

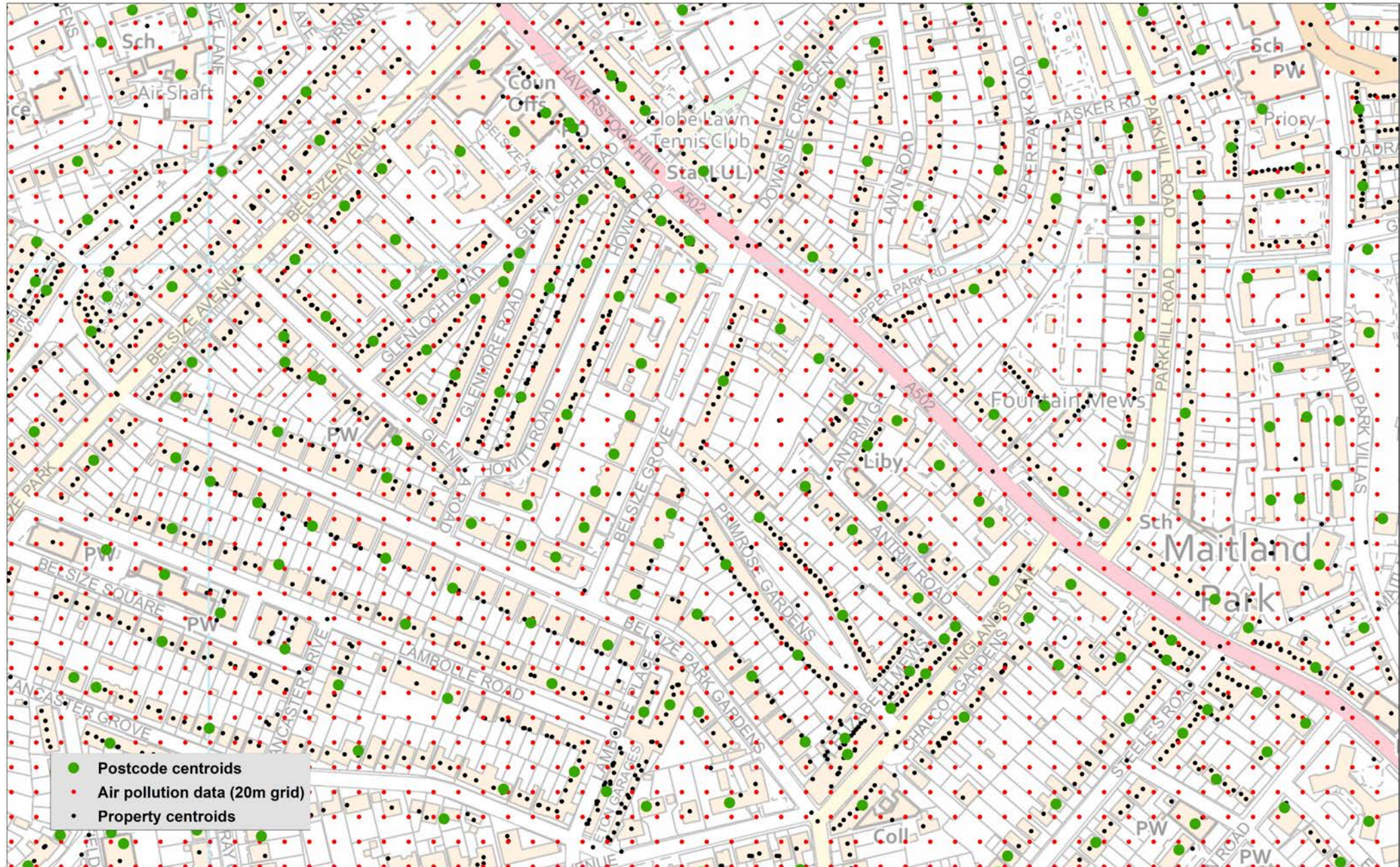
CLOSER Preparing for the future III: Tackling key challenges facing longitudinal population studies in a post-COVID world

Enhancing the birth cohort studies through address-level geographic linkages

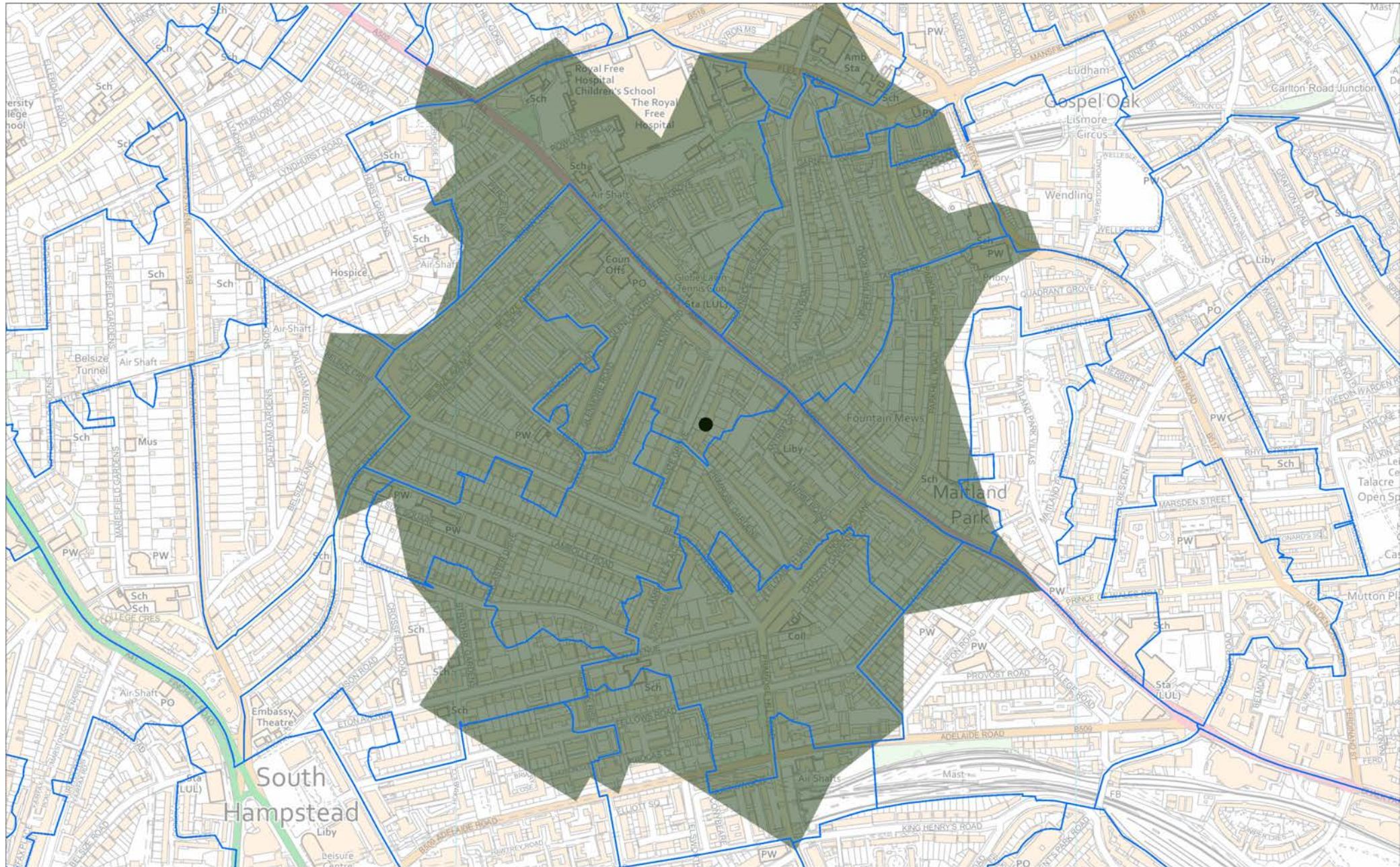
Benefits of address-level geocoding

- Increased accuracy when assigning environmental exposures to cohort members (e.g. air pollution)
- Ability to move away from arbitrary areal units (e.g. wards, LSOAs) and create more meaningful, neighbourhoods (e.g. by creating network-based buffer zones around each cohort member)
- Calculate more accurate distance measures between cohort members and a set of destinations of interest (e.g. parks, main roads, crime events)
- Eliminate assignment errors – e.g. up to 2 percent of cohort members at each survey sweep are allocated to the wrong LSOA based on their postcode centroid

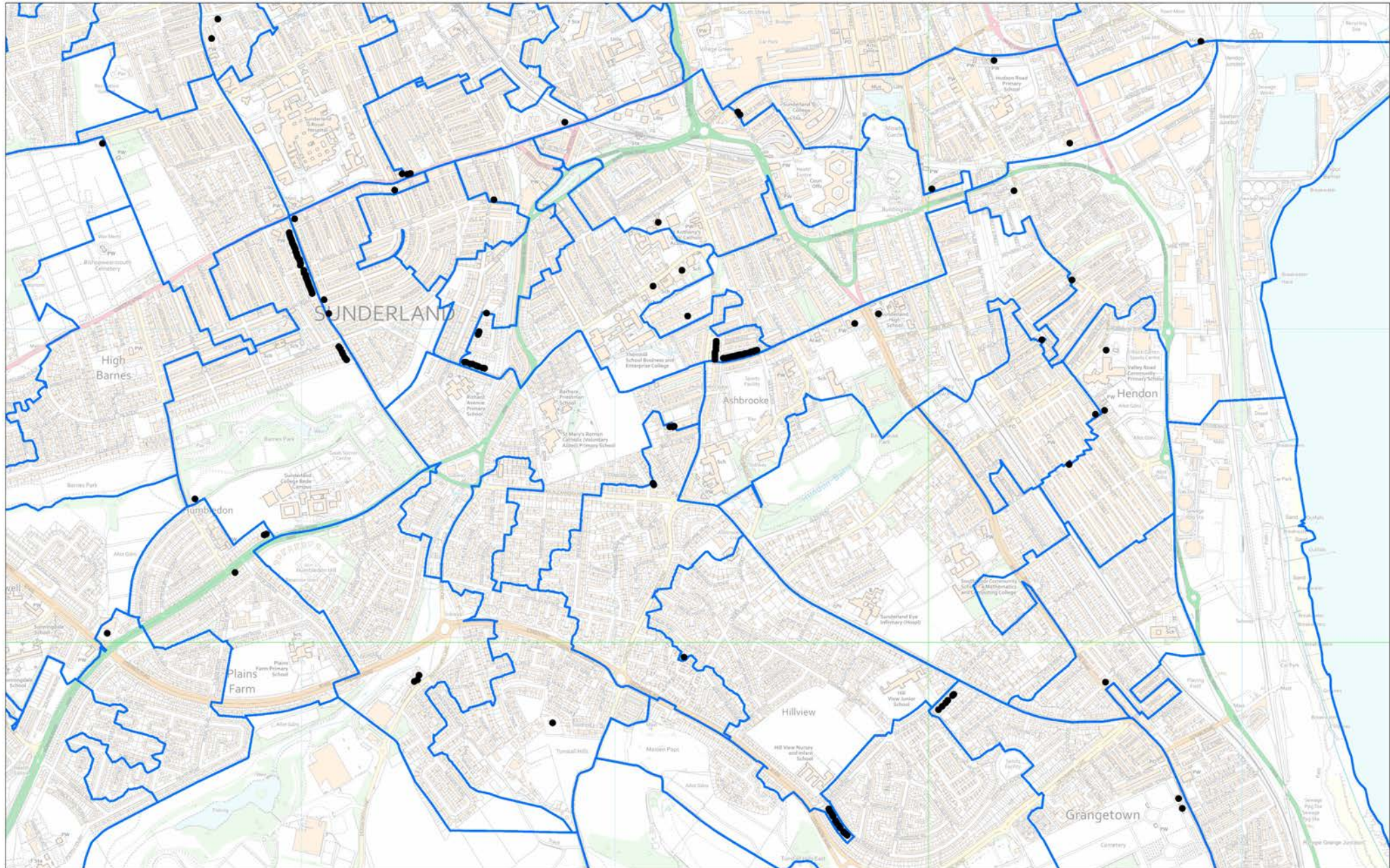
Allocating air pollution exposures - property vs postcode centroids



Self-defined neighbourhood (10 minute walk time from property) vs arbitrary areal units (2011 LSOAs)



Misallocation of properties to LSOAs based on postcode centroids



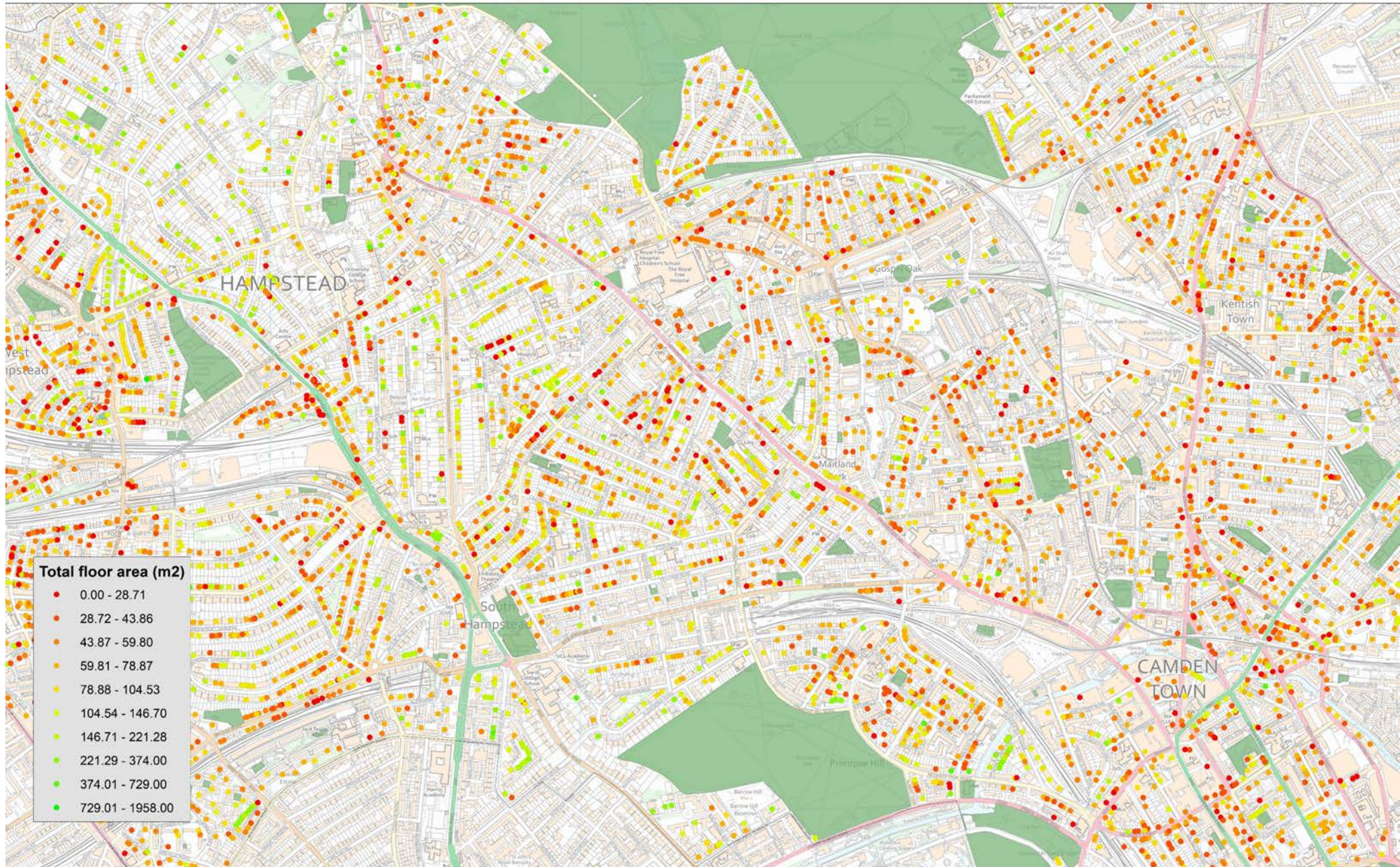
0 0.125 0.25 0.5 0.75 1 Kilometers



What datasets can we link cohort data to at the address level?

- Land Registry Price Paid data
- Energy Performance Data
- Historic OS mapping to enable geocoding of the earliest birth cohort study sweeps (e.g. NCDS 1958, 1965, BCS 1970)
- OS Points of Interest (e.g. supermarkets, bookmakers, fast-food outlets, cash points etc.)

Energy Performance Certificate data - Property floor area



Challenges in the address-level geocoding process

- Reference data (OS AddressBase Premium) are not freely available – need to establish relationship with Ordnance Survey
- Address data are continually updated/refreshed on a 6-weekly basis, therefore there are potential replicability/reproducibility issues as old properties are demolished and new ones are built (e.g. new build properties are most popular with the Next Steps cohort study)
- Still patchy use of the UPRN in non-OS datasets, which necessitates fuzzy matching based on single-field address
- Need to preserve cohort member confidentiality – address-level data cannot be deposited ‘as is’ at the UK Data Service, so deposited data need to be aggregated or researchers need a secure environment to work in (e.g. UCL Data Safe Haven)

Data sources

- **Ordnance Survey MasterMap** (Topography, Aerial, Network, Green spaces, POI, Historic): <https://digimap.edina.ac.uk/>
- **Ordnance Survey AddressBase Premium**: available to university researchers on application to OS (universityenquiries@os.uk)
- **Land Registry Price Paid data**, 1995 – present: <https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads>
- **Energy Performance Certificate (EPC) data**: <https://epc.opendatacommunities.org/>
- **Street lighting/trees data**: check relevant local authority websites, e.g. <https://opendata.camden.gov.uk/Environment/Trees-In-Camden/csqp-kdss>
- **Street-level crime data, 2011 – present**: <https://data.police.uk/data/>
- **DfT road accident data**: <https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data>
- **Income data** (2011, 2014, 2016, 2018): <https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/earningsandworkinghours/datasets/smallareaincomeestimatesformiddlelayersuperoutputareasenglandandwales>
- **Census Interaction data** (1981, 1991, 2001, 2011): <https://wicid.ukdataservice.ac.uk/>

My contact details:

David Church, GIS Developer, Centre for Longitudinal Studies, UCL Social Research Institute

utnvdch@ucl.ac.uk