



USING ADMINISTRATIVE DATA TO ESTIMATE THE POPULATION AND APPLICATIONS

Gillian Harper *Mayhew Harper Associates Ltd and
Les Mayhew* *Cass Business School and Mayhew Harper Associates Ltd*

There is considerable interest in the exploitation of administrative data to estimate the UK population instead of traditional methods based on a decennial census. This stems from the problem of population undercounting in parts of London and other English cities following the 2001 UK Census, the ten year delay between each census that renders the results out-of-date as soon as they are published two years later, and the substantial cost of around £500 million over the ten year cycle. Further, the outputs are inflexible and unsuitable to support local level service planning and delivery (Westminster City Council, 2002). A 2008 House of Commons Treasury Committee report considered official population statistics to be 'unfit for purpose'. According to the literature, the three main purposes of population counts are:

- for central revenue resource allocation to local authorities and primary care trusts (PCTs that are now to be re-structured under the new Coalition Government);
- for use as denominators; and
- for use by local authorities for planning and delivering services.

These issues have become even more pertinent subsequent to this research being completed with the Coalition Government announcing in July 2010 the intention to scrap the census in its existing format, deeming it as "an expensive and inaccurate way of measuring the number of people in Britain" (The Daily Telegraph, 9 July 2010).

Long before this announcement however, recognition of these issues led Mayhew Harper Associates to adapt their data linking 'neighbourhood knowledge management' technique (see www.nkm.org.uk) to estimate whole populations of local authorities. This technique utilises existing administrative data available in all local authorities and PCTs at the household level, thereby

offering a population estimation alternative which is similar in principle to 'population registers' that are found in Scandinavian and some other countries. The aim of the project was to fully describe and systemise this alternative approach, focusing on the methods and algorithms for merging diverse data sources and to assess its fitness for the three core purposes of population counts.

Key findings

- A methodology for combining local administrative datasets to create a population count was established using a formal system of logic to ensure replicability and rigour: a rule-based sequence of truth tables.
- The methodology can be used in any local authority in England and Wales.
- The administrative data methodology figures are consistent with other non-ONS statistics such as Child Benefit counts of children aged under 16.
- The statistics derived are timelier than the ONS Census because they use current data sources.
- The statistics derived by this route are more economical to produce than the ONS Census because they do not involve labour intensive and costly surveys, and therefore can be repeated frequently.
- The end product contains a wealth of demographic and socio-economic information at the individual and household level, including the age and sex demographic of each individual, which is unavailable elsewhere.
- This flexible and granular output provides greatly improved local planning intelligence (e.g. flexible spatial units, household demography and type).



- In the absence of consistent unique personal identifiers in the UK, data matching techniques are required.
- Quality improvements to the input administrative data (e.g. improved addressing) would lower the methodology's data matching requirements and reduce the number of residual unmatched records.
- Outputs are not identical to the census, but offer a wealth of socio-economic variables at the household and individual level that are not cross-referenced in a pre-determined manner.
- As a census alternative, individual local authorities could use these techniques to provide a population count to be fed into a national system. Certain procedures would need to be put in place to cover the whole country.
- We estimate the cost of an annual administrative data population count could be a tenth of the cost of the current census entire ten-year cycle.

Data sources

The input datasets are critical to the process. We accessed a number of standard administrative datasets and registers held by local authorities and PCTs, each of which records address information for every client (Table 1).

Metadata for each dataset was obtained to ensure the purpose and content and weaknesses and strengths of each dataset were fully understood. In the absence of one single comprehensive register that captures the entire local population, combining these different sources is essential to maximise coverage. Using the datasets in this way adds value beyond their original purposes. The GP Register is the most comprehensive of these datasets because it records the majority of a population and contains age and gender information, and is therefore used as the foundation of the methodology. The Local Land and Property Gazetteer (LLPG) or equivalent is critical by providing a base set of addresses to adhere to and provide standardised address formats and labels known as Unique Property Reference Numbers (UPRNs). UPRNs are the common denominator used to link the datasets.

Methodology

We needed the methodology to be systematic and rule based so that all assumptions are transparent and therefore replicable. The stages are set out in a series of truth tables to represent how all the datasets are incorporated to create a single final population count and database (Figure 1). Truth tables are used in Boolean algebra to test whether a logical expression is true or false for all legitimate input values (e.g. Lipschutz, 1998). These express when a person should be classified as a current resident at an address or not, based on the binary combination of the relevant factors relating to them from the input datasets.

Prerequisites are that the datasets are all current at the same snapshot in time, there are no duplicate people on the same dataset, and that every address is represented by

Dataset	Source	Purpose
GP Register	PCT	Records everyone registered with an NHS GP Practice
School Census	Local Education Authority	Records all children attending maintained schools in a Local Authority area (regardless of where they live) every January
Electoral Register	Local Authority	Records those aged 18 (or almost 18) and over who are eligible and registered to vote in local, European and General Elections, published every December
Council Tax Register	Local Authority	Records every domestic and mixed property liable for Council Tax, the name of the liable person(s) and the property's tax band
Council Tax and Housing Benefits	Local Authority	Records any locally administered benefit claims linked to a Council Tax property
Births	PCT	Public health birth records provided by ONS to PCTs at address level
Deaths	PCT	Public health death records provided by ONS to PCTs at address level
Housing Waiting List	Local Authority	Records people aged 16 and over and their dependants (not subject to immigration control) who are on the waiting list for a property in the Local Authority
Local Land and Property Gazetteer	Local Authority	Records all property addresses and land parcels in a Local Authority in BS7666 (British Standard) standardised format

TABLE 1. FEATURES OF AVAILABLE LOCAL ADMINISTRATIVE DATASETS

a UPRN from the property gazetteer. Each residential address (UPRN) on the property gazetteer is regarded as a household unit and current residents for each one counted. To summarise the above steps, the methodology address matches each dataset, takes the GP Register as the base, then cross-references the datasets by UPRN to assess who is current at each address, finally adding extra births and removing deaths. Sequential logical assumptions are used at each stage to determine who to include or exclude (Figure 1). The connectives are as follows:

^ and, v Or, ~ Not, → if-then

The first stage is to determine who on the GP Register can be classified as current residents at UPRNs and so can be

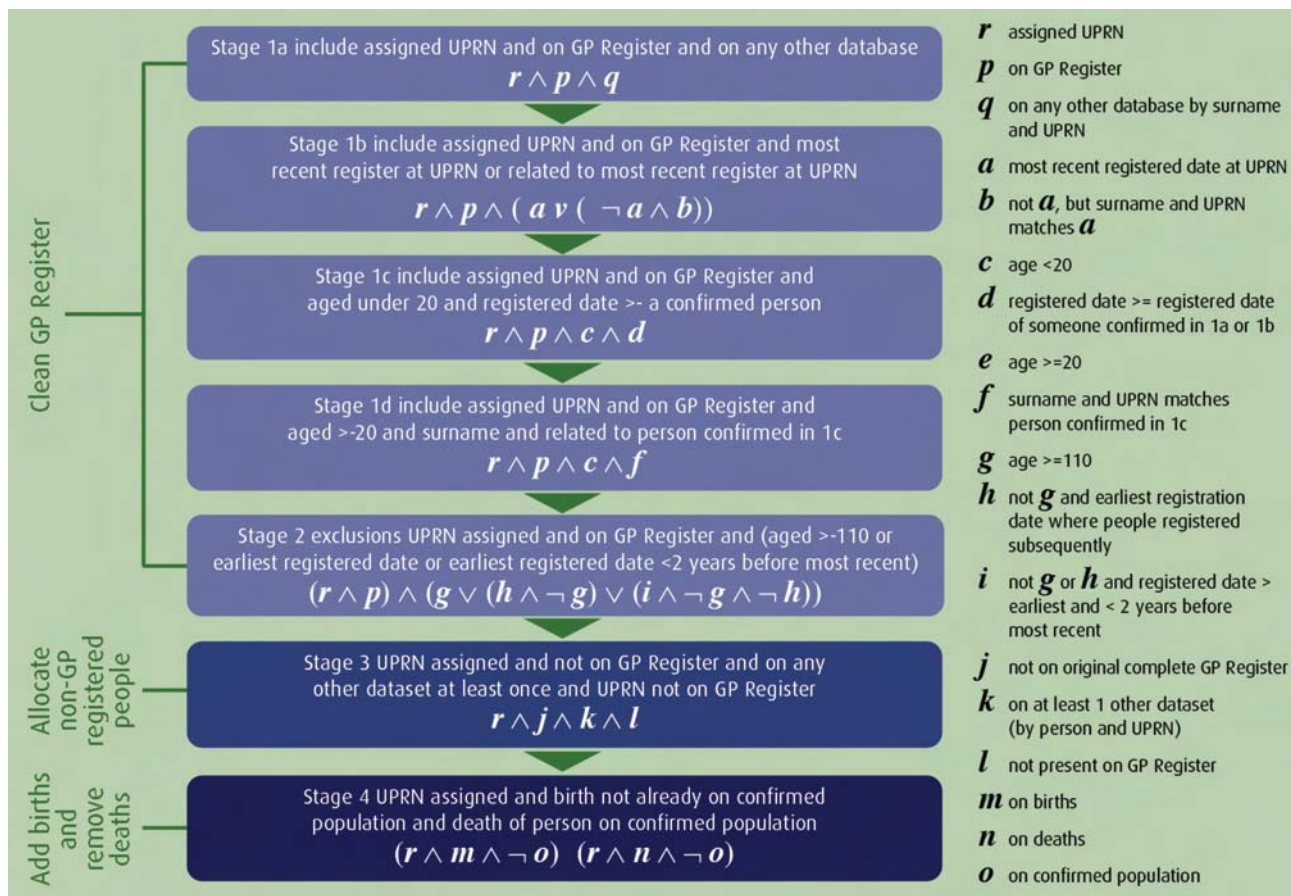


FIGURE 1. SUMMARY OF POPULATION ESTIMATION METHODOLOGY STAGES

included. The next stage of processing the GP Register is to identify who can definitely be excluded, that is, who no longer lives at an address and are part of any list inflation. The third stage looks to fill in any gaps in the population not covered by the cleaned GP Register, by allocating people on the other datasets into UPRNs that remain unused. The fourth and final stage is a last check at filling

in gaps that the other datasets have not been able to fill and to remove people who have died. The end result is a final dataset of the minimum confirmed population according to the rules of the algorithm, with each record representing a confirmed current resident, their age and sex and UPRN. The route to confirming a person as a current resident is summarised in Figure 2.

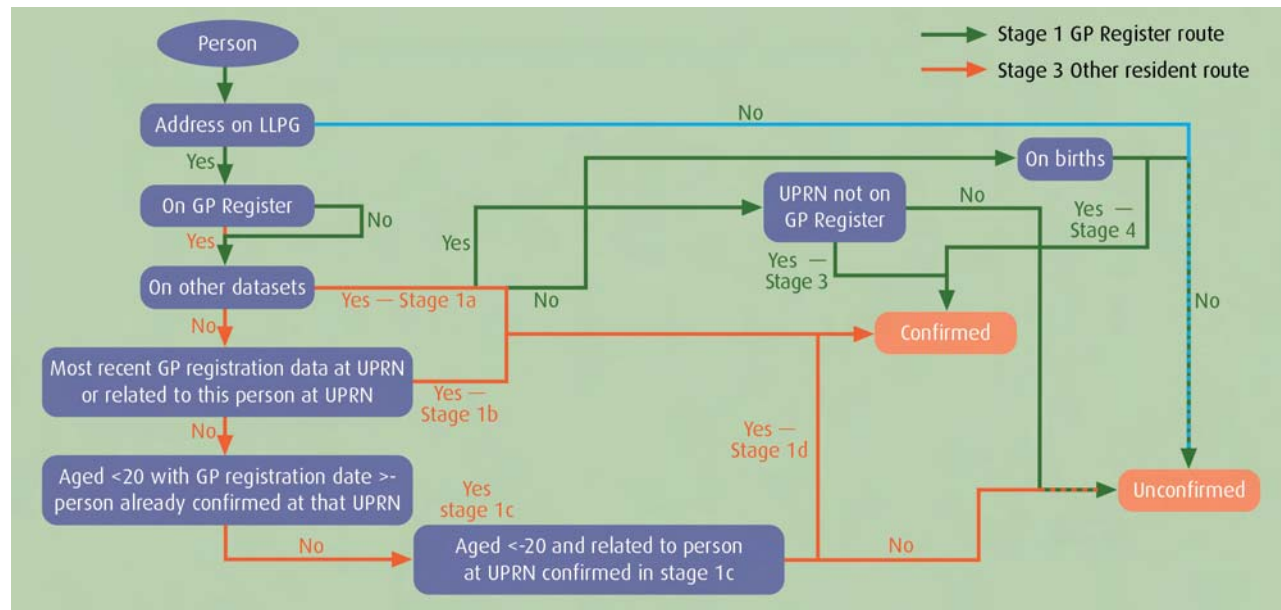


FIGURE 2. PATHWAY TO DETERMINE IF A PERSON IS A CURRENT RESIDENT AT A UPRN OR NOT

Evaluation

There is no single benchmark against which estimates can be compared. Instead, a number of ‘reasonability’ checks are carried out on the final population count to ensure that the results are sensible, taking into account timing and definitional differences. Since there are no absolute benchmarks that can be used it is necessary to refer to a range of sources. These include:

- Child Benefit numbers published by HMRC for children aged 0-16;
- State Pension claimants by males (65+) and females (60+);
- comparing the vacant UPRN rate with a local authority’s own figures or Council Tax records;
- UPRNs with high occupancy levels, greater than nine people, are identified and checked for being multiple-occupancy;
- the number of children aged <16 without an adult at the UPRN are counted and checked for possible explanations; and
- compared to other sources from similar date snapshots e.g. ONS Mid Year Estimates or GLA figures if the authority is situated for example in the London area.

The case studies showed consistent reasonable comparison figures in each case.

Residuals

Residuals are defined as records that have not been able to be included or verified, and are an important measure of the completeness of the methodology, and are represented as a Venn diagram in Figure 3. Each circle corresponds to the three main elements of the methodology — the GP Register, the property gazetteer (i.e. a record can be assigned a UPRN) and all other datasets.

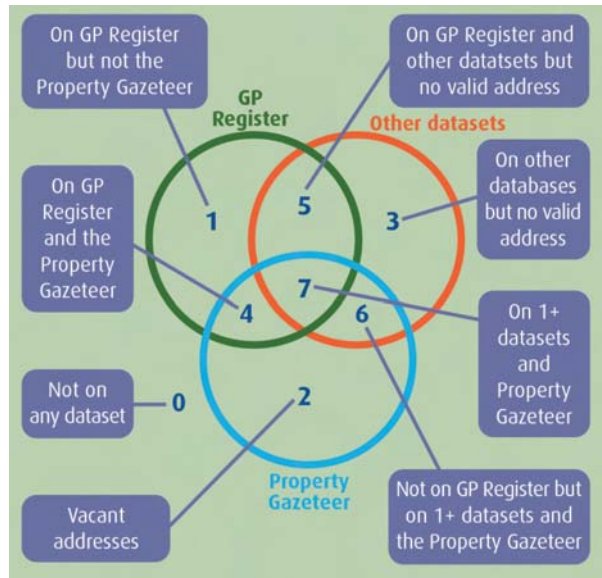


FIGURE 3. VENN DIAGRAM OF METHODOLOGY PRINCIPLES

Categories 4, 6, 7 are part of the confirmed population if they meet certain criteria. Categories 0, 1, 2, 3 and 5 are not part of the confirmed population and are instead treated as residuals. Residuals consist of dataset records for people who were not able to be assigned a UPRN, records for people who were assigned a UPRN but were not confirmed as current residents, and also duplicate records across the datasets for any of these aforementioned people, because people are liable to be present on more than one dataset. Those in category 0 are, by definition, unobservable and unquantifiable by this route. Each of the eight categories in the Venn diagram can be studied individually to examine why they have been created and propose any recommendations to deal with them.

Figure 4 is a flow diagram summarising the residuals and possible changes to how they are handled. The main outcome is maximizing UPRN assignment at the beginning of the process.

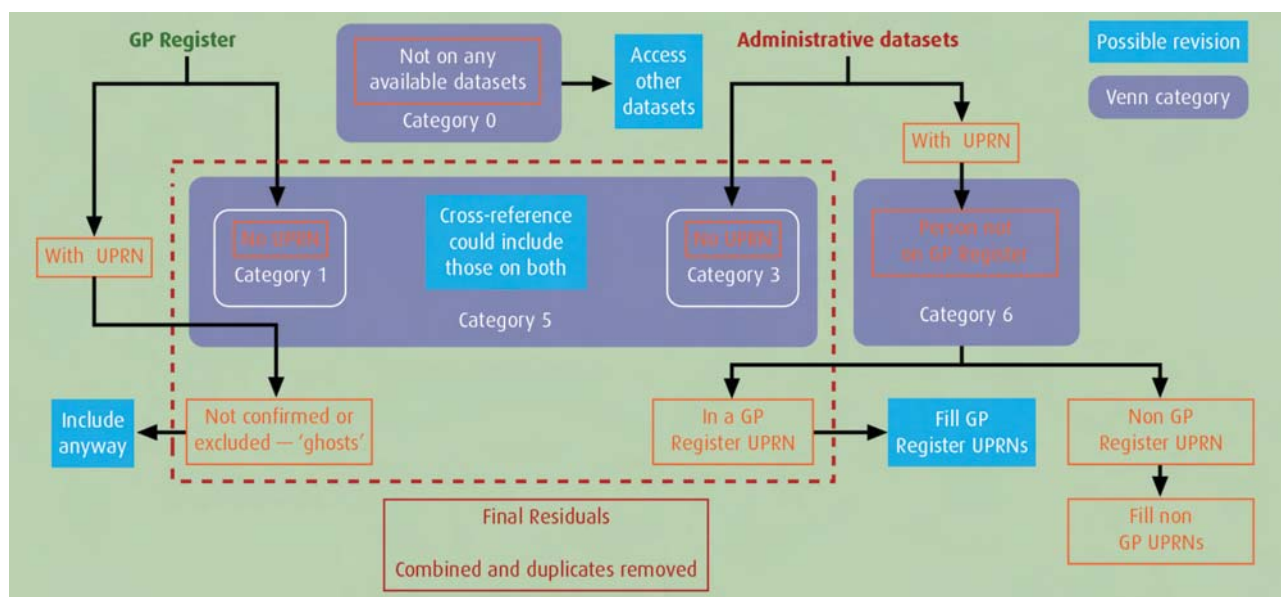


FIGURE 4. RESIDUALS AND POSSIBLE REMEDIAL ACTIONS

These residual sources are grouped together at the end to form a possible population 'extension'. The total number is an absolute maximum the confirmed population could be extended by, and the actual number of these that should be added is unknown and could in fact be zero. That is why the final result is called the 'minimum' confirmed population.

Algorithms

An important part of the methodology is data matching large datasets often with hundreds of thousands of records. A crucial consideration is that different datasets may be collected on different occasions for different purposes, and were not designed for easy, accurate matching. A number of algorithms have been developed to enable accurate and efficient data matching to support our approach.

Address Matching: For the purpose of the population estimation, every data record needs an address to act as a proxy for a household and to be used as the unit for capturing current residents. To ensure that the correct match is identified across datasets, the addresses are standardized by finding each address in the available property gazetteer and represented by a UPRN. A purpose-built address matching algorithm has been designed to do this. Unavoidably, a small percentage will remain that cannot be matched in this way. These tend to be formatted so differently from the gazetteer version that they are processed manually to choose the correct match. This can be a laborious process and so a further 'address finding' interface has been developed to facilitate this.

Person Matching: Person matching is used in the population estimation to ensure that the same person is matched across multiple datasets, particularly between the GP Register and other datasets. There is no single unique person identifier on the datasets to allow full exact matching, so a fuzzy technique is employed using the forename, surname and date of birth. We note that effective person matching techniques will become critical as the value of linking administrative data is increasingly recognised and if future censuses are to be constructed in this way.

Application

Once created, the population database can be used in numerous ways for analysing the population. The core database consists of the basic population count and the age and sex breakdown and combines this with additional variables extracted from the input datasets. The typical variables and if they apply to the person (P) or the UPRN are listed in Table 2.

The core database (personal identifiers are removed) provides a rich source of information on the population at a very high level of granularity. Variables can be cross-referenced at person or household level, either of which is under the control of the user. This is in contrast to ONS population counts which are typically constrained to wards or output areas. When combined with geographic co-ordinates for every household, users can select from the database any sub-group of the population they are

Dataset	Additional Variables
Council Tax Liable	Council Tax band (UPRN) Single Adult discount (UPRN)
Benefits	Receipt of Council Tax or Housing Benefit (UPRN)
Electoral Register	Elector type
School Census	Eligible for Free School Meals (P and UPRN) Ethnicity (P and UPRN) Language (P and UPRN) Special Educational Needs (P)
Property Gazetteer	Eastings (UPRN) Northings (UPRN) BLPU class i.e. property type (UPRN)
Births	Low birth weight (P)
Deaths	Cause of death (P)
Housing	Housing tenure (UPRN)

TABLE 2. TYPICAL VARIABLES FOUND IN CORE INPUT DATASETS

interested in based on any demographic and any geography. This is particularly useful for local analysis where areas of interest can be small and unconventional.

An example of applying the core database is given in Figure 5. Here, the database is used to assess access to child care in the London Borough of Brent. The local policy aspiration is that child care facilities are expected to be within pram pushing distance of children's homes. This is represented by mapping the location of every household containing a child aged under 5 and colour coded according to whether there are 0, 1, 2 or 3+ nurseries within short walking distance (i.e. a 500 metre radius representing a 10 minute walk). Areas with the least access and choice are identified by the dark blue dotted areas (e.g. cells E11, I13 or R12).

This information provided the local authority concerned with high quality information about potential geographic gaps in the service in which the numbers of children affected could be accurately quantified and their needs profiled. The accuracy of the information was seen to be a major advance over what is possible with official data.

CONCLUSION

The research has put forward a case for utilising and linking local administrative data to estimate local populations and as a better basis for intelligence-led policy and service planning at a local level. The method is current, economical and frequently repeatable, and also has the advantage of capturing people directly from extensive databases based on their presence at an address rather than relying on enumerating heads of households with postal surveys and depending on them to complete and return the forms.

The value of the use of administrative data over surveys for empirical sociology is discussed by Webber (2009) and Savage and Burrows (2009). Our research takes this further and demonstrates innovatively how the problems

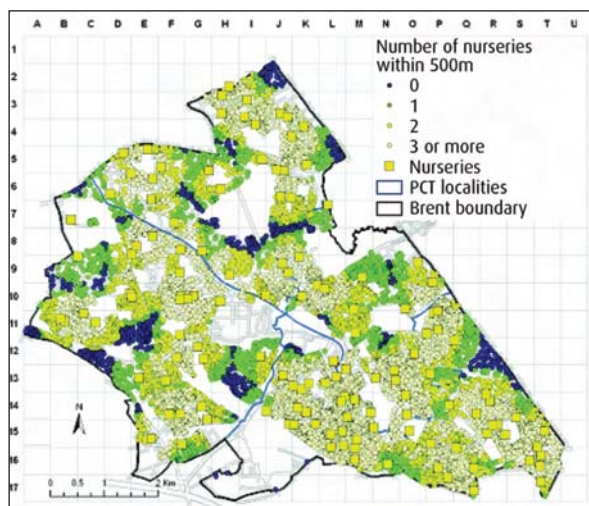


FIGURE 5. HOUSEHOLDS CONTAINING CHILDREN AGED UNDER 5 PLOTTED AND SHADED BY DISTANCE FROM NURSERIES

Source: London Borough of Brent;
Crown Copyright: Ordnance Survey

associated with the onus being on the citizen to self-report and self-return a census survey can be bypassed.

Implementing the methodology at a national level has not yet been attempted but can be considered as a matter of carrying out the population estimation for each of the local authorities in England and Wales, and combining them into national coverage. For this to happen, a number of key steps would be required to facilitate data sharing across organisational boundaries and between areas. Currently these challenges are greater than the technical challenges but if successful could be implemented at a fraction of the cost of conducting a census.

As an estimate, if an administrative data population count costs on average £100,000 per authority, the total cost for the 348 authorities in England and Wales would be £34.8million, plus say an additional £2million for co-ordination costs. This could be streamlined and made more economical over time. Theoretically then, an update could be carried out every year and the results disseminated the same year for the same cost that one ten year cycle of the census would cost.

Next steps

To improve the methodology the research could be developed in a number of ways. For example:

- by assessing if a minimum selection of datasets and fields are sufficient to achieve an accurate population count to make the method more efficient;
- by further developing the address matching routine so that they can deal with assigning UPRNs to

problematic addresses and therefore reduce residuals, and to further optimise the person matching routine;

- a study into the best way to adapt and extend the methodology to be appropriate over multiple local authorities;
- assessing the value of other datasets e.g. owned by central government that could contribute directly or act as confirmatory evidence at a person, household, output area or local authority level;
- more work on the assignment of people to different ethnic categories. Currently the methodology uses a combination of self-reported ethnicity and name recognition (Mayhew and Harper, 2010).

Acknowledgements

We acknowledge contributions from Sam Waples of Mayhew Harper Associates Ltd and thank Richard Verrall of Cass Business School and John Eversley of PPRE Ltd for their comments and support.

References

- Burrows, R. and Savage, M. (2009) Some further reflections on the coming crisis of empirical sociology, *Sociology*, 43: 762-772.
- House of Commons Treasury Committee (2008) *Counting the Population, Eleventh Report of Session 2007-8 Volume 1*, 14 May.
- Lipschutz, S. (1998) *Schaum's Outline of Set Theory and Related Topics*.
- Mayhew, L. and Harper, G. (2010) Counting with Confidence — The Population of Waltham Forest, March. www.walthamforest.gov.uk/index/community/wf-statistics/mayhew-report.htm
- The Daily Telegraph (2010) National Census to be axed after 200 Years, 9 July.
- Webber, R. (2009) Response to The coming crisis of empirical sociology: An outline of the research potential of administrative and transactional data, *Sociology*, 43: 169-178.
- Westminster City Council (2002) *Evaluation of Accuracy and Reliability of 2001 Census*, November.

Contact details of authors

Gillian Harper
Mayhew Harper Associates Ltd
Email: harpergill@gmail.com

Les Mayhew
Mayhew Harper Associates Ltd and
CASS Business School, City University London,
106 Bunhill Row, London, EC1Y 8TZ.
Email: lesmayhew@googlemail.com