritarian regimes manipulate official data to legitimize their authority
- Systematically missing or biased data may jeopardize research integrity and lead to false inferences
- Official statistics may...
  - Exaggerate development progress (Sandefur & Glassman 2015)
  - Reflect politicized population counts (Akinyoade, Appiah and Asa 2017; Elemo 2018)
  - Reflect politicization of macroeconomic indicators (Martinez 2019; Rawski 2001; Tsai 2008; Wallace 2016)
- Why release data?
  - Rewarded by voters (Maerz 2016; Little 2017)
  - Ensure future access to government information (Berliner 2014)
  - Discredit opposition parties (Carlitz & McLellan 2021)

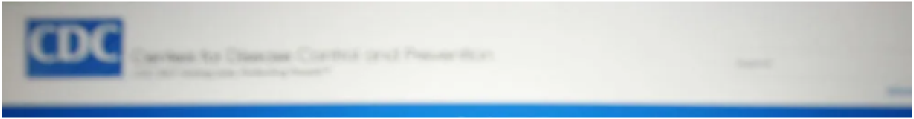
TIME

SUBSCRIBE

POLITICS • DONALD TRUMP

Federal Webpages Go Dark as Trump Administration Removes Public Data

2 MINUTE READ



Example: “open data” on racially biased policing

Example: “open data” on racially biased policing

MPs rebuke police for ‘systemic failure’ to improve record on race

Failings have led to ‘unjustified inequalities’, says landmark report that finds little progress in 22 years since Macpherson

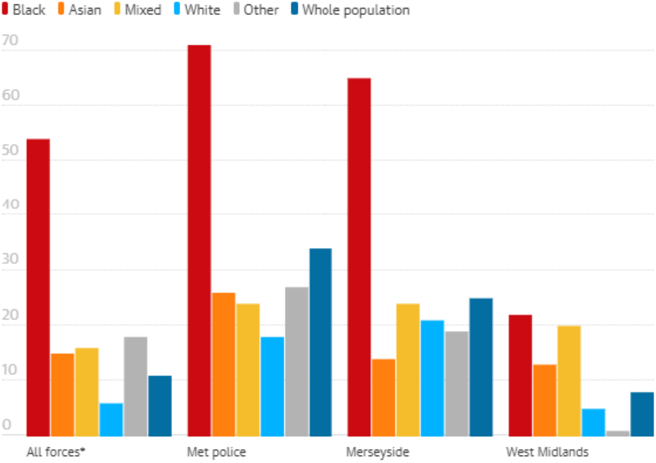
- **Analysis: failure at the top of police and of governments Tory and Labour**

Black people are nine times more likely than white people to face stop and search, with most

Example: “open data” on racially biased policing

Police carried out 54 stop and searches for every 1,000 black people in England and Wales in 2019-20

Stop and search per 1,000 people



Guardian graphic | Source: The Macpherson report. Note: data unavailable for Greater Manchester, City of London and British Transport Police. * Includes BTP but excludes Greater Manchester. Selected forces shown

Metropolitan Police Dashboards (London Open Data)

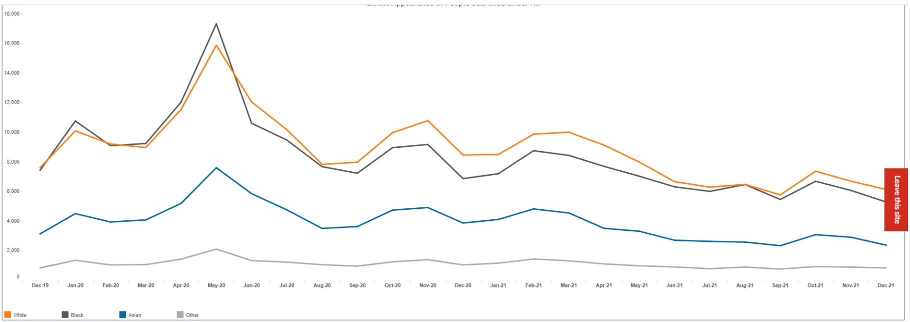
Crime, Stop & Search and Taser data

Dashboards below or full data and accessible versions are available to download from the London datastore.

- Business crime dashboard
- Crime data dashboard
- Crime data dashboard - previous crime categories
- Hate crime or special crime dashboard
- Stop and search dashboard
- Taser data dashboard – historic (not updated)
- Use of force dashboard

https://www.met.police.uk/sd/stats-and-data/

Stop & search by ethnicity (London Open Data)



EA	Dec 2019		Jan 2020		Feb 2020		Mar 2020		Apr 2020		May 2020	
	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.
White	7,583	1.6	10,065	2.1	9,180	1.9	8,946	1.8	11,523	2.4	15,851	3.2
Black	7,407	5.8	10,734	8.4	9,070	7.1	9,216	7.2	12,006	9.4	17,300	13.6
Asian	3,113	2.1	4,479	3	3,914	2.6	4,056	2.7	5,167	3.5	7,582	5.1
Other	813	1.6	1,324	2.5	1,014	1.9	1,032	2	1,407	2.7	2,082	4
All	18,916	2.3	26,602	3.3	23,178	2.8	23,250	2.8	30,103	3.7	42,815	5.2

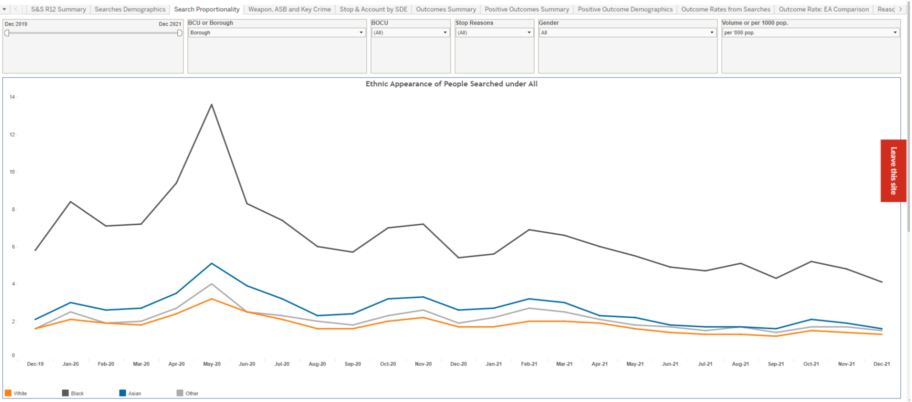
EA	Jun 2020		Jul 2020		Aug 2020		Sep 2020		Oct 2020		Nov 2020	
	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.
White	12,024	2.5	10,151	2.1	7,806	1.6	7,953	1.6	9,947	2	10,761	2.2
Black	10,584	8.3	9,448	7.4	7,651	6	7,206	5.7	8,937	7	9,149	7.2
Asian	5,828	3.9	4,733	3.2	3,479	2.3	3,605	2.4	4,717	3.2	4,889	3.3
Other	1,309	2.5	1,202	2.3	1,027	2	932	1.8	1,224	2.3	1,368	2.6
All	29,745	3.6	25,534	3.1	19,963	2.4	19,696	2.4	24,825	3	26,167	3.2

EA	Dec 2020		Jan 2021		Feb 2021		Mar 2021		Apr 2021	
	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.	Search Volume	Search per '000 pop.
White	8,435	1.7	8,476	1.7	9,846	2	9,978	2	9,102	1.9
Black	6,841	5.4	7,172	5.6	8,729	6.9	8,411	6.6	7,671	6
Asian	3,845	2.6	4,084	2.7	4,795	3.2	4,526	3	3,485	2.3
Other	1,013	1.9	1,132	2.2	1,414	2.7	1,293	2.5	1,081	2.1
All	20,134	2.5	20,864	2.6	24,784	3	24,208	3	21,339	2.6

Ethnic Appearance of People Searched, total volume.

Source: `https://www.met.police.uk/sd/stats-and-data/met/stop-and-search-dashboard/`

Stop & search by ethnicity (London Open Data)



Ethnic Appearance of People Searched, per 1,000 population.

Source: `https://www.met.police.uk/sd/stats-and-data/met/stop-and-search-dashboard/`

Stop & Search in the UK (Vomfell & Stewart 2021)

- The majority of officers over-search Asian and Black people relative to the ethnic composition of crime suspects and of the areas they patrol.
- Due to both over-patrolling (targeting areas based on ethnic composition) and over-searching (targeting individuals based on ethnic appearance)

Vomfell, L., Stewart, N. Officer bias, over-patrolling and ethnic disparities in stop and search. *Nature Human Behaviour* 5, 566–575 (2021).