NCRM National Centre for Research Methods

Engineering geoprivacy using automated zone design

James Robards (NCRM), David Martin (NCRM, ADRC-E), Chris Gale (ADRC-E)

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
UNIVERSITY OF Southampton
MANCHESTER 1824 The University of Manchester
THE UNIVERSITY OF EDINBURGH

Overview: Engineering geoprivacy using automated zone design

- What is zone design and what is the utility of zone design?
- Zone design for new forms of data
- Application of zone design to a “disclosive” synthetic dataset
- Initial results
- Applications to other data / the bigger picture

What is zone design?

- Depending on the purpose, size and position of boundaries may matter in many different ways.
- Geographers know this as the “Modifiable Areal Unit Problem” (Openshaw, 1984).
- Comprises “scale” and “aggregation” problems.
- Divisions of geographical space, usually defined in terms of polygons – often though of as just shaded areas on a map.
- Choice of the number and configuration of zones.

What is zone design?

- Locational information is an important key to disclosure.
- Individual census records are potentially identifiable.
- Standard approach is to aggregate over geographical areas to meet a required level of comfort for one-off release of aggregated data.
- Additional protection from other techniques (e.g. record swapping, collapsing classes, minimum thresholds).
- Examples:
  - 2011 Output Areas
  - 2011 Lower Layer Super Output Areas
  - 2011 Workplace Zones

Zone design for new forms of data

- Growth, and desire to make better use, of administrative data.
- Importance of data linkage because administrative records often domain-specific and attribute poor.
- Power of linkage across domains and to location.
- Widespread investment in access (Administrative Data Research Network, Farr Institute, VML, secure data labs…).
- Pressure and potential: to realise societal benefits and make best use of data investment.

Using Zone Design with linked data

- What is the utility of automated zone design for use with new forms of linked and administrative data?

- Create zones where there is sufficient information on the building blocks for research to proceed, yet the locations of individuals are protected.
- Potentially – quantify the process leading to the zones released to researchers.

Using Zone Design with linked data

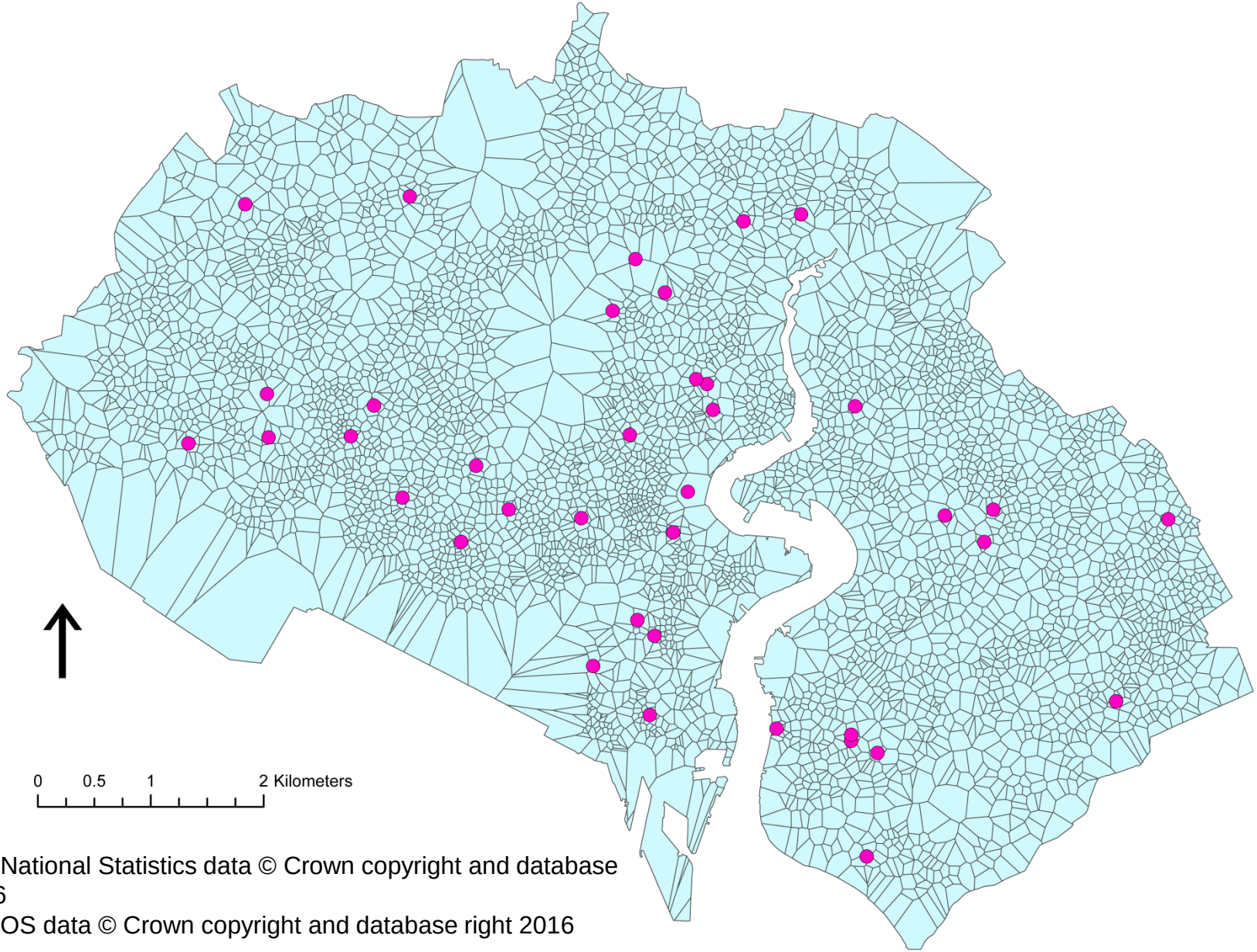
- Middle layer super output area (MSOA) coded individual level synthetic dataset (2011 Census) with 2015 Index of Multiple Deprivation (IMD).
- Postcode Thiessen polygons.
- Health & Social Care Information Centre – location of General Practitioners (GPs) in Southampton.

Using Zone Design with linked data

Geography	N units
Households in Southampton (2011)	98,244
GP locations (November 2015)	45
MSOAs	32
LSOAs	148
OAs	766
Postcodes	5,047

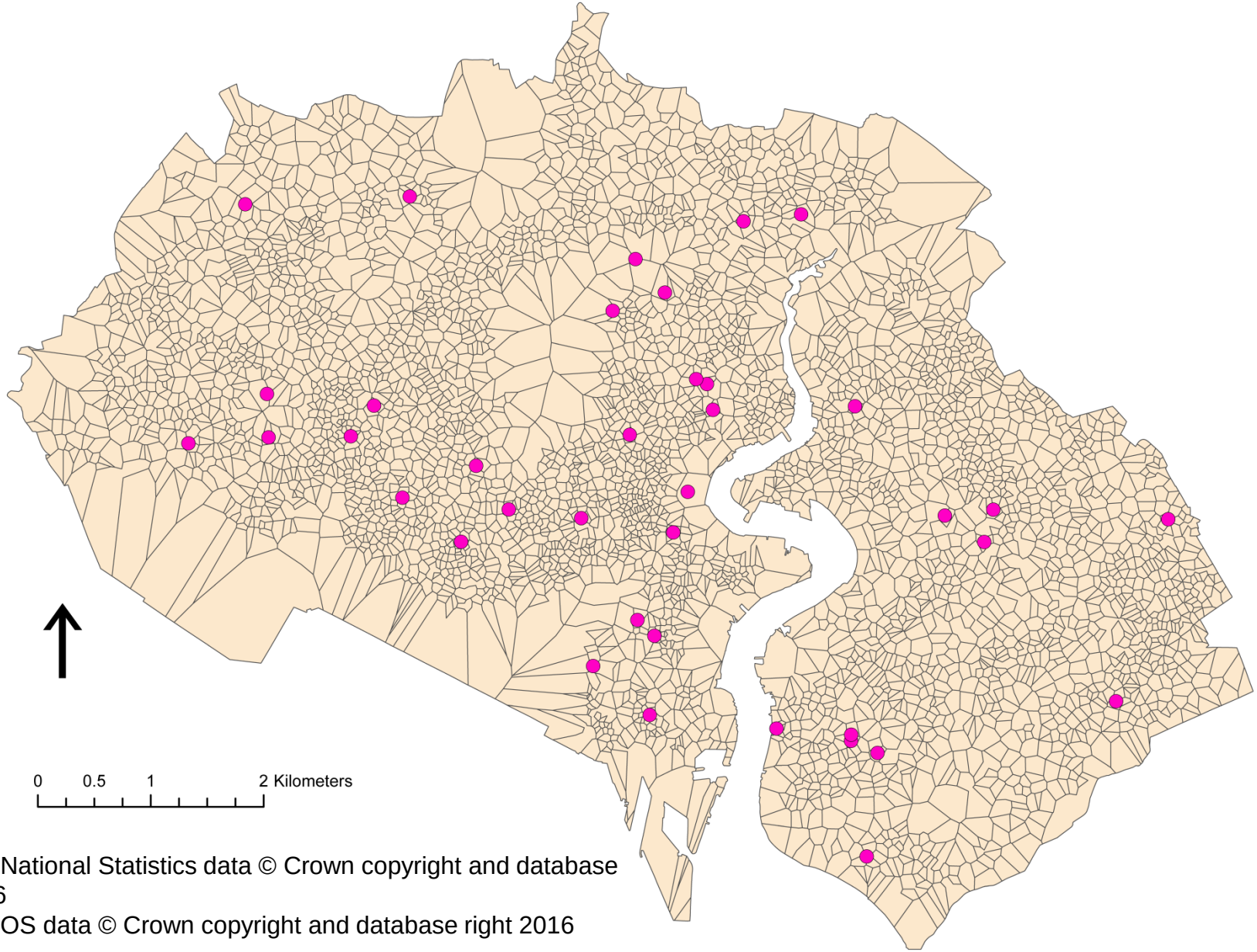
Postcodes and GPs

NCRM National Centre for Research Methods



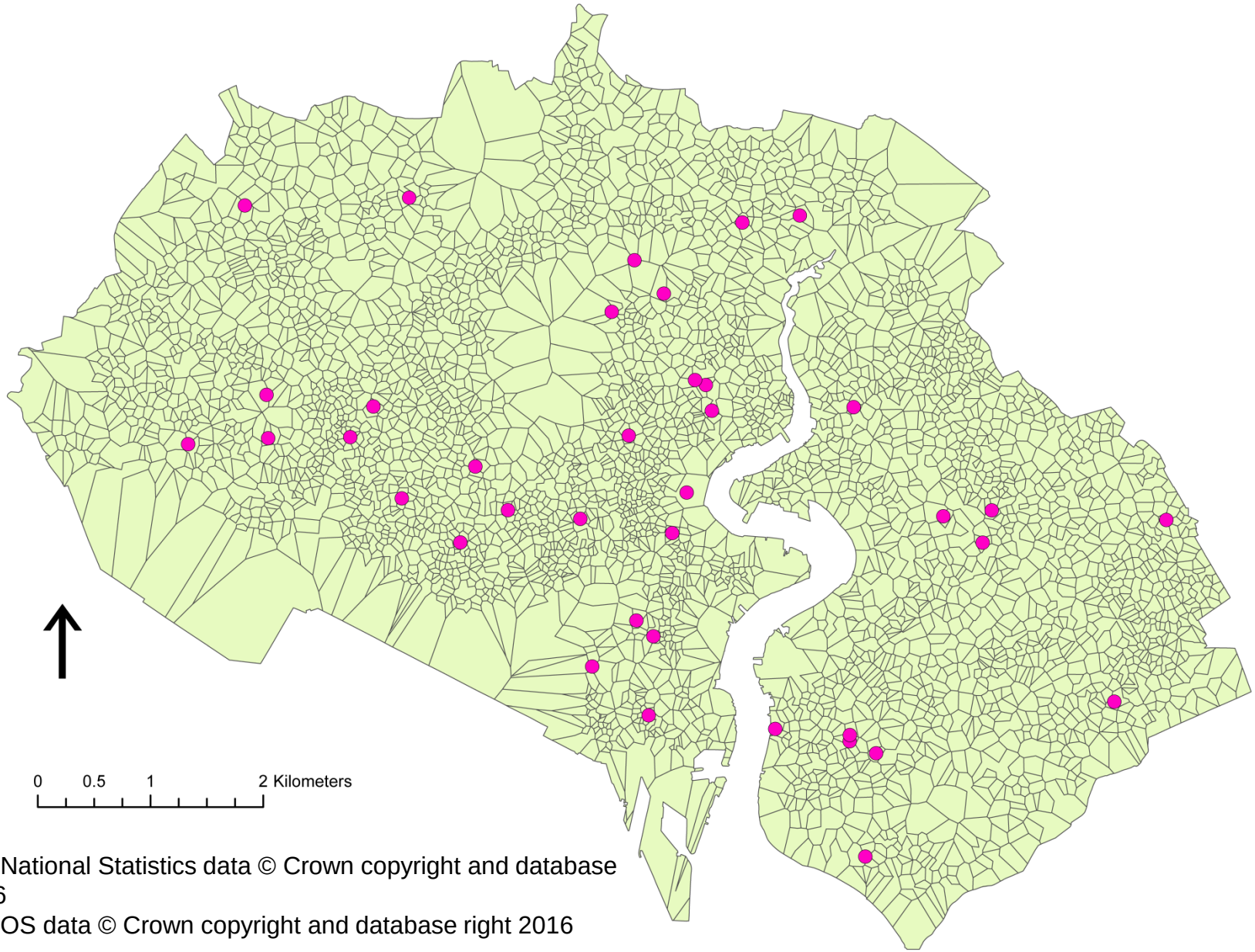
ZD100

NCRM National Centre for Research Methods



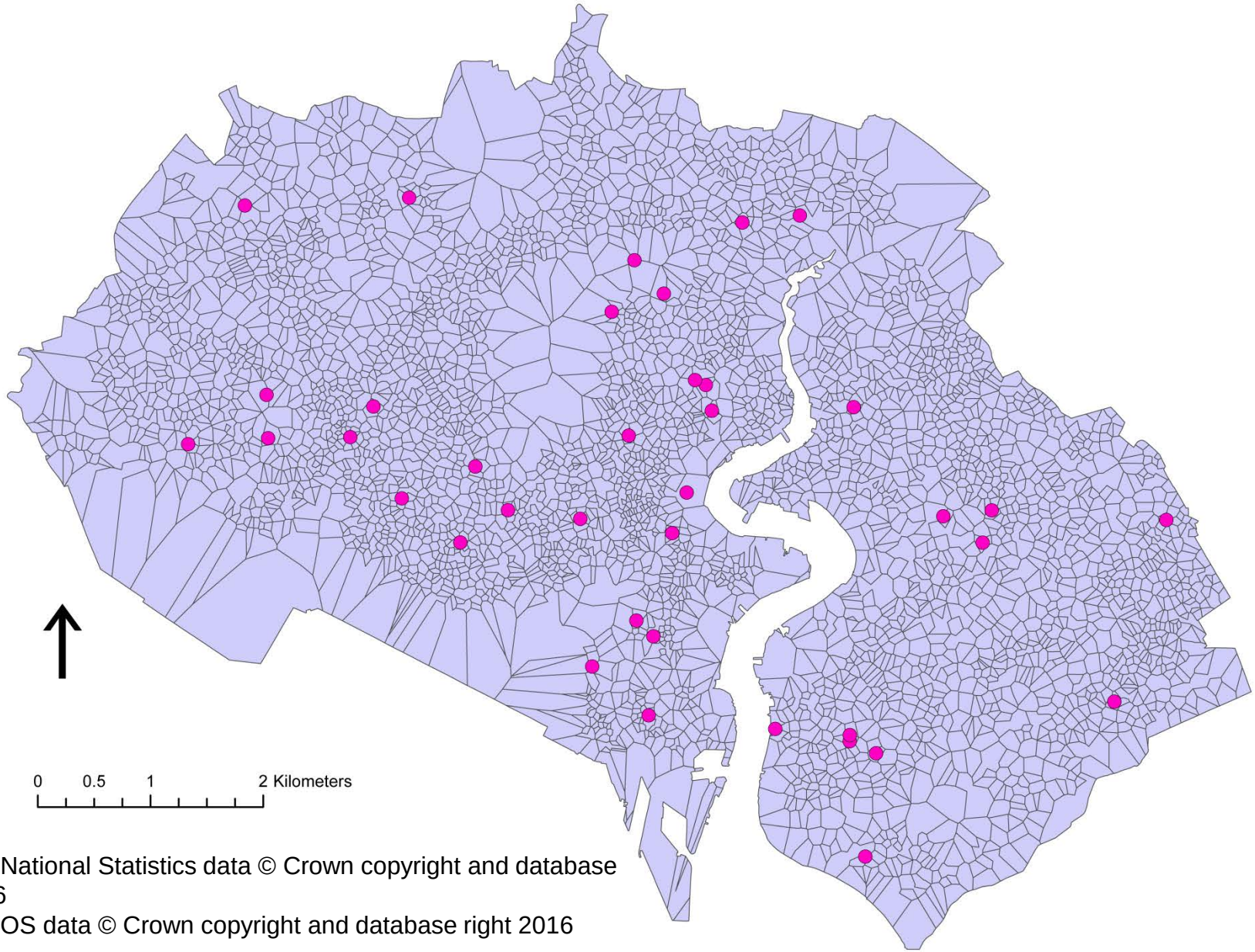
ZD150

NCRM National Centre for Research Methods

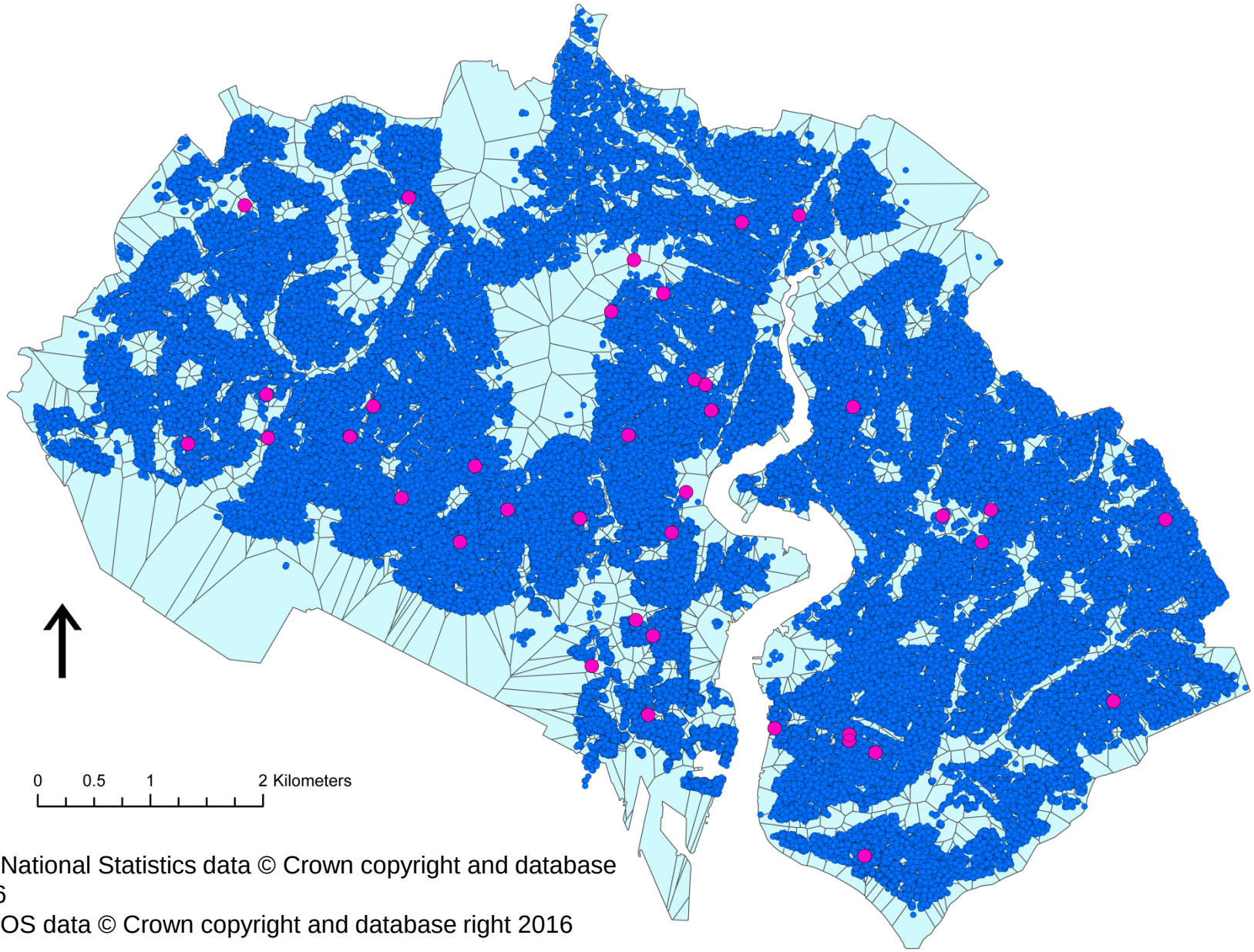


ZD200

NCRM National Centre for Research Methods



Households



NCRM National Centre for Research Methods

Example of using Zone Design with linked data

Geographic Unit	Number of units	Average population	Minimum population	Maximum population	Mean distance to GP (m)	Distance to GP – IMD rank score (R)
MSOA	32	47	34	70	828	0.176
LSOA	148	49	24	107	826	0.113
OA	766	53	15	246	817	0.081
Postcode	4,999	45	1	392	818	0.116
SZ 100	2,186	60	9	392	818	0.091
SZ 125	1,800	59	8	392	817	0.107
SZ 150	1,516	54	9	392	826	0.109
SZ 175	1,490	54	9	227	825	0.102
SZ 200	1,488	53	9	227	825	0.093
SZ 225	1,491	53	11	249	827	0.114

Conclusions

- Highlighting the pervasive nature of geographical units and the influence they have over analyses.
- Ongoing work. Next step is to control the lower population threshold – key disclosure constraint for data providers.

Conclusions

- Our research and technique is seeking to understand how to:
    1. Maximise protection for the data provider.
    2. Maximise data fidelity for the researcher.

- Transparency? There is the potential to quantify the zone design solution for the researcher.

- Zone design has a wider applicability to administrative and linked data.

Acknowledgements/references

- David Martin and James Robards are supported by NCRM, ESRC Award ES/L008351/1
- David Martin and Chris Gale are supported by ADRC-E, ESRC Award ES/L007517/1
- The data for this research have been provided by the Consumer Data Research Centre, an ESRC Data Investment

- AZTool http://www.geodata.soton.ac.uk/software/AZTool/
- Openshaw, S. (1984) *The Modifiable Areal Unit Problem,* Geobooks, Norwich, England.