# Administrative and big data – opportunities and challenges for the 2021 Census and beyond

Andy Teague
Jane Naylor
Office for National Statistics

# Overview

- Context and background

- Aspirations, expectations and challenges in using admin and big data

- Research outputs using admin data

- Case studies using alternative data

- Further information

# Beyond 2011 Programme - 2011-2014/15

- Programme to identify the best way to provide small area population and socio-demographic statistics in future (England and Wales*)

- ONS consulted on two options in autumn 2013

  - Census once a decade
  - Census based on administrative data and large annual surveys

- Made a recommendation in March 2014

* *Three Census taking authorities in UK – ONS (England and Wales), National Registers for Scotland, Northern Ireland Statistics and Research Agency – all work closely together*

# National Statistician's recommendation (March 2014)

- An **online census** of all households and communal establishments in 2021 - with support for those who are unable to complete the census online.

  **AND**

- Increased use of **administrative data and surveys** in order to enhance the statistics from the 2021 census and improve annual statistics between censuses.

  - ➢ Make the **best use of all available data** to provide the population statistics required.

    AND

  - ➢ Offer a **springboard to** the greater use of administrative data and annual surveys in **the future.**

# Way forward agreed with Government

- "The Government welcomes the recommendation for a **predominantly online census** in 2021 supplemented by further use of administrative and survey data.

- **Government recognises the value of the census** and its history as a bedrock of statistical infrastructure. The census provides information on the population that is of **fundamental importance to society**....

-  Our ambition is that **censuses after 2021** will be conducted **using other sources of data** and providing more timely statistical information .... dependent on the dual running sufficiently validating the perceived feasibility of that approach."

*Minister for the Cabinet Office, July 2014*

# Census Transformation Programme Key Strands

1.  **2021 online Census operation**

    - Developing and implementing 2021 online Census

    - Maximise online response (target 75%), minimise digital exclusion

    - Enhancing census outputs with admin & survey data – research outputs

2.  **Beyond 2021/Admin Data Census**

    - Acquire new administrative data

    - Develop new methods using admin data and surveys

        - Evaluation against 2021 census outputs

    - Research Outputs from 2015 onwards

        - Population estimates and characteristics from admin data and surveys

# What do we need to switch to an admin data approach post 2021

- Easy/flexible/rapid access to existing and new data sources – data access legislation
- Ability to link data efficiently and accurately
- Methods to produce statistical outputs of sufficient quality that meet user needs
- Acceptable to stakeholders (users, public, suppliers and parliament)

# Admin Data - Aspirations and Expectations

- Lots of potential with admin data but admin (and big) data alone won't be the answer, we will still need surveys

- Aiming to replicate as many census outputs as possible using admin data (and surveys) by 2021 - compare with 2021 Census

- Research outputs published each year, subject to data access and quality, started Autumn 2015

- Publish assessment of progress each year starting Spring 2016

- Building towards a recommendation in 2023 on whether a Census based on admin data and surveys can provide statistics of the required quality

- Would be first country in the world to move to an admin data based approach without a population register.

# Potential of administrative sources - population characteristics

| Individual variables | Earnings and benefits data | Health data | Education data | Vehicle and driver data | Property attributes |
|---|---|---|---|---|---|
| Sex | G | G | G | G | R |
| Date of birth | G | G | G | G | R |
| Marital Status | A | R | A | R | R |
| Household relationships | A | R | R | R | R |
| National identity | R | R | G | R | R |
| Ethnicity | A | A | G | R | R |
| Language (ability to speak English or Welsh) | RA | RA | A | R | R |
| Religion | R | R | G | R | R |
| Qualifications | R | R | AG | R | R |
| General health | R | A | RA | R | R |
| Disability/long-term health conditions | A | A | A | RA | R |
| Carers (number of) | A | R | R | R | R |
| Economic Activity | A | R | A | R | R |
| Industry of occupation | A | R | R | R | R |
| Mode of travel to work | R | R | R | R | R |
| Place of work | A | R | R | R | R |
| Country of birth | RA | RA | R | R | R |
| Internal or international migrant | A | G | A | A | R |
| Term time address | R | R | G | R | R |
| Income | AG | R | R | R | R |

# Potential of administrative sources - housing characteristics

| | Earnings and benefits data | Health data | Education data | Vehicle and driver data | Property attributes |
|---|---|---|---|---|---|
| Size of accommodation (e.g. number of rooms) | R | R | R | R | G |
| Type of dwelling | RA | R | R | R | G |
| Tenure | RA | R | R | R | R |
| Number of cars and vans | R | R | R | AG | R |

# Aims of publishing admin based research outputs

New annual admin based research outputs beginning Autumn 2015 to enable:

- feedback from users on quality

- methods to be improved through time

- operationalise procedures

- users to derive early benefits

Aim to demonstrate improvements each year in:

- breadth (topics) and/or

- (geographical) detail and/or

- accuracy/timeliness

# Publication of new (research) outputs (indicative - subject to data access and quality)

| Output topic | Geographic level | 2015 (done) | 2016 | 2017 | 2018 | ........... | 2021 |
|---|---|---|---|---|---|---|---|
| **Population estimates by age and sex** | National | ✔ | ✔ ✔ | ✔ ✔ | ✔ ✔ | | |
| | LA | ✔ | ✔ | ✔ ✔ | ✔ ✔ | | |
| | MSOA | | ✔ | ✔ | ✔ | | |
| | LSOA | | ✔ | ✔ | ✔ | | |
| | OA | | | ✔ | ✔ | | |
| | Postcode | | | | ✔ | | |
| **Household estimates** | National | | ✔ | ✔ | ✔ ✔ | | |
| | LA | | ✔ | ✔ | ✔ ✔ | | |
| | MSOA | | | ✔ | ✔ | | |
| | LSOA | | | ✔ | ✔ | | |
| | OA | | | | ✔ | | |
| **Measures of income** | National | | ✔ | ✔ | ✔ | | |
| | LA | | ✔ | ✔ | ✔ | | |
| | MSOA | | | | ✔ | | |
| **Ethnicity – combined admin and survey data** | National | | | ✔ | ✔ | | |
| | LA | | | ✔ | ✔ | | |
| | MSOA | | | | ✔ | | |
| **Qualifications (under 30s)** | National | | | ✔ | ✔ | | |
| | LA | | | ✔ | ✔ | | |
| | MSOA | | | | ✔ | | |
| **Industry of employer** | National | | | | ✔ | | |
| | LA | | | | ✔ | | |
| **Housing characteristics – eg number of rooms** | National | | | ✔ | ✔ | | |
| | LA | | | ✔ | ✔ | | |
| **Alternative population bases for example, commuter flows, daytime/night-time populations)** | National | | | | ✔ | | |
| | LA | | | | ✔ | | |

# 2011 Admin data based population counts compared to the 2011 Census
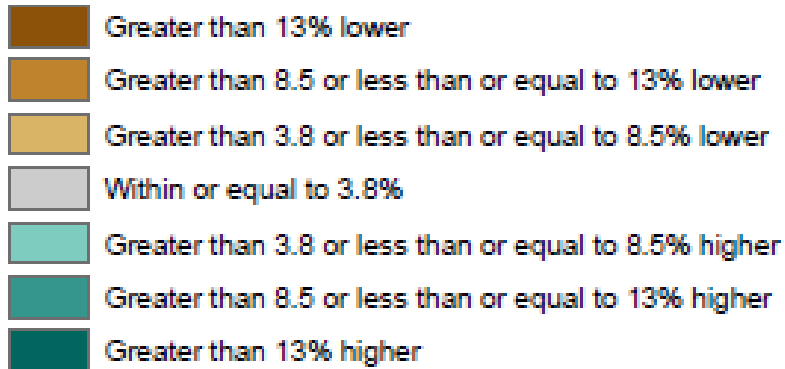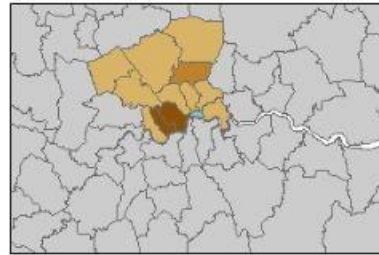


**Admin data method lower than 2011 Census**

**Admin data method higher than 2011 Census**

Greater than 13% lower

Greater than 8.5 or less than or equal to 13% lower

Greater than 3.8 or less than or equal to 8.5% lower

Within or equal to 3.8%

Greater than 3.8 or less than or equal to 8.5% higher

Greater than 8.5 or less than or equal to 13% higher

Greater than 13% higher
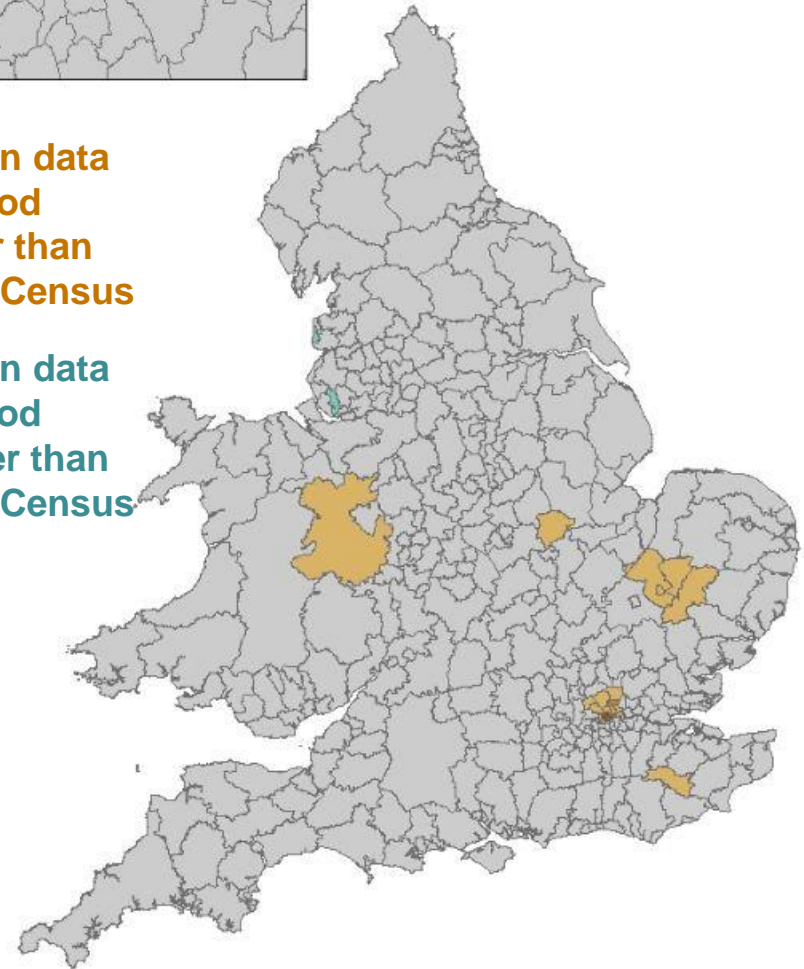
94% of LA total population counts within 3.8% of Census estimate in 2011

# 2014 Admin data based population counts compared to the 2014 official population estimates



Greater than 13% lower

Greater than 8.5 or less than or equal to 13% lower

Greater than 3.8 or less than or equal to 8.5% lower

Within or equal to 3.8%

Greater than 3.8 or less than or equal to 8.5% higher

Greater than 8.5 or less than or equal to 13% higher

Greater than 13% higher
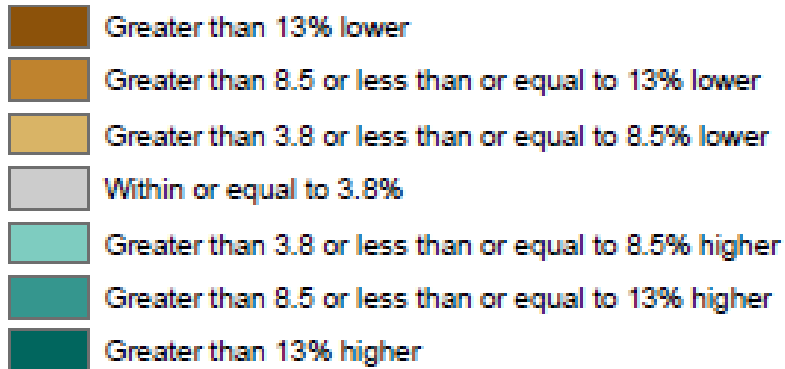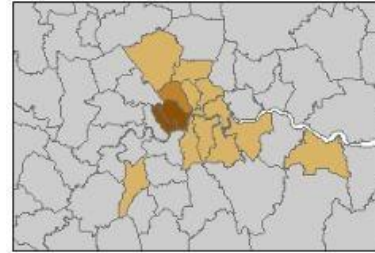
**Admin data method lower than 2014 MYE**

**Admin data method higher than 2014 MYE**
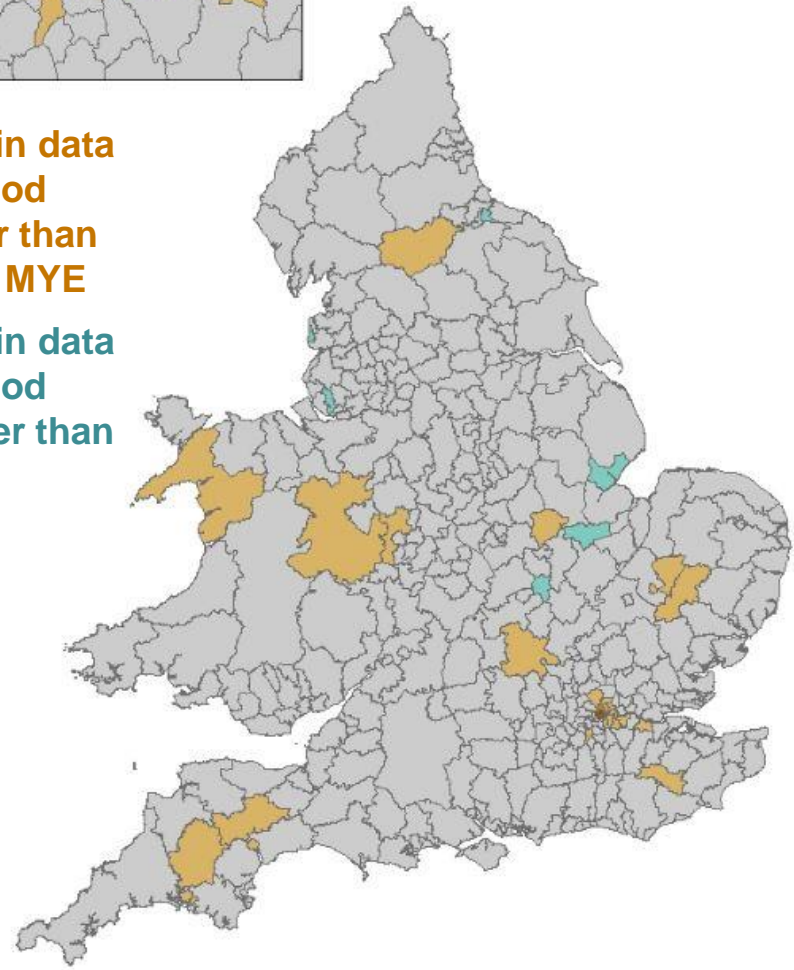
90% of LA total population counts within 3.8% of mid-year estimate in 2014

# Interactive content



Administrative data based estimates (SPD v1.0) compared to official population estimates
2011 to 2014 by Local Authority, England & Wales
Note that these outputs are not official statistics on the population
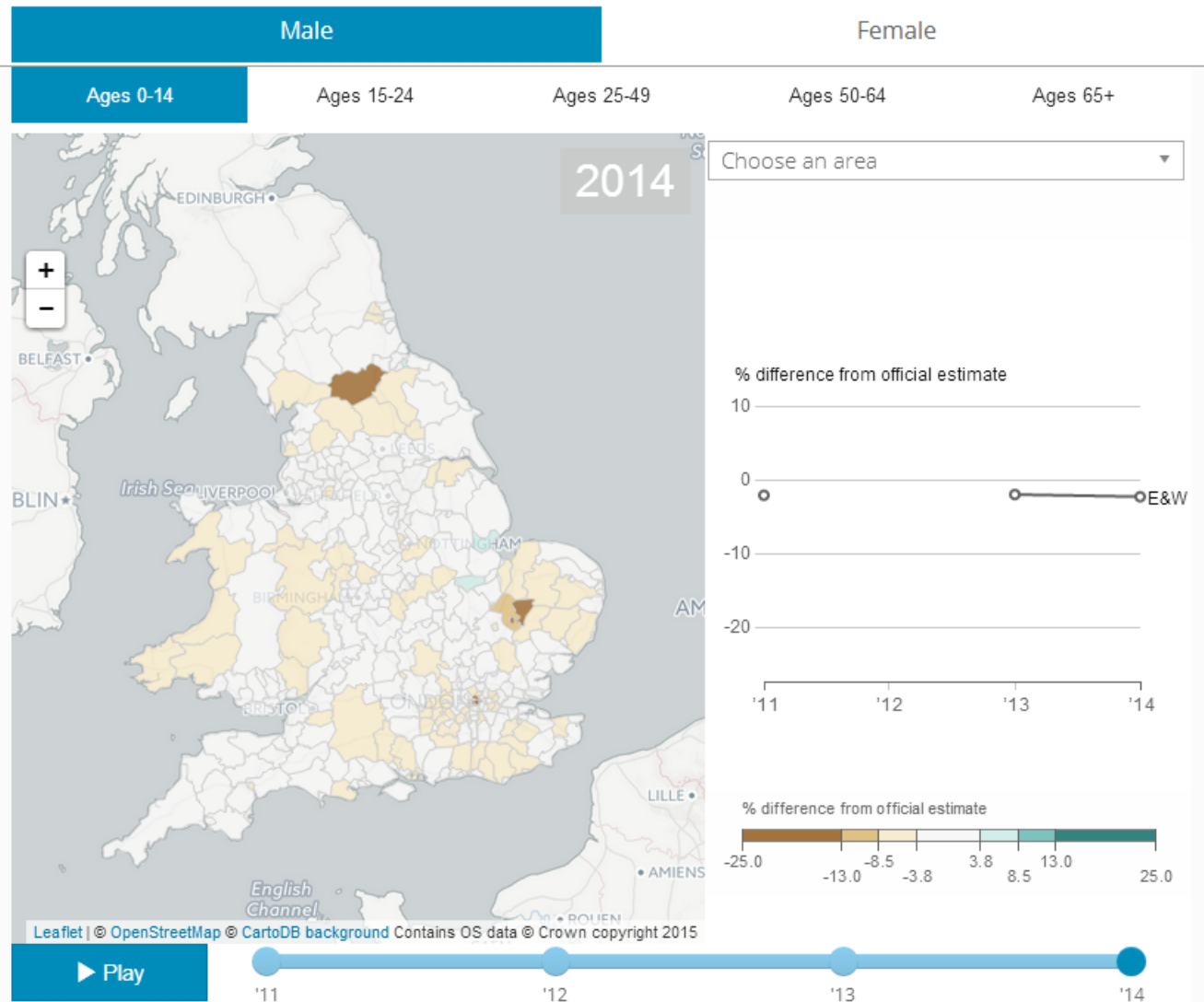
# Big data pilot projects

Demographics: Twitter

Electricity: smart meter

Property: Zoopla Data

Demographics, population flows : mobile phone data

# Smart meters

**Rationale: Potential of data from electricity smart-type meters to model occupancy**

- Support more efficient field operations
- Data from smart meter trials in Great Britain and Republic of Ireland
- A range of potential methods identified
- Privacy and ethics

# Electricity: smart meter

**Half hourly consumption over a week**
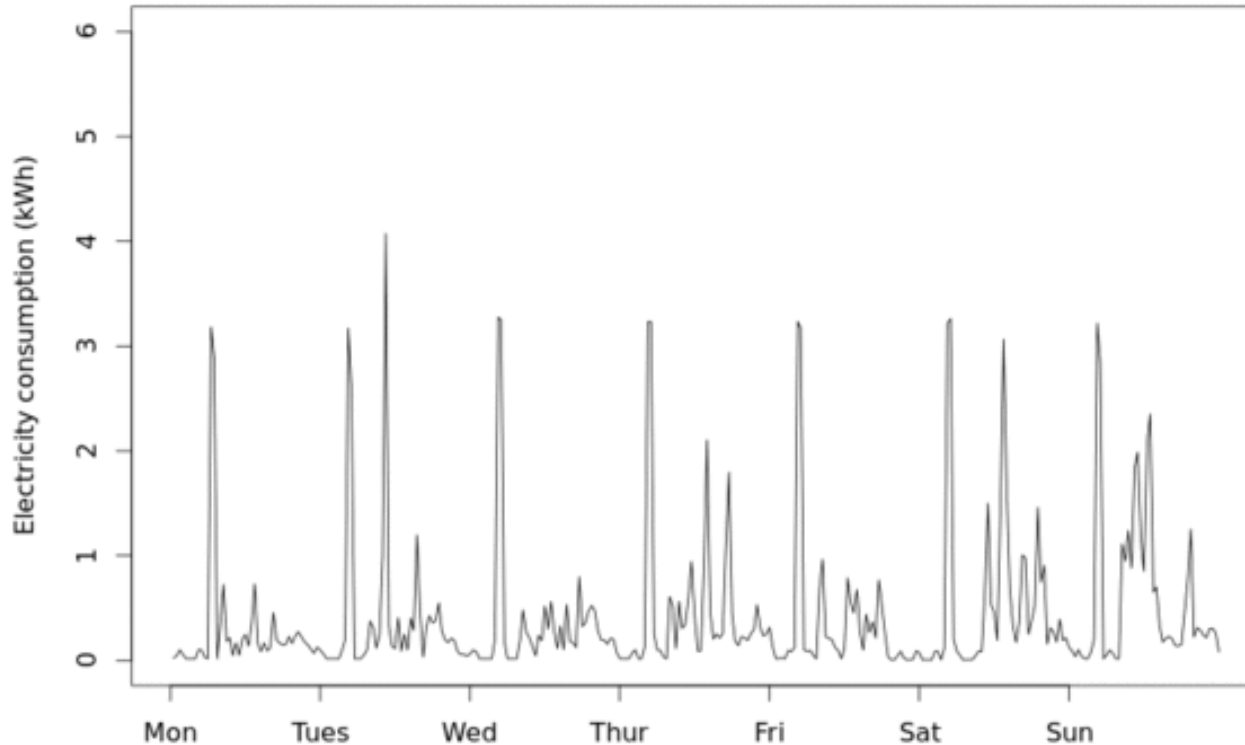


Half hourly electricity consumption over 7 days at one meter, through 28 consecutive 7 day periods.

# Housing websites: Zoopla

# Mobile Phones

**Rational: Modelling population density and population flows, e.g. Commuting statistics**

- Building relationships with mobile network operators and other parts of UK Government
- No data yet
- Privacy and ethics

# Demographics, population  flows : mobile phone data



Call density on 17-04-2015 at 00:00, Milano

<= 338.4/km²

<= 67.1/km²

<= 31.6/km²

<= 13.0/km²

<= 3.7/km²

No calls made

km

0    5    10

Source of the Dataset: TIM Big Data Challenge 2015, www.telecomitalia.com/bigdatachallenge
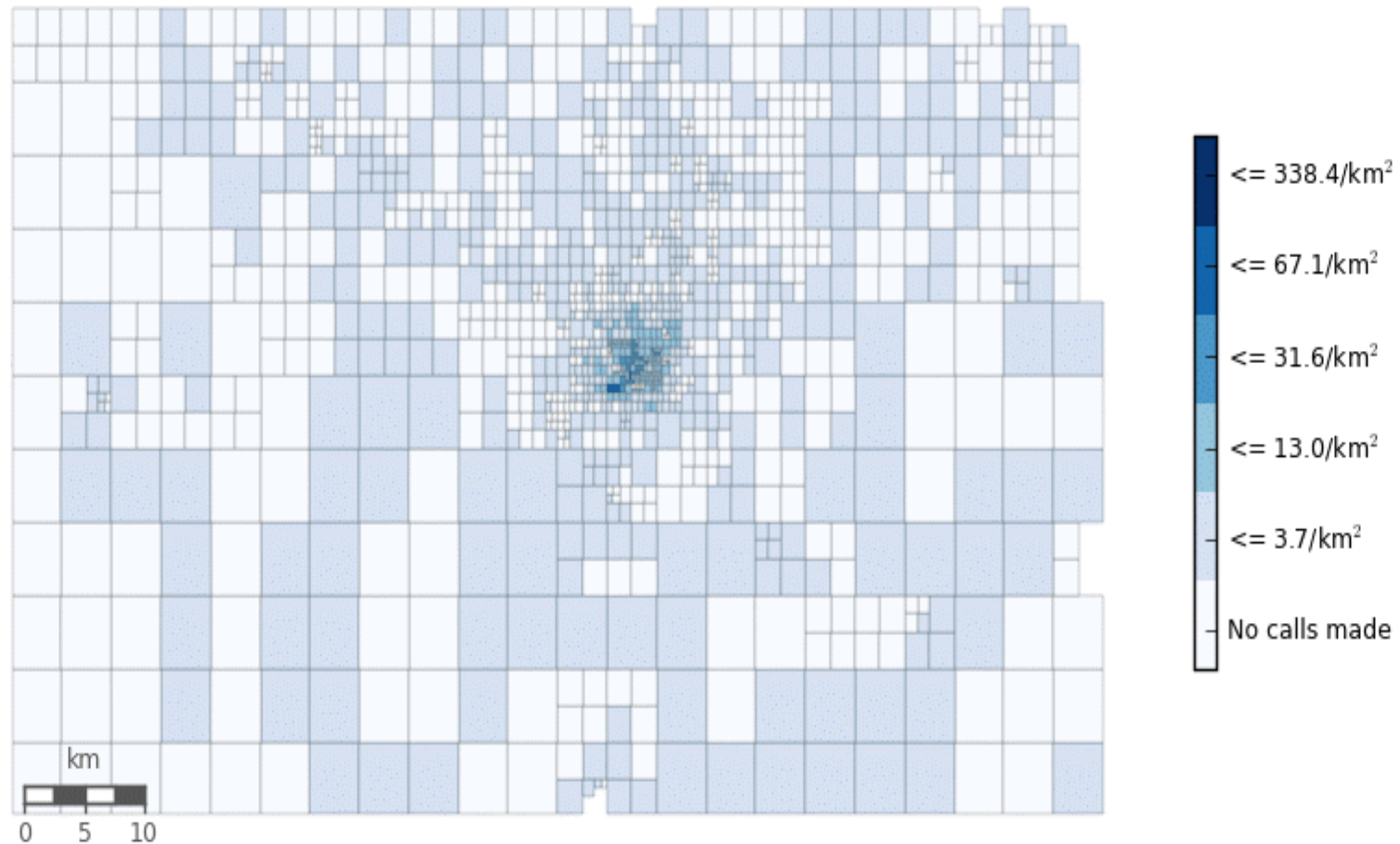
# Twitter

**Rationale: Using geo-located Twitter to gain new insights mobility and migration**

- 7 months of geo-located tweets within Great Britain (about 100 million data points)

- Methodology to infer place of usual residence:

-      Identify user 'anchor points' by clustering tweets using a DBSCAN algorithm

-      Identify residential anchor points using AddressBase and nearest neighbour analysis

**Geolocated penetration rates by local authority**



Legend
- 0.0029 - 0.0045
- 0.0045 - 0.0055
- 0.0055 - 0.0069
- 0.0069 - 0.0126
- 0.0126 - 0.0217

# Use case: Student mobility



* Net flow of Twitter users inferred to be living in a different local authority in the previous month

# Demographics: Twitter Data



Use python to extract demographics for twitter user metadata

Representative sample of UK twitter users

Apply adjustment factors taken from Opinions Survey (ONS)

Office for National Statistics

Extract user information from twitter API

Use R package ReGenesis to calibrate to population

# Opportunities and challenges

## Opportunities

- Reduce respondent burden
- Improve quality – timeliness, frequency, new or complimentary outputs
- More efficient/lower cost – re-use of data

## Challenges

- Statistical
    - Bias
    - Coverage
    - Definitions
    - New methods - combining with surveys
- Data access
- Ethical/privacy

- Can admin/big data _help_ replace the Census? Not proven yet but, in conjunction with surveys…

Population estimates – highly likely

Structure of the population – possibly

Characteristics of the housing stock and population - mixed

# Privacy

- All admin and big data used by CTP is anonymised before use and held in secure research environments

- First iteration of privacy impact assessment for 2021 Census and admin data published

http://www.ons.gov.uk/census/censustransformationprogramme/beyond2011censustransformationprogramme/privacyandconfidentiality

- Safeguarding data policy paper published

http://www.ons.gov.uk/ons/about-ons/who-ons-are/programmes-and-projects/beyond-2011/reports-and-publications/beyond-2011-safeguarding-data-for-research-our-policy--m10-.pdf

- National Statistician's Data Ethics Committee

# Further information

New ONS website https://www.ons.gov.uk/

Census Transformation Programme

https://www.ons.gov.uk/census/censustransformationprogramme

ONS Big Data project

https://www.ons.gov.uk/aboutus/whatwedo/programmesandprojects/theon
sbigdataproject

Data Access Legislation Consultation

https://www.gov.uk/government/consultations/better-use-of-data-in-
government

Email

andy.teague@ons.gsi.gov.uk

jane.naylor@ons.gsi.gov.uk