A Study of the Use of Linked Routinely Collected Administrative Data at the Local Level to Count and Profile Populations

by
Gillian Harper

A Dissertation Submitted to

Cass Business School
City, University of London

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

April 2017

Contents

1 Introduction 11
1.1 Overview 11
1.2 Papers 14
1.2.1 Paper 1 – Using administrative Data to Count populations 14
1.2.2 Paper 2 - Applications of Population Counts Based on Administrative Data at Local Level 16
1.2.3 Paper 3 - Using Administrative Data to Count and Classify Households with Local Applications 17
1.2.4 Paper 4 - Impact of Asthma on Educational Attainment in a Socioeconomically Deprived Population: A Study Linking Health, Education and Social Care Datasets 20
1.3 Individual contribution to the co-authored papers presented in the thesis 21
1.3.1 Personal contribution to the work presented in chapter 2 21
1.3.2 Personal contribution to the work presented in chapter 3 23
1.3.3 Personal contribution to the work presented in chapter 4 23
1.3.4 Personal contribution to the work in chapter 5 24

2 Using Administrative Data to Count Local Populations 26
2.1 Introduction 26
2.2 Background 29
2.3 Data Sources 32
2.4 Methodology 34
2.5 Residuals 38
2.6 Evaluation of Results 40
2.7 Matching Algorithms 41
2.7.1 Address Matching 42
2.7.2 Person Matching 43
2.8 A Worked Example 44
2.9 Conclusions 51

3 Applications of Population Counts Based on Administrative Data at Local Level 56
3.1 Introduction 56
3.2 Limitations of Official Population Statistics 58
3.2.1 Resource Allocation 58
3.2.2 Use as Denominators 59
3.2.3 Use in Delivery of Local Services 61
3.2.4 Geography as a Barrier 62
3.2.5 Use of Administrative Data in Practice 64
3.3 Data Structures Using Administrative Data Sources 66
3.4 Application Example 70
3.4.1 Background to Case Study 70
3.4.2 Take-up Rates of Free Eye Tests Under the NHS 74
3.4.3 Geographical Access to Free Eye Tests 75
3.4.4 Impact of Geographical Access on Take-up Rates 77
3.4.5 Evaluation of an Alternative Service Configuration 79
3.4.6 Implications for Resource Allocation 81
3.5 Discussion 82
3.6 Conclusions 83

4 Using Administrative Data to Count and Classify Households with Local Applications . 86
4.1 Introduction 86
4.1.1 Why Analyse Households? 86

4.1.2 Present Arrangements and the Case for Change ... 88
4.1.3 Aims of Paper ... 91
4.2 Creating Household Statistics from Administrative Data ... 92
4.2.1 Background to Demographic Counts ... 92
4.2.2 Alternative Household Classification Systems Using Administrative Data ... 93
4.2.3 Enumerating Household Types ... 95
4.2.4 Mapping Household Counts on to Standard Types ... 98
4.2.5 Examples of Household Enumeration ... 99
4.3 A Case Study: Child Poverty in Hackney ... 102
4.3.1 Access to Children's Centres in Hackney ... 107
4.4 Administrative Counts Versus Official Household Statistics ... 109
4.5 Comparison of Household Types Using Official Figures ... 114
4.6 Reasons for Differences Between Sources ... 117
4.7 Discussion ... 118
4.8 Appendix 4.A - Measuring Statistical Quality of Population Estimation and Household Counts from Administrative Data Method ... 120

5 Impact of Asthma on Educational Attainment in a Socioeconomically Deprived Population: A Study Linking Health, Education and Social Care Datasets ... 123
5.1 Introduction ... 123
5.2 Methods ... 124
5.2.1 Study participants ... 124
5.2.2 Outcome variables ... 124
5.2.3 Predictor variables ... 125
5.2.4 Data linkage ... 126
5.2.5 Statistical methods ... 128
5.3 Results ... 130
5.3.1 Primary analysis ... 131
5.4 Discussion ... 133
5.4.1 Summary ... 133
5.4.2 Strengths and weaknesses ... 134
5.4.3 Comparison with other studies ... 135
5.4.4 Clinical and policy relevance ... 135
5.4.5 Conclusion ... 136
5.4.6 Acknowledgments ... 136
5.5 Appendix 5.A - Sensitivity analyses ... 136

6 Conclusions ... 139
6.1 Impact of the research ... 140
6.1.1 Academic penetration ... 140
6.1.2 Commercial implementation ... 141
6.1.3 Informing national statistics and trailblazing administrative data strategy ... 145
6.1.4 Research Excellence Framework (REF) recognition and award ... 151
6.2 Critical reflection ... 151
6.3 Overall Summary ... 152

List of Tables

Table 2.1: Features of available local administrative data sets 32
Table 2.2: Example of a simple truth-table based on Figure 2.1. Key: A accept; R reject 36
Table 2.3: Population count audit trail for a case study 45
Table 2.4: Comparison of case study population age breakdown from different sources 48
Table 2.5: Enumeration of rejected records for case study 51
Table 3.1: Typical structure of currently available official population data (OA = Output Area) 68
Table 3.2: Typical structure of databases using administrative data sources (OA = Output Area) 68
Table 3.3: Household structure, population, tenure and benefit status 73
Table 3.4: Table segmenting the population of Tower Hamlets by access to eye test centres according to the given risk factors 77
Table 3.5: Alternative resource allocation scenarios 82
Table 4.1: Specific examples of households defined by size and age group (Key: O indicates a person) 95
Table 4.2: Possible combinations of household demographic types based on size and age (see text for details of the highlighted cells) 97
Table 4.3: Mapping household demographic combinations on to the eight standard types, A to H for the case of three age groups and up to four occupants 99
Table 4.4: Summary table showing a breakdown of households across the six Olympic Boroughs and selected key attributes 100
Table 4.5: Average household age and occupancy 102
Table 4.6: Summary table showing a breakdown of households in Hackney according to the number of children, housing tenure and benefit status according to three different communities as at 2011 105
Table 4.7: Odds of income deprivation by risk factor including 95 % confidence intervals (CI) 106
Table 4.8: Comparison based on total number of households by local authority using administrative, DCLG, GLA and ONS Census data, and the % difference of each compared to the administrative data counts 111
Table 4.9: Comparison of count and % of vacant dwellings by local authority using administrative, DCLG, GLA and ONS Census data, ... 113
Table 4.10: The government household typology scheme 115
Table 4.11: Household type counts in London Borough of Hackney using administrative, DCLG, and ONS Census data, and the % difference of each compared to the administrative data counts 116
Table 5.1: Key stage tests for the UK's National Curriculum 124
Table 5.2: Characteristics of 12,136 pupils that sat Key Stage tests in 2002 to 2005. 128
Table 5.3: Coefficients, 95% confidence intervals and P-values for the effect of socio demographic and clinical variables on standardised attainment scores in Key Stage tests 1, 2 and 3, from multiple regression model allowing for clustering 131
Table 5.4: Table 5.3: Coefficients, 95% confidence intervals and P-values for the effect of socio demographic and clinical variables on standardised attainment scores in Key Stage tests 1, 2 and 3, from multiple regression model allowing for clustering 133

List of Figures

Figure 2.1: Simple Venn diagram partitioning different categories of administrative data with and without addresses 35
Figure 2.2: Summary of population count methodology stages 37
Figure 2.3: Pathway to determine if a person is a current resident at a UPRN or not 38
Figure 2.4: Residuals and possible remedial actions 39
Figure 2.5: Extended UPRN assignment flow chart. Key: SAON = Secondary Address Object, PAON = Primary Address Object 43
Figure 2.6: Distribution of high UPRN occupancy levels resulting from the case study 46
Figure 2.7: Chart showing the differences in estimates by age group between the administrative count and ONS and GLA 49
Figure 3.1: The flow of administrative data at national and local level 66
Figure 3.2: Density of Bangladeshi population in Tower Hamlets by Lower Super Output Area (LSOA) (Contains Ordnance Survey data © Crown copyright and database right 2010, and data sourced from London Borough of Tower Hamlets) Note: LAPs are Local Area Partnerships 71
Figure 3.3: Geographical access to eye testing centres based on 10-minute walk time or 500 m. Round symbols indicate locations of households with one or more persons aged 60+ (Contains Ordnance Survey data © Crown copyright and database right 2010, and data sourced from London Borough of Tower Hamlets) 75
Figure 3.4: Free eye test take-up in the 60+ population based on distance from nearest eye test centre 78
Figure 3.5: Geographical access to GP practices based on 10-minute walk time or 500 m. Round symbols indicate locations of households with one or more persons aged 60+ (Contains Ordnance Survey data © Crown copyright and database right 2010, and data sourced from London Borough of Tower Hamlets) 79
Figure 3.6: Predicted change in eye test take-up in the 60+ population following re-configuration 80
Figure 4.1: Stages in the production of person and household level data and policy domains supported 94
Figure 4.2: The eight standard household types from administrative data 96
Figure 4.3: Scatter-gram of household types showing occupancy versus average age by output area 101
Figure 4.4: Map showing the locations of households on benefits with children aged<5 that are outside pram pushing distance from the nearest children's centre 108

Acknowledgements

The research material in this thesis is from papers published previously in peer-reviewed journals. Chapters 2, 3, 4 and 5 were originally published in Harper and Mayhew (2012a), Harper and Mayhew (2012b), Harper and Mayhew (2016), and Sturdy et al (2012) respectively. The author is grateful to the editors of the journals involved (Journal of Applied Spatial Analysis and Policy, and PLOS ONE) for giving their permission for the papers to appear in this thesis. Thanks also to the anonymous referees for their comments and suggestions that led to the improvement of the published papers.

The author would like to thank ESRC (ESRC RES-163-27-0019: 'Using Administrative Data to Estimate the Population and Measure Deprivation') for funding to be able to convert the methodology into the academic papers in chapters 2 and 3.

The author would like to acknowledge all the helpful comments and suggestions made by her colleague Professor Les Mayhew, and Professor Richard Verrall in his capacity as supervisor.

Further, the author would also like to acknowledge the contributions to the research of Professor Les Mayhew and John Eversley, and the Blizard Institute, Queen Mary, University of London, and the support of Cass Business School.

Lastly, the author is grateful to Jon and her friends for their support and always believing in her.

Declaration

The author grants powers of discretion to the University Librarian to allow this thesis to be copied in whole or in part without further reference to the author. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

Abstract

There is increasing evidence that official population statistics are inaccurate at the local authority level, the fundamental administrative unit of the UK. The main source of official population statistics in the UK comes from the decennial census, last undertaken in 2011. The methodology and results of official population counts have been criticised and described as unfit for purpose. The three main purposes of population statistics are resource allocation, population ratios, and local planning and intelligence.

Administrative data are data that is routinely collected for administrative purposes by organisations, government departments or companies and not for statistical or research purposes. This is in contrast with surveys which are designed and carried out as a specific information gathering exercise. This thesis describes a methodology for linking routinely collected administrative data for counting and profiling populations and other purposes at the local level.

The benefits of this methodology are that it produces results more quickly than the decennial census, in a format that is more suitable for accurate and detailed analyses. Utilising existing datasets in this way reduces costs and adds value.

The need and the evolution of this innovative methodology are set out, and the success and impact it has had are discussed, including how it has helped shape thinking on statistics in the UK. This research preceded the current paradigm shift in the UK for research and national statistics to move towards the use of linked administrative data. Future censuses after 2021 may no longer be in the traditional survey format, and the Office for National Statistics are exploring using a similar administrative data method at the national level as an alternative. The research in this thesis has been part of this inevitable evolution and has helped pave the way for this.

List of Abbreviations

ADC	Administrative Data Census
ADRN	Administrative Data Research Network
BTS	British Thoracic Society
BY2011	Beyond 2011
CIS	Customer Information System
CTP	Census Transformation Programme
DCLG	Department for Communities and Local Government
DWP	Department for Work and Pensions
ESS	European Statistical Service
FTE	Full-time Education
GIS	Geographic Information Systems
GLA	Greater London Authority
GP	General Practice
GROS	General Register Office for Scotland
HESA	Higher Educational Statistics Agency
HMO	Houses in Multiple Occupation
HMRC	Her Majesty's Revenue and Customs
IMPS	Improving Migration and Population Statistics Programme
JSNA	Joint Strategic Needs Assessments
LAP	Local Area Partnerships
LARIA	Local Area Research and Intelligence Association
LB	London Borough
LLPG	Local Land and Property Gazetteer
LSOA	Lower Super Output Area
MAUP	Modifiable Areal Unit Problem
MYE	Mid-year Estimates
NHS	National Health Service
NISRA	Northern Ireland Statistics and Research Agency
nkm	Neighbourhood Knowledge Management
NLPG	National Land and Property Gazetteer
NROS	National Records of Scotland
NTTS	New Techniques and Technologies for Statistics
ONS	Office for National Statistics
PAON	Primary Addressable Object Name
PCT	Primary Care Trust
PRS	Private Rented Sector
REF	Research Excellence Framework
RSS	Royal Statistical Society
SAON	Secondary Addressable Object Name
SAR	Sample of Anonymised Records
SHLAA	Strategic Housing Land Availability Assessment
SPD	Statistical Population Dataset

SQL	Structured Query Language
UPN	Unique Pupil Number
UPRN	Unique Property Reference Number
UPTAP	Understanding Population Trends and Processes

1 Introduction

This thesis comprises four research papers on the use and linkage of routinely collected administrative data for counting and profiling populations and other purposes. These papers are titled:

1. Using Administrative Data to Count Local Populations
2. Applications of Population Counts Based on Administrative Data at Local Level
3. Using Administrative Data to Count and Classify Households with Local Applications
4. Impact of Asthma on Educational Attainment in a Socioeconomically Deprived Population: A Study Linking Health, Education and Social Care Datasets

The first, second and third papers were published in the journal Applied Spatial Analysis and Policy in June 2012 (online April 2011), September 2012 (online March 2011) and December 2016 (online August 2015) respectively. The third paper was also initially a Cass Business School Actuarial Research Paper. All these papers were co-authored with Professor Les Mayhew.

The fourth paper was published in PLOS ONE in 2012, co-authored with Pat Sturdy, Stephen Bremner, Les Mayhew, Sandra Eldridge, John Eversley, Aziz Sheikh, Susan Hunter, Kambiz Boomla, Gene Feder, Keith Prescott, and Chris Griffiths.

Chapter 1 gives an overview of the research in section 1.1, and a focus on each of the individual papers in section 1.2. My individual contribution to each paper is described in section 1.3. Chapters 2 to 5 contain each of the above four papers in order, with a discussion and conclusion in chapter 6.

1.1 Overview

This thesis sets out an innovative methodology of linking routinely collected administrative data to create new information on local populations. It also demonstrates the value and impact of the methodology through a range of different applications.

The approach described in this thesis came about due to growing evidence that official population statistics are inaccurate at the local authority level, the fundamental administrative

unit of the UK. The main source of official population statistics in the UK comes from the decennial census, last undertaken in 2011. This population count is conducted as a household survey, and disseminated as aggregated counts for small geographical areas. The population counts are projected for each of the interim years as Mid-Year Estimates (MYE) based on assumptions about future fertility, mortality and migration. The methodology and results of these population counts have been criticised and were described as 'unfit for purpose' by the House of Commons Treasury Select Committee in May 2008 (House of Commons Treasury Committee, 2008).

The three main purposes of population statistics are resource allocation, population ratios, and local planning and intelligence (House of Commons Treasury Committee, 2008).

The growing dissatisfaction with the census, including its high cost, led to an announcement by the Rt. Hon Francis Maude MP, Minister for the Cabinet Office, that the 2011 Census would be the last. However, in 2014 the National Statistician recommended a predominantly online census in 2021, supplemented by increased use of administrative data and surveys (Office for National Statistics, 2014).

Administrative data are information that is routinely collected for administrative purposes by organisations, government departments or companies and not for statistical or research purposes. This is in contrast with surveys which are designed and carried out as a specific information gathering exercise.

Together with Professor Les Mayhew, I have been developing a system for the exploitation of administrative data and counting local populations since 2000. The approach links together person and address level administrative data that are routinely collected by local authorities and health trusts, to assess who are recorded as the current residents at each address on each dataset. The confidence of the accuracy of this being the correct address at a fixed point in time for each person is determined by which dataset they are on, and how many datasets in total they are on.

An algorithm applies a set of rules to establish this, and results in a final count of the population for that point in time, defined as the 'confirmed minimum population' resident in that local authority area.

The output is a population database where each record corresponds to a person, and contains the age and gender and address of each person. This is a higher level of geographical granularity than official population statistics outputs which are aggregated at small area geographical level, thus providing more detail and enabling more flexible analyses. This is because the address level data can be built up into any geography of the user's choice.

Additionally, other datasets of interest can be linked to the population database to build a more detailed profile e.g. service use and other socio-economic variables relevant to that person or household. This gives the user control of which population to study and which variables to cross-reference, something that is not possible with existing official population statistics.

The research in this thesis argues that the use of locally-available administrative datasets for counting populations produces results more quickly than the decennial census, in a format that is more suitable for accurate and detailed analyses. Utilising existing datasets in this way reduces costs and adds value.

The potential for these data to be accessed for the purposes of social science research and population statistics is increasingly recognised, although it has not as yet been fully exploited.

The objective of this thesis is to demonstrate the need and the evolution of this innovative methodology of linking and modelling administrative data to create new information on local populations, and the value and impact of its applications.

The papers in chapters 2, 3, 4 and 5 describe the administrative data methodology for counting local populations, and how capturing and organising data in this way benefits local decision makers and service providers and communities.

Some of the context in the papers in chapters 2 to 5 are specific to the time of publication, and has since changed. Chapter 6 gives an update on these aspects and a general review of current population data science.

Chapter 6 critically reviews the research, considering how the method is a trade-off in that it loses some beneficial aspects of a traditional census such as consistent variables for longitudinal research, and nationally comparable data, to achieve its distinctive advantages

mentioned previously. In this way, it is explained how the method is most suitable for the purposes of local planning and intelligence.

The success and impact of this work, including how it has helped shape thinking on statistics in the UK, is also discussed in chapter 6.

Each of the papers has its own introduction and literature review. The purpose of the rest of this section is to provide a short summary of the more important aspects of the four papers, and to describe the connecting features.

1.2 Papers

1.2.1 Paper 1 – Using Administrative Data to Count Populations

The first paper (Harper and Mayhew, 2012a) contained in chapter 2 sets the scene as to why existing official population statistics derived from surveys have become increasingly unsatisfactory for some users. The emphasis is on users of population statistics at the local level, such as the English and Welsh local authority geography or smaller.

The methodology originally came about to meet a demand for more accurate local population estimates. This was to provide evidence in response to the Census 2001 results which were significantly under-counted in some cities in England and parts of London, resulting in serious under-funding to these areas (Bowley, 2003; Statistics Commission, 2004).

The paper was timely in that it coincided with the publication of a 2008 House of Commons Treasury Committee report (House of Commons Treasury Committee, 2008) that declared current population statistics to be 'unfit for all purposes required', and the beginning of a move by official population statistics to explore how the use of administrative datasets could improve accuracy and reduce costs and output dissemination times.

The research sets out an alternative methodology that combines locally available administrative data sources with different population coverage according to a defined set of rules. It is a logical systematic methodology that is reliable and replicable, and benchmarks and quality assures the results as far as possible. A case study is used to demonstrate this.

Several key concepts relating to this method emerge from the paper. These are:

- The necessity of good metadata and understanding the purpose and scope and quality of the input data sources
- The importance of confident data linkage in the absence of consistent unique identifiers as is the case in England
- The value of linking data and creating outputs at the individual person and/or household level rather than aggregations
- The problems associated with choosing appropriate definitions of e.g. a household
- The effects of geographical scale
- The fact that trade-offs are required and are acceptable if the outcome is fit for purpose

As the first attempt of this kind for this purpose in the UK, it was important to assess the strengths and weaknesses of the methodology. It was decided that sensitivity analysis was not practical. Instead, the results were compared to a range of other sources to act as benchmarks and were found to be highly aligned, with differences of less than 2% at the aggregate level.

Most encouragingly, revised ONS population estimations for the case study area became more closely aligned with the administrative data results. Importantly, the administrative data methodology returned these results within approximately three months, but the ONS took three years to produce their figures.

Overall, the paper concludes that a more accurate population estimate for local areas is likely to be obtained from the administrative data methodology that relies on current data rather than the decennial census procedure which is out of date and synthetically adjusted. The former is also lower cost, requires less resources and has a quicker turn-around. Weaknesses are its reliance on input data quality and consistency and a need for more thorough quality assurance testing. These advantages and disadvantages are part of the trade-off when choosing this method.

1.2.2 Paper 2 - Applications of Population Counts Based on Administrative Data at Local Level

The second paper (Harper and Mayhew, 2012b) contained in chapter 3 follows on from the first paper by illustrating in more detail how the outputs from the methodology can be applied, and importantly, setting out real-life examples of applications that would not be possible otherwise using official population statistics.

The limitations of official population statistics are set out within the framework of the three main purposes of population statistics, which are resource allocation, population ratios and local planning and intelligence (House of Commons Treasury Committee, 2008). And within the two main components of official population statistics: the demographic population spine and the socio-economic characteristics and variables. The population spine is the basic demographic count of the population with age and gender. The socio-economic variables are the extra information relating to each person or household, captured by the ONS census with survey questions including e.g. employment status, housing tenure, limiting long-term illness. These are considered crucial by some census users, who are resistant to any changes in how the census captures information, so that these socio-economic variables remain consistent to support longitudinal analysis.

It is demonstrated in the paper that the administrative data methodology does provide an accurate population spine and a rich array of socio-economic variables, albeit not identical to those on the census.

The paper argues that an advantage of the methodology over official population statistics are that by outputting a population spine at the individual person and household level rather than aggregated, the data are more flexible because it can be assigned to any user-defined geographical area. Any population or geography of interest can be studied without being restricted to fixed boundaries and without the bias and causality issues of the ecological fallacy and the Modifiable Areal Unit Problem (MAUP).

Another main advantage is that the granular level of the output enables direct data linkage to other sources to create new variables of interest so that cross-referencing is not pre-determined as in the census, but controlled by the user. The results can also be produced within a much shorter timescale meaning that the outputs are timelier and can be updated more quickly.

It becomes clearer throughout the paper that the methodology is particularly suited to the third purpose of population statistics – local planning and intelligence. For this, flexible, timely, targeted and bespoke population intelligence is required to inform the increasing demand for innovative policy, decision making and service planning.

A statistical method called 'risk ladders' is used that calibrates the influence of individual risk factors on outcomes. This is only possible with the granular linked population data created by the methodology.

Overall, the paper points out the significant deficiencies and disadvantages of official population statistics and shows how administrative data can be captured and structured as a solution to these.

This contributes to the debate as to whether administrative data can provide a full national census replacement in terms of the variables available in addition to a population spine. This is discussed more fully in chapter 6.

1.2.3 Paper 3 - Using Administrative Data to Count and Classify Households with Local Applications

For the third paper (Harper and Mayhew, 2016) contained in chapter 4, we turn our attention fully to one of the applications touched upon in the second paper – how the administrative data methodology can be used to count and classify households. This is an important topic because households rather than individuals are being increasingly used for research and to target and evaluate public policy.

The census and other official statistics sources provide household counts and typologies derived from survey methods. Therefore, the administrative data methodology also needs to be assessed in its suitability to provide this function if it is to be considered as a full replacement to these sources.

The literature review in the third paper sets out the reasons why knowing the attributes and types of households in an area or population is needed across economic, deprivation, political, health and housing applications. It also sets out why official housing statistics are not always able to meet that need sufficiently.

In the same way as official population statistics as discussed in paper 2, official housing statistics are also too aggregated, out of date, inflexible, and unable to be linked easily to other data sources for effective local planning and policy.

An important advantage of administrative data is that it is possible to add attributes to households that are not available in official data. The paper gives examples of this for six local authorities in London that made up the 2012 Olympic area, all of which had administrative data population estimates carried out previously.

The research describes a system for producing flexible classifications and enumerations of household types using locally collected administrative data at address level and compares this with official sources to highlight similarities and differences.

In England, official housing counts and classifications are provided by the Department for Communities and Local Government (DCLG)[1]. DCLG use the census and subsequent population projections as a baseline to provide indicative figures of future numbers by household type if past demographic trends were to continue, using household composition proportions (Department for Communities and Local Government, 2010b).

The administrative data approach to counting households takes the administrative data population estimates for the study area as the base, which contains the age and gender of residents for every address within the local authority areas. These demographic attributes are summarized and decomposed into a typology of eight comprehensive mutually exclusive household categories.

By applying this typology to the study area, it is demonstrated that this alone provides a useful profile breakdown of the population. An example of how linking additional household level variables to the typology to assess which types of household have the highest propensity to be living in low income and what are the risk factors with the most influence on this is given. This is a common local authority policy evidence requirement, and the example illustrates that the administrative data methodology can be used to support better service planning.

[1] This function was taken over by ONS in January 2017
https://www.ons.gov.uk/news/news/transferofhouseholdprojectionstoons

The paper then turns to exploring how much confidence users can have in the administrative data enumeration of households. This is done by comparing total household counts and household typology counts with three different sources of official figures.

To do this, the administrative data for the study area had to be moulded into the same household type definitions used in the official statistics. While this was possible, it highlighted definitional issues within and between the sources. Examples of this include: whether 'couple' households include same sex couples or not; whether we need to know the marital status; and whether a household is a housekeeping unit or a dwelling.

In comparison, the counts from the administrative data methodology are close to the counts from the official sources. Even more encouraging is that the biggest discrepancies are seen to reduce with revised versions of the official counts that fall more in line with the administrative data results. These revisions were required to deal with errors in the baseline Census 2001 population counts.

These results gave an initial confident indication that administrative data sources of household counts and types could be a satisfactory replacement to official sources.

To gain a more thorough grasp on the confidence that can be placed in the results, a further quality control checklist based on Eurostat's European Statistical Service (ESS) six dimensions of quality was carried out on the administrative data methodology as far as possible. Standard statistical measures of confidence intervals were not applicable in this case, and a reliance on external comparators was required instead.

Of the six dimensions of quality, the methodology was found to be very strong on the dimensions of relevance, timeliness and accessibility. It was less strong on the dimensions of accuracy, comparability and coherence, but it was concluded that this was not critically so. This is another component of the trade-off in the administrative data methodology.

Again, the approach is found to be particularly relevant for local authority planning and intelligence applications where accuracy, timeliness and detail are important.

1.2.4 Paper 4 - Impact of Asthma on Educational Attainment in a Socioeconomically Deprived Population: A Study Linking Health, Education and Social Care Datasets

The fourth paper in chapter 5 (Sturdy et al., 2012) takes the research down a slightly different path, into an important epidemiological application studying the impact of asthma on educational attainment.

This research does not employ or assess the administrative data population or household count methodologies described in the previous papers. Instead, it explores the value of using administrative data in research by linking administrative health, education and social care data together to enable the statistical analysis of the hypothesis that asthma adversely affects performance in national school examinations. Without this linkage, the study would not have been possible.

The same data linkage techniques were used as in the previous papers to link disparate general practice (GP), housing and education databases for a local authority in London. This enabled the study to address a wide range of clinical and socio-demographic factors and importantly, to explore relationships between clinical factors and social outcomes.

From this, it was possible to use a wealth of variables in the regression model at the individual child level. It was also possible to tease out very specific influencing factors.

This paper found no evidence for an adverse effect of asthma or asthma severity on examination performance. Instead, ethnicity (Bangladeshi children), social adversity (eligibility for free school meals, living in social housing, one parent households and households with a smoker) and those with mental health problems and special educational needs were related to poorer examination performance.

This analysis would not have been possible without the availability of the administrative datasets at the individual person and household level, and the linkage between them. Like the previous papers, it demonstrates the new information and value and granular detail that can be created from routinely collected data. Most significantly, this research made use of such data and techniques to make a valuable contribution to health research and provide important evidence for health care policy makers.

1.3 Individual contribution to the co-authored papers presented in the thesis

To give context, the author and Professor Les Mayhew have been developing a system for the exploitation of administrative data and counting populations since 2000. This collaboration is entitled Neighbourhood Knowledge Management (*nkm*) for commercial purposes. The key initiating project that led to the population estimation work was for the London Borough of Brent, who were looking for assistance in providing evidence that their count of population from the 2001 Census was too low, resulting in a reduced funding allocation from central government.

Through both our ideas, we established the methodology to link administrative datasets that capture different parts of the population, dealing with the overlap between these, de-duplication, and confirming who are current residents and who are not. The results were well received by the London Borough of Brent and set off a demand from many other local authorities for similar analyses.

Over time, the innovative nature of the methodology and its results, and its value and impact became clearer, and an opportunity was sought by myself to structure the work into an academic research framework of a standard that could be published in a peer-reviewed journal. This opportunity came about with ESRC UPTAP (Understanding Population Trends and Processes) Fellowship funding, where I was funded from June 2008 to December 2009 as an UPTAP Research Fellow at Cass Business School to do further research and convert the work initially into an academic report. This was titled 'Using Administrative Data to Estimate the Population and Measure Deprivation' (ESRC RES-163-27-0019), and was published as an UPTAP report and presented at the UPTAP conference.

My thesis is comprised of four papers, all of which are co-authored, three with one other author and one with eleven other authors. My specific contribution to each is set out in the next section. Co-authors provided signed declarations agreeing with the contributions as described.

1.3.1 Personal contribution to the work presented in chapter 2

The first paper presented in chapter 2 was authored by Gillian Harper and Les Mayhew.

The UPTAP Research Fellowship allowed for further time to be spent adapting the UPTAP report into the form of this first paper.

My contribution to the methodology was built on my academic background in geography and geographic information science, and my practical skills in data analysis and management. Specifically, I provided expertise in administrative and address data and data linkage, understanding the content and limitations of each individual dataset, and creating, carrying out and quality assuring an effective and accurate data linkage procedure.

I devised the majority of the population estimation rules that establish who is the same person across datasets, and who is the 'current confirmed' resident at an address. This was based on my knowledge of what information was available in each dataset and how they overlap, and my breakdown of the residual records that remain after the 'current confirmed' resident is identified.

I was responsible for all data management, and design and implementation of the methodology and the creation of the final definitive population database. These aspects were managed and carried out in a database environment using SQL (Structured Query Language).

As well as the practical data aspects, I was involved in establishing a theoretical background to the research and undertook the detailed literature review. It was evident that there was no one else doing a similar type of methodology in the UK at that time, so there was little to compare it to. Instead, the review had to look towards what else had been done with administrative data and why it was suitable for the demand we were witnessing for an alternative way to count local populations, and what quality was required.

Professor Les Mayhew brought statistics and policy and planning expertise to the work, and had input into these aspects of the methodology and the paper, in particular the Venn diagram concept in figure 2.1, and the truth-table binary framework. These were used to validate the methodology, provide a conceptual and theoretical context and to verify that the methodology was replicable across different datasets.

I wrote the full paper content myself including undertaking the literature review as the main author, with Professor Les Mayhew providing feedback and edits where appropriate.

After the successful feedback of the UPTAP final report, the paper was filled out with more detail, re-edited several times, and successfully submitted to and published in the journal Applied Spatial Analysis and Policy.

1.3.2 Personal contribution to the work presented in chapter 3

The second paper presented in chapter 3 was authored by Gillian Harper and Les Mayhew.

After the completion of the first paper it was obvious that a subsequent stand-alone paper could follow on focusing on the applications of the methodology set out in the first paper. This had already been done to some degree under the UPTAP funding and time allowance, but needed to be polished and edited.

These applications were devised from real-life cases from consultancy work undertaken by myself and Professor Les Mayhew using data mostly sourced from Tower Hamlets Primary Care Trust and the London Borough of Tower Hamlets, which is one of the most deprived boroughs in the country.

Professor Les Mayhew wrote the bulk of the new version and conceived the examples. The paper included the first published example of Professor Mayhew's risk ladder methodology and household typology. I implemented the tables, maps and other supporting analyses that were included in the final paper and contributed to the editing and finalisation of the submission.

1.3.3 Personal contribution to the work presented in chapter 4

The third paper presented in chapter 4 was authored by Gillian Harper and Les Mayhew.

While the previous paper explored a variety of applications of the methodology, one application emerged as highly valuable in local authority applications from this research and other consultancy work. This was converting the administrative data population database into a relevant household typology by summarising the demographic information available for each property address. From our experience of working with client local authorities, we considered this typology to be more useful and flexible than the official statistics typology available at the time, and decided to explore this further as a research paper.

I undertook the first draft of the paper at Cass Business School as a Researcher in 2012. Initially this was a Cass Business School Actuarial Research Paper published in 2012 (Actuarial Research Paper number 128). It was then submitted to and successfully published in the journal Applied Spatial Analysis and Policy.

Professor Les Mayhew was responsible for the middle, mainly methodological sections including the combinatorial formulation of the household typology itself and the devising of the local application based on data from the London Borough of Hackney. He also helped to edit down and revise the introduction, and devised the policy context in the introductory paragraphs. I was responsible for a) produced all the data used; b) for the comparison of administrative data with official household statistics and types; c) the interpretation of reasons for any differences between them and d) producing the final tables and maps. The rest of the paper, i.e. the literature review, comparisons with official sources, quality issues and the discussion sections were carried out and written up by myself.

An important element of the paper is the quality measure of the methodology in Appendix 4.A. This was carried out on my initiative as I felt this was something that so far had been lacking in the previous papers. The methodology by nature is difficult to assign quality to using standard statistical confidence measures, and the European Statistical Service Dimensions of Quality seemed appropriate, as they are also used by official statistics providers including ONS.

1.3.4 Personal contribution to the work in chapter 5

The third paper presented in chapter 4 was authored by Pat Sturdy, Stephen Bremner, Gill Harper, Les Mayhew, Sandra Eldridge, John Eversley, Aziz Sheikh, Susan Hunter, Kambiz Boomla, Gene Feder, Keith Prescott and Chris Griffiths.

My contribution to the fourth paper took a slightly different form, as part of a large academic research team of twelve people. My role was to provide administrative data and linkage expertise, and I was solely responsible for the creation of the definitive database used for analysis and modelling by the statisticians in the team. This involved intensive data management and linkage and quality assurance checks to combine the disparate data sources of GP patient records and clinical data with education and socio-economic datasets using an SQL database environment.

The study would not have been possible without this. I composed the section on 'data linkage' in the paper, while the other sections were contributed to by the other authors as appropriate. The other authors were statisticians, clinical experts, asthma experts and air pollution experts, led by Professor Chris Griffiths.

It is stated in the paper that I contributed to conceiving and designing the experiments; performing the experiments; analysing the data; contributed reagents/materials/analysis tools; and writing the paper.

2 Using Administrative Data to Count Local Populations

Preface: Content in this chapter consists of an exact reproduction of the article published in the Journal of Applied Spatial Analysis and Policy in 2012. Only minor edits have been made to make numbering consistent throughout the thesis. As such there may be some dated references or statements. Developments in population data science since the time of publication of this paper are described in Chapter 6.

2.1 Introduction

There is considerable interest in the exploitation of administrative data to count 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 10-year gap between each census that renders the results out-of-date as soon as they are published 2 years later, and the substantial cost of around £500 m over the 10-year cycle. These counts are used as the basis for subsequent annual Mid-Year Estimates (MYE) between censuses and so contribute to a range of problems further down the line until the next census. In 2008, a House of Commons Treasury Committee report, noting that there had been substantial problems in generating accurate population estimates in some areas during the 2001 Census, declared population statistics to be 'unfit for all purposes required' (House of Commons Treasury Committee, 2008). In addition, users complain that the outputs are inflexible and unsuitable to support local level service planning and delivery (Westminster City Council, 2002; Keohane, 2008).

The first censuses of sorts, such as the Domesday Book[2], took place before the first official Great Britain Census in 1801. This was enabled by the Census Act 1800, driven by a growing concern about the population of Britain and its demand for food (Malthus, 1888). In the 20th Century, the demand for population statistics increased steadily, in large part due to the gradual transfer of powers, including control over funding, from local to central government over many decades in areas such as health and education, and social security. Although population statistics have a wide range of uses, it is only in recent decades that their accuracy has been recognized as a critical factor in certain applications. One of these applications is the formulaic basis for allocating money from the government to local authorities and key public

[2] http://webarchive.nationalarchives.gov.uk/20160110200228/http://www.ons.gov.uk/ons/guide-method/census/2011/how-our-census-works/about-censuses/census-history/early-census-taking-in-england-and-wales/index.html

services such as health[3]. Modern formula-based allocation methods are technically sophisticated, containing variables that are linked one way or another to population counts so that if these are inaccurate results will be skewed. Since the mid-1990s population statistics have acquired further uses in the governing of the country through the widespread growth in the use of targets for holding a wide range of public services to account. Targets are often expressed as ratios with population as the denominators and the function or activity of interest in the numerator (e.g. the percentage of adults who are economically inactive).

Although the new Coalition Government (2010) has now abolished targets, the 'target culture' became pervasive under Labour (1997–2010) with hundreds of examples drawn from areas as diverse as law enforcement, education, housing, employment, health, social services and waste disposal. However, if anything the Coalition has increased the demand for local data due to the onus on public services to make themselves more transparent to consumers. This is expected to add to the already growing range of other applications at sub-local authority level in which accurate population counts are needed to effect policy, ensure value for money and be more accountable to citizens. The problem is that many of the claims promulgated for service improvements are based on local population statistics that are spurious at best because of the poor quality of the data.

These issues have become even more pertinent subsequent to this research being completed with the announcement in July 2010 of 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' (Hope 9th July 2010). Long before this announcement however, recognition of these issues led Mayhew Harper Associates to adapt their data linking 'Neighbourhood Knowledge Management (nkm)[4] technique to count whole populations for local authorities. This technique utilises existing administrative data available in all local authorities and primary care trusts (PCTs) at the household level, thereby offering a population count alternative which is similar in principle to 'Population Registers' that are found in Nordic and other countries.

In this paper, we describe a methodology for combining local administrative data sets to create a population count using a formal system of logic to ensure reliability, established on a rule-based sequence of truth tables. In a practical application of the methodology, we show

[3] In health sector, the history begins in 1970 with the Labour Government's Green Paper on NHS reorganisation which included a commitment to a new method of resource allocation. This led to the Crossman formula and then later to the RAWP formula in the same decade. For subsequent history see Thompson (2010).

[4] See www.nkm.org.uk

that the administrative data methodology figures are consistent with other administrative data sources such as Child Benefit and state pension counts. Because it is quicker to do than a census, data derived from this process are timelier than the census conducted by the Office for National Statistics (ONS). The process is more economical than a full census because it does not involve labour intensive and costly surveys, and therefore can be repeated frequently. However, the approach does not rule out the use of smaller scale surveys where this would supplement data derived from administrative data or other sources. The end product is not identical to the census, but it produces core demographic data by individual and household that in practical terms can be linked to a wide range of other administrative data.

By working at a household level, the flexible and granular output obtained provides greatly improved local planning intelligence (e.g. flexible spatial units, household demography and type of household). However, in the absence of consistent unique personal identifiers in the UK, data matching techniques are required, both for names and addresses. We find that 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. Individual local authorities could use these techniques to provide a population count to be fed into a national system. However, certain procedures would need to be put in place to cover the whole country. We will describe how commonly available administrative data sets available at local level can be used to count populations for local authority areas. Our findings are split into two papers, both published through this journal.

This first paper focuses on describing the methodology, understanding its merits and the contribution it can make to counting populations more accurately and at lower cost. It considers the nature and the strengths and weaknesses of key locally available administrative data sets and how they may be joined in such a way as to produce a replicable, credible and verifiable data set that is accurate at local level.

The following sections provide further background, describe the data sources and explain the methodology; a worked example using actual data is evaluated and a discussion section at the end briefly considers wider issues of implementation and data access. Key strengths of the present approach lie in the applications which go far beyond what is possible with official population statistics, and which can be performed more quickly, accurately and with fewer resources. The second paper (Harper and Mayhew, 2012b), elsewhere in this journal, provides details and examples of applications using these new data sources and contrasts them with existing sources and uses.

2.2 Background

Concerns about the accuracy of population figures have been prominent in debates about statistics, for example whether national level figures derived through a census of the population are acceptably accurate at a local level (Cook, 2003). It is accepted that for areas in population flux the figures are more problematic and therefore less acceptable at local authority level (House of Commons Treasury Committee, 2008). Increasingly however, local policy makers are demanding an understanding of their populations in a more disaggregated, local context in order to better understand their local needs (Freedman et al., 2008; Keohane, 2008). The 2001 UK Census showed that it had not been possible to capture all addresses where people live and so coverage was incomplete even before postal survey forms were dispatched (the first ever census in which they had been used).

Substantial under-counting was also the result of low response rates to the postal survey, particularly in inner city areas. Well publicised cases of this included the cities of Manchester and Westminster (Bowley, 2003; Statistics Commission, 2004). The consequence of these shortcomings was that imputation techniques were needed to fill assumed population gaps. Although the 2011 Census preparation process has taken steps to overcome the addressing problem, including a dedicated address register and huge input from local authorities to help identify hard to count areas and encourage local community support, it is evident that local authorities continue to be concerned about the possibility of low response rates (Central London Forward, 2010; Pharoah and Hale, 2007). Further specific criticisms of the census are that it is only carried out every 10 years and because the results are not published until 2 years later they are already out-of-date. From a user's perspective, statistical outputs and geography are inflexible and do not align with local needs; the data cannot be linked to other data sets except in crude ways; and inter-census MYE population estimates are widely believed to be unreliable due to intervening population fluxes (House of Commons Treasury Committee, 2008).

Redfern (1986, 2004), Ericksen and Kadane (1986) and Keohane (2008) concur with this analysis and point to the burden on the public and the lack of cost- effectiveness, with a typical census costing around £500 million over a 10-year cycle. According to Redfern the census is no longer appropriate in that people are more mobile with second homes and the concept of the 'usual address' is too fuzzy. Keohane agrees that Britain's population is getting harder to count, due to second homes, inaccessible properties, complex residential structures, and

migration and student populations. The Treasury Committee Inquiry was substantially in agreement with these points concluding that the 2007 Census test had shown that even well tried methods will be stretched to the limit by the nature of contemporary society (House of Commons Treasury Committee, 2008). Redfern (2004) proclaims that estimates of the national population need substantial revision and that a new census strategy is required. In particular, he sees the creation of a population register over a period of years as 'probably the only chance to return to quality population statistics' (p.222).

Replacing or enhancing the census of population with administrative data is one suggestion (House of Commons Treasury Committee, 2008 p41), whilst running an administrative data check in a sample of areas in parallel to the 2011 Census is another (Martin, 2006). ONS's position on the use of administrative data has varied over the last 10 years. In 2003, ONS recognised the need for change and improvement. This was envisaged as an 'Integrated Population Statistics System' (Office for National Statistics 2003a) that would combine census, survey and administrative data together into a person- level population statistics database to provide superior population counts, annual estimates and 'Neighbourhood Statistics' to replace the 2011 Census and beyond. This would build upon work already underway to develop a high-quality address register, and be combined with a population register that included administrative data linkage. Since then, they have back-tracked from this position in favour of a traditional census in 2011, with no population register in sight. The use of administrative data would be primarily to improve migration data for the MYEs (Office for National Statistics 2009) and for the Census Coverage Survey. No parallel use of administrative data to the 2011 Census has been confirmed or a decision on how the traditional method will be replaced. The 'Beyond 2011' programme however is intended to assess the integration of existing and new data sources (Office for National Statistics, 2010a) to meet the new demands of population statistics.

The use of administrative data is not new. It has been experimented with since the late 1960s in the USA (Burghardt and Geraci, 1980) and exemplified in existing population registers of the Nordic countries. A population register relies on administrative records as the primary source of census type statistics. This method was pioneered in Denmark in 1981 and utilises administrative data already held in the public sector and combines them by personal identification numbers for the census (Redfern, 1986; see Finnish example in Myrskyla, 1991) and others in Poulsen, 1999. A population register may be limited in scope to how many people are resident in a country alongside basic demographic information such as age and sex, or it may be extended into a full 'census' in the sense that it also records more detailed socio-

economic circumstances. For example, the Dutch Population Register has been available electronically since 1995 (de Bruin et al., 2004) and was used to carry out their full 2001 Census using this and other administrative data sets and surveys, reducing the cost from 300 million Euros to 3 million Euros (Nordholt, 2005). There are also other administrative spin offs; these include less administrative burden on the citizen, increased tax yields and reductions in the over- payment of benefits (e.g. see Redfern 1990; de Bruin et al. 2004).

Clearly, a population register is most effective where there are central files that contain the same consistent personal identifiers, where there is a supportive legislative framework, and where citizens notify the authorities of any changes.

Unlike Scandinavian countries, the UK does not have the benefit of a single personal identification number that is fully universal (Redfern, 1990). Because it covers all ages, the NHS number[5] is the closest the UK comes to this and would be undeniably useful but only if it can be accessed for statistical purposes. While much data are available in government departments that could be used as a basis for a national count, there has been relatively little progress in accessing these data, although following the Statistics and Registration Service Act of 2008, this situation has begun to improve by allowing removal of many legal barriers to data sharing between public authorities and the UK Statistics Authority for statistical purposes.

In our methodology, we use only local readily available administrative sources whose use for statistical and research purposes has been agreed under the Data Protection Act of 1998 and sanctioned by local data owners. These data sets are in use at a local level for a variety of purposes such as tax collection and registration and are part of a national system that is replicated in all local authorities. Of course, it would be even more preferable if data sets such as those held in different government departments were also to be made more available. In line with its desire to make government more transparent in future, the Coalition Government's programme states that, 'Setting government data free will bring significant economic benefits by enabling businesses and non-profit organisations to build innovative applications and websites' (HM Government, 2010). However, whether the data that are released would be suitable for population estimation purposes is unclear at this stage, since much depends on the level of detail that they are prepared to release.

[5] The NHS or The National Health Service number is assigned at birth or when a person registers for the first time with a doctor (for example a foreign migrant).

2.3 Data sources

Whilst administrative data sets and registers at the household level may be a viable source for capturing the population, the data need to be linked and analysed systematically before they can be used for statistical purposes. Local authorities and health trusts hold a wealth of such data on their local populations that can have added value by linking them together and using them in this way. Typical universally available data sets at a local level in the UK are listed in Table 2.1. These should be considered the basic minimum but the list could be extended to include others especially those relating to special populations (e.g. students, armed forces, prisons, and people in institutions).

Data set	Source	Purpose
GP Register	Primary Care Trust (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	Primary Care Trust (PCT)	Public health birth records provided by ONS to PCTs at address level
Deaths	Primary Care Trust (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 2.1: Features of available local administrative data sets

In the absence of one single comprehensive register that captures the entire local population, combining these different sources is essential to maximise coverage. However, each data set has strengths and weaknesses. Combining them becomes a key part of the process in order to remove people that have moved away, are duplicates, or have died. It is hence extremely important to understand the basis for information held in administrative data sets before administrative data can be used successfully. The GP Register, for example, is the most comprehensive of these data sets because it records the majority of a population and contains

age and gender information. Its compilation is illustrative of the detailed considerations that need to be factored in when using it for population counting.

The General Practice (GP) Register is based on the right of everyone living in the UK to register with a GP based solely on residency and not citizenship or payment of taxes. However, patients must only be registered with one practice at any one time and generally need to reside in the UK for more than three months. However, there are several issues to be considered before the GP Register can be used successfully for population counting. For example, a patient is expected to notify a GP of a change of address, but since there are lags in the system of re-registering upon moving to a new area, some records may contain the wrong address for a patient for a period. The net effect of this phenomenon is sometimes called list inflation (or deflation), i.e. when people who have moved (or have died) are not removed (for further amplification of the GP register see discussion section later).

Further considerations apply to other administrative data sets in the list. So, for example, the locally available school pupil census does not cover independent or private schools or pupils that are educated in neighbouring boroughs (unless local authority neighbours have data sharing arrangements); the electoral register only includes registered voters and only the edited version is publically available; the Council Tax Register is based on a single named person per taxable unit and not necessarily reflecting a whole or single household; benefits data contains only people eligible to receive benefits and so on. In addition, data sets such as the school census and electoral register are compiled at regular intervals whereas others such as Council Tax are updated daily.

Births and deaths data are different and these are supplied through the ONS via the local primary care trust. These contain information on all registered births and deaths in an area and can be used to verify whether a person on any of the other data sets has died or whether births have occurred that have not yet appeared on the GP register. The Local Land and Property Gazetteer (LLPG[6]) serves a different purpose to the other data sets. Its purpose is to provide a base set of addresses to which people can be assigned and provide standardised address formats and labels known as UPRNs (Unique Property Reference Number). These are

[6] A LLPG forms a central or corporate address list that provides a unique and unambiguous identifier for each entry in the gazetteer. This central address list will be made up from key Creating Authority service areas responsible for the official street naming and numbering and revenue collection processes. Additional Address Change Intelligence (ACI) is also introduced from other Local Authority statutory functions such as building control, planning and land charges which affect the real-world objects included in the gazetteer (www.nlpg.org.uk).

the common denominator which we use to link data sets together via the address as the core unit of analysis.

There are other address registers available but the LLPG is the most convenient for local authority users because it is created and updated internally and is freely available to them. It also contains other useful information such as when a property was registered and the use of the property (e.g. residential or commercial). Differences between address sources are well documented (Office for National Statistics Geography, 2007) and no one source is able to capture all properties. A 'super' address register using available sources is being constructed for use by the ONS in the 2011 Census, but we understand it will not be made available to local authorities, who will continue to rely on their LLPGs[7].

2.4 Methodology

In comparing information held on different administrative data sets, it is necessary to conceptualise how the information may be categorised. For example, a person may be on one data set and not on another; a person may have a valid address that can be identified on the LLPG or the address may be invalid (the road or house number does not exist) or only partial (a house number may be missing). A person may not be on any of the data sets and is therefore 'invisible' for enumeration purposes. Figure 2.1 is a Venn diagram representing each possible circumstance a record may fall into based on the combination of the three main administrative data sources. In our methodology, we aim to confirm as many people as possible who are current at an address; by definition 'invisibles' are uncountable and so it follows that the more data sets that can be used the better the chance of enumeration in this regard.

[7] It has been recently announced that the Office of Fair Trading (OFT) has given the green light to plans unveiled by Eric Pickles MP, Secretary of State for Communities and Local Government in December 2010, to create a definitive national address database for England and Wales. This will bring together addressing information from local government and Ordnance Survey. See www.nationaladdressgazetteer.co.uk.

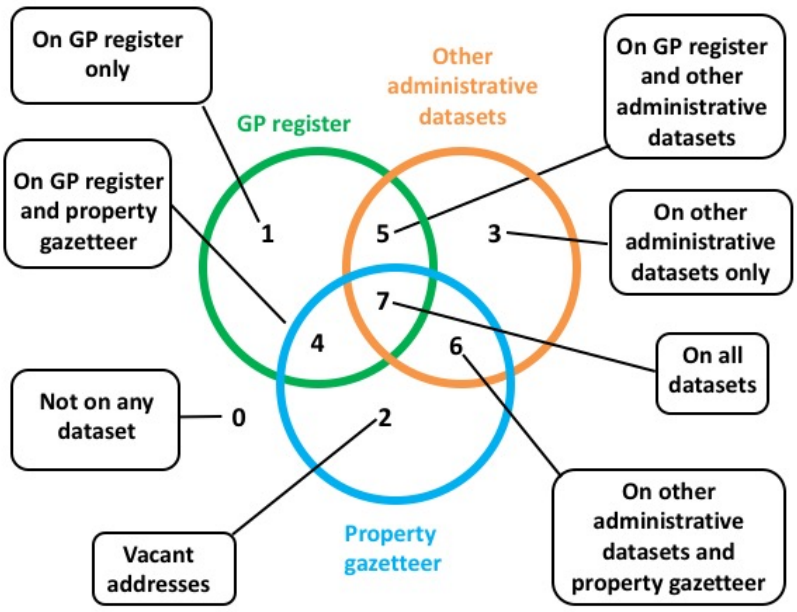

Figure 2.1: Simple Venn diagram partitioning different categories of administrative data with and without addresses

In combining the data sets in Table 2.1, we need 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 data sets are incorporated to create a single final population count and database. Truth tables employ Boolean algebra which can be implemented in freely available software to test whether a logical expression is true or false for all legitimate input values (e.g. Lipschutz, 1998, Chapter 10). 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 data sets.

Prerequisites are that the datasets are all current at the same snapshot in time, that there are no duplicate people on the same data set, and that every address is represented by 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. In summary, the methodology address matches each data set, takes the GP Register as the base, then cross-references the data sets 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.

The logical connectives used in the logical expressions are as follows:

^ and

v Or

¬ Not

→ if-then

Table 2.2 is an example of the simplest kind of truth-table based on the elements in Figure 2.1. In Boolean terms, the combination of factors a and b and c in the logical expression (a v b) ^ c can be represented in a truth table as in Table 2.2 in which '1' represents the condition that a person appears on a, b or c and 0 that a person does not; a for example, might represent the GP register, b other data sets and c the LLPG. A person can be in any one of the seven categories shown in Table 2.2 and represented in the Venn diagram (the eighth category, row zero, is the 'invisible' category). A person is either accepted ('A') or rejected ('R') based on this simple example.

Venn element	*a*	*b*	*c*	decision	comment
0	0	0	0	R	not on any data set
1	1	0	0	R	on the GP register only
2	0	0	1	R	empty property
3	0	1	0	R	on other data set only
4	1	0	1	A	on GP and address register
5	1	1	0	R	on GP register and other data set
6	0	1	1	A	on other data set and on address register
7	1	1	1	A	on GP register and other data set and address register

Table 2.2: Example of a simple truth-table based on Figure 2.1. Key: A accept; R reject

The rules used in the actual methodology are more involved and are applied in a series of stages with the outputs from one stage carrying forward to the next (see Figure 2.2). Brief summaries of each rule are given in the boxes, together with the accompanying Boolean notational form. These rules are designed to ensure that any person identified at an address is current and can be verified, that duplicate persons are eliminated, and as many addresses as possible are filled with confirmed people. Each variable is defined in the column to the right of Figure 2.2, so for example r, 'assigned UPRN', means that a person has been identified as having a valid address.

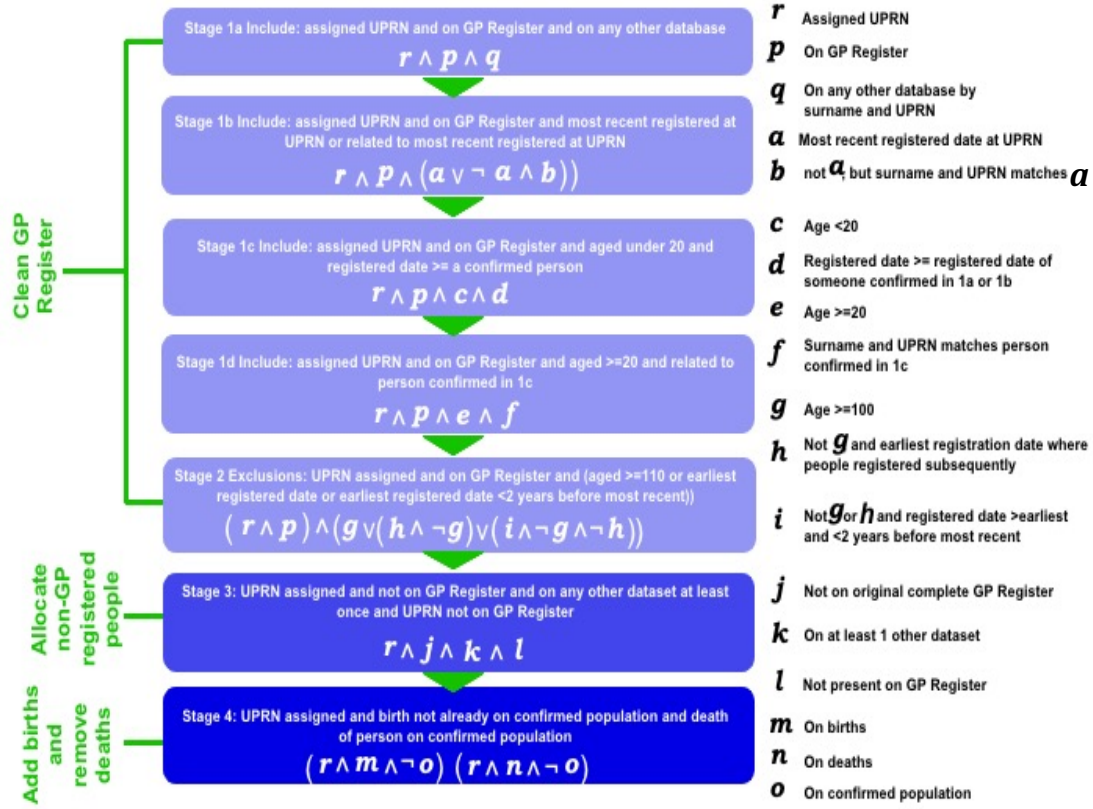

Figure 2.2: Summary of population count methodology stages

The first stage is to 'clean' the GP Register, that is, to determine who on the GP Register can be classified as current residents at UPRNs and so can be included. The rules take account of whether a person is the latest at a given address or if not, if a person is related to someone by name to someone that is current; the cut off for children and young adults is taken to be 20 (i.e. up to age 19). 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 is designed to fill in any gaps in the population not covered by unused records. The fourth and final stage is a last check aimed at filling in gaps that the other data sets have not been able to fill and to remove people who have died but have not yet been removed from other data bases. The end result is a data set, which we define as 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 and therefore 'confirmed' is summarised in Figure 2.3.

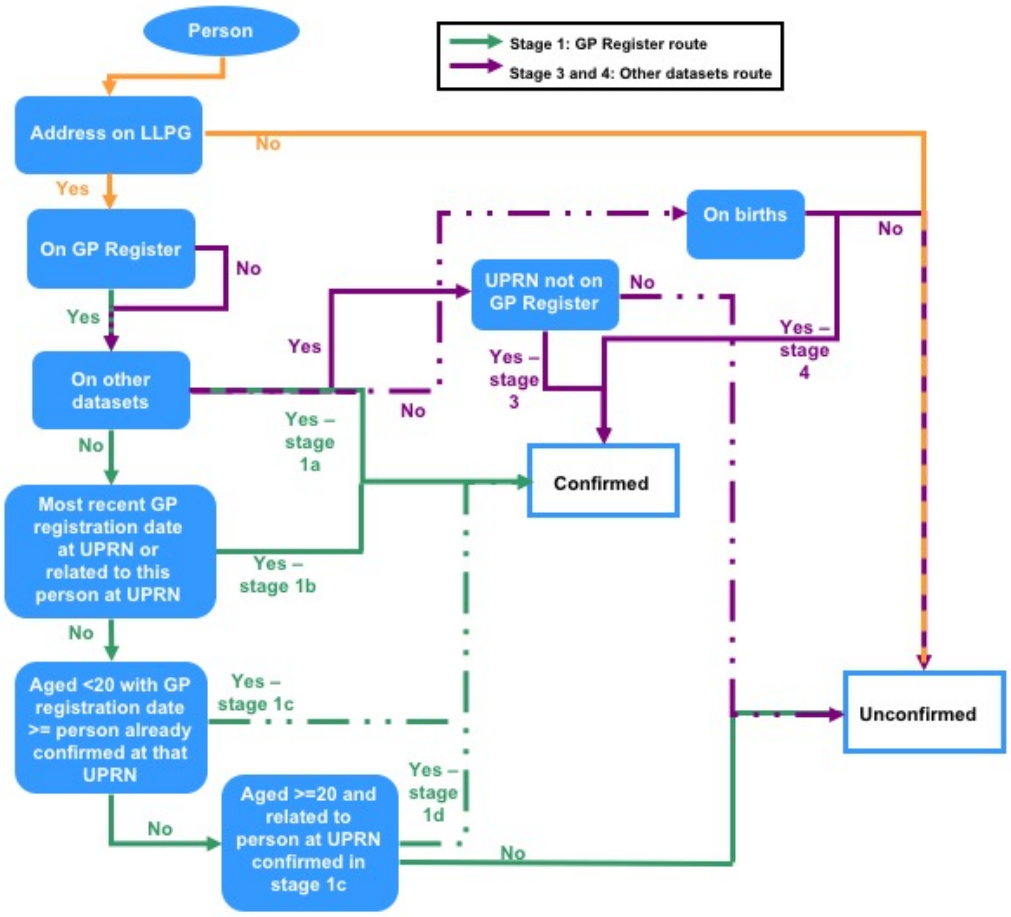

Figure 2.3: Pathway to determine if a person is a current resident at a UPRN or not

2.5 Residuals

Residuals are defined as records that have not been able to be included or verified. They are an important indicator of the completeness of the methodology, and are represented in the simple example in Table 2.2 in rows labeled 'R' (rejected). Each circle in Figure 2.1 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 data sets. Categories 4, 6, 7 are part of the confirmed population if they meet the stated criteria, i.e. they are labeled 'A'. Categories 1, 2, 3 and 5 are not part of the confirmed population and are instead treated as residuals.

The number of residuals tends to rise with the number of data sets used and so is not of itself a measure of matching success, but is more an insight into the compilation of the individual data sets. Residuals consist of data set 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 data sets for any of these aforementioned people, because people are liable to be present on more than one data set. The main sources

of residuals are records which cannot be assigned a UPRN. Therefore, techniques designed to decrease the number of residuals through the correct assignment of addresses are required. Residuals are not immediately discarded but can be evaluated to examine why they have been created and strategies developed for dealing with them. Note that those who are homeless but on a data register recorded as living at 'no fixed abode' or at e.g. their local GP surgery, are considered residuals because they cannot be assigned a UPRN. However, they can be separated out and quantified if necessary.

Figure 2.4 is a flow diagram summarising the residuals and possible changes to how they are handled. Colour shaded boxes refer to the corresponding Venn category in Fig. 2.1. Boxes in black summarise what actions could be taken to reduce or include the residual records. For example, where a person is not included because they are not recorded on the existing input datasets, the suggested revision is to access other datasets that such a person may be recorded on. Residual sources are grouped together at the end to form a possible population 'extension' to indicate the range of uncertainty in any count.

The total number of residuals is the theoretical 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. In practice, many could be duplicates of other records that have been confirmed but could not be matched due to spelling or other differences. It is for these reasons that the final result is called the 'minimum' confirmed population, but the theoretical maximum will always be uncertain due to reasons that can frequently be traced to quality issues within the source data.

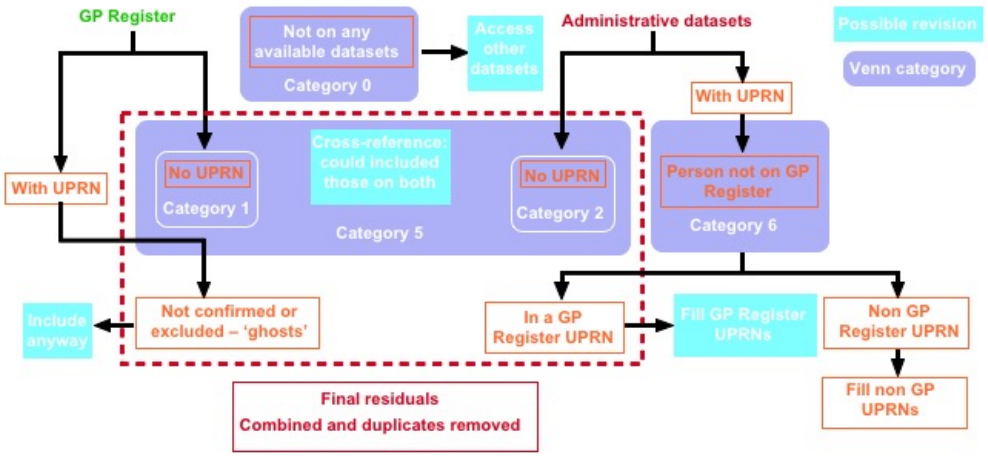

Figure 2.4: Residuals and possible remedial actions

2.6 Evaluation of results

In testing the accuracy of any administrative count, it is important to recognise that there is no single gold standard 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. The best sources, if possible to obtain, are often those which involve financial transactions or transfers of one kind or another (e.g. benefit or pension payments) since these are arguably more likely to be accurate. In addition, accuracy also needs to be considered in relation to why a population count is needed. For example, is it to assess the need for public transport or the number of state school places? The relevant population could be very different in each case.

Obviously, sources should be contemporaneous with the administrative snapshot where possible, although sometimes there may be a lag. Also, administrative sources may be subject to changes of definition or eligibility as in the recent case of Child Benefit which was universal to the age of 16 but the Government is now intending to withdraw it from households with a higher rate tax payer. One can also use ONS MYEs or their equivalent such as Greater London Authority (GLA) estimates, although clearly there is a danger of circularity here since the purpose of an administrative count is to replace counts by other methods. However, their use for such purposes seems unavoidable until and unless they are replaced.

In practice, there are relatively few readily available administrative or other comparators, none of which is perfect and all of which are partial in coverage. Examples include:

- Child Benefit numbers published by HM Revenue and Customs 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 9 people, are identified and checked for being multiple-occupancy
- Comparison with other sources from contemporaneous snapshots e.g. ONS MYEs or GLA figures, if the local authority is situated for example in the London area
- Number of children aged <16 without an adult at a UPRN is checked for possible explanations (e.g. parent or guardian is not on the GP register).

The question arises as to whether it is possible to create measures of confidence in estimates based on this approach using standard statistical methods and assumptions. In this regard, different approaches can be envisaged. It is well known for example that the veracity of

individual data sets varies both in completeness and coverage as well as accuracy, often in unknown ways. Sensitivity analysis can be undertaken by relaxing or varying certain assumptions in the methodology or by systematically adding or removing data sets; however, the approach which we find makes most practical sense is to split up the population into groups with strict rules of association and assigning labels such as 'confirmed' or 'probable'.

Small surveys can then be undertaken to assign probabilities to a sample of members in each group to establish whether they should be included or not, with a given level of statistical confidence; in theory, these could piggy-back on other routine surveys, for example housing or health and life style surveys and we have some experience of this. Although we have not designed and conducted such a survey ourselves, we are aware of at least one occasion of where our data was used by local emergency services to check on people living in streets that had been severely impacted by a small localised tornado. Although hardly a model on which to build, the feedback we received was that the data were the most accurate they had ever seen!

2.7 Matching algorithms

Thus far, we have said little about the data matching process itself which comprise the techniques needed to link people to addresses and between data sets. In an ideal world, each record on every data set would have one or more unique identifiers and so matching would be straight forward, e.g. a person identifier such as a national insurance number, NHS number, and a UPRN. In practice, the GP register is the only data set to have a unique person identifier in the form of the NHS number. The Local Property Gazetteer has UPRNs for each address and the School Pupil Census a UPN or Unique Pupil Number, but this covers only a narrow age range.

With the cancellation of the planned national identity card system, it is unclear whether there will ever be a universal basis for uniquely identifying individuals or a citizen's index that could be used as a basis for a population register. Councils typically match council tax information to the UPRN, but matching records to UPRNs is still not common practice across other data sets. This means that we must resort to other methods of matching people either to addresses or to each other until other solutions are found. Since data sets may comprise many thousands of records, it is important that the matching process should be automated as far as possible, but also that the processes should also be accurate.

Data sets are variable in their quality and standards of completion. With addresses, the same address can be captured in varying ways either through data entry mistakes, misspelling or the existence of aliases. With individuals, sources of error are variations in spellings, data coding and preparation, use of name synonyms and nicknames, Anglicisation of foreign names, double-barrelled names, cultures that commonly incorporate the same title in the name, e.g. Singh or Kaur, use of initials, truncation and abbreviation, forename and surname swapped round, missing words and extra words (Gill, 2001). Dates of birth may not be reliable either; the day or month may be substituted with a default value if it is not known, or have a character entered incorrectly.

A crucial consideration is that different data sets may be collected for different purposes, and so were not designed for easy, accurate matching. Matching methods therefore need to reflect this and algorithms must recognise common differences and formats. While these algorithms are suited for matching local administrative data, the processes can become very technical and there is a substantial literature on record linkage that goes into more detail (Ericksen and Kadane, 1986; Winkler, 2011; Gill, 2001; de Bruin et al., 2004; Jenkins et al., 2008; Office for National Statistics, 2010b). In our approach the two main categories used in record linking are address matching and person matching, as described below.

2.7.1 Address matching

For the purpose of the population count, 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 data sets, the addresses are standardised by finding each address in the available property gazetteer and representing each with its unique property reference number (UPRN) on the database. A purpose-built address matching algorithm has been designed to do this.

Unavoidably, a small percentage of addresses will remain that cannot be matched in this way. These tend to be formatted so differently from the gazetteer version that they need to be processed manually to choose the correct match. This is facilitated in our methodology by a semi-automatic process with manual over-ride. If after this a UPRN can still not be confidently assigned, the record becomes a 'residual' as defined and discussed in the previous section.

A record is designated a residual due either to the address being outwith the study area, the address is missing a vital discriminatory piece of information, usually the SAON (Secondary

Addressable Object Name), the address contains a SAON that has not yet been recorded in the gazetteer, or the address is too 'noisy' or incomplete to assign a match with any confidence. Figure 2.5 sorts these cases into five categories (0 to 4) and suggests solutions to improve UPRN assignment for each as matching proceeds. For example, for addresses that contain a SAON (usually a flat number) that is not recorded in the property gazetteer, but the PAON (Primary Addressable Object Name, usually the street number and name) does exist, a 'dummy' UPRN will be generated.

2.7.2 Person matching

Person matching is used in the population count to ensure that the same person is matched across multiple data sets, particularly between the GP Register and other data sets. There is no single unique person identifier on the data sets to allow full exact matching, so a technique is employed using the forename, surname and date of birth fields. Gill (2001) and others review the issues in person matching and our methods entail similar considerations; however, we note in passing 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. In particular, names can offer clues to a person's nationality or ethnicity especially when used in combination with a range of administrative data sources. We exploit this property in applications of our methodology (not discussed here).

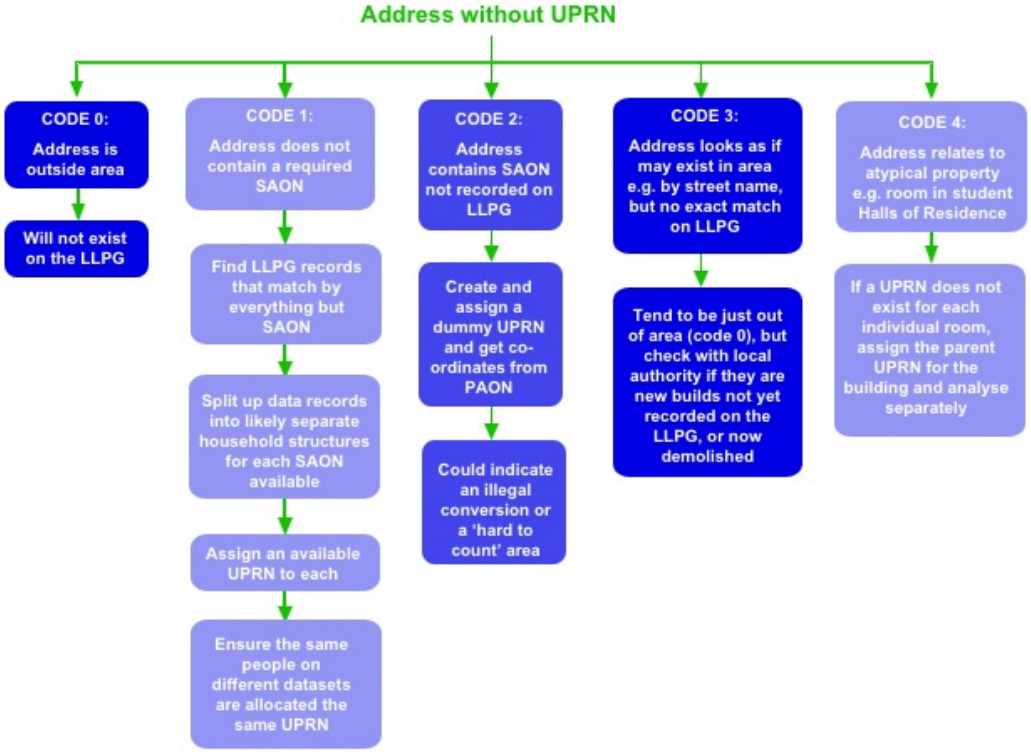

Figure 2.5: Extended UPRN assignment flow chart. Key: SAON = Secondary Address Object, PAON = Primary Address Object

2.8 A worked example

The methodology is now illustrated by means of the following case study which is based on the London Borough of Barking and Dagenham and uses an administrative snapshot date taken at 30th September, 2008. In this case, the UPRN assignment rate to addresses in the data sets was very high at around 98%, and so it was possible to include practically all available data records in the analysis. A summary of the audit trail for this case study is given in Table 2.3 based on each of the stages in Figure 2.2, in which the confirmed additions to the population for each of the four stages are shown and also the numbers of records eliminated. It shows that the final population count obtained was 171,851 people.

For this case study, reasonability checks using data available at the time yielded the following results:

- 44,258 children aged 0-16 were counted, compared to 44,985 on Child Benefit August 2008 (source: HMRC)
- 7,492 males aged 65 and over compared to 7,830 males aged 65 and over claiming state pension as at August 2008 (source: DWP)
- 13,915 females aged 60 and over compared to 14,050 females aged 60 and over claiming state pension as at August 2008 (source: DWP)
- 23,801 single occupancy UPRNs compared to 20,720 on Census 2001
- Vacant UPRN rate = 3.9% compared to 2.8% from Valuation List March 2008 (source: Communities and Local Government)
- 152 UPRNs of the 68,247 allocated UPRNs have > 9 people, covering 1,829 people in total

Stage	Summary	Main comments	Population count
1 and 2 – Clean GP Register	Identify current registered patients at each UPRN to be included	❑ 1,607 GP patient records could not be assigned a UPRN ❑ 59,730 UPRNs have *current* patients to include ❑ 11,269 UPRNs have no *current* GP patients to include ❑ 21,520 GP patients can be excluded	+ 156,764
3 – Identify additional people from other data sets and allocate to as yet unfilled UPRNs	Eliminate people on Council Tax, Benefits, Electoral Register and School Census who are already on GP Register. Then identify which of the remaining 55,562 records are in the 11,269 unfilled UPRNs, and remove duplicates	❑ Eliminated 167,455 duplicate people using person matching across all data sets ❑ Leaves 55,562 records to check ❑ 20,194 records across data sets have 'unfilled' UPRNs ❑ Reduced to 14,496 people after removing duplicates ❑ Leaves 35,368 records to check that do not have a non-GP Register UPRN	+ 14,496
4 – Add births and remove deaths		❑ 2,381 of the 3,005 births are already included ❑ 624 births are additional, 604 with UPRN ❑ Subtract 13 deaths from existing population base*	+ 604 - 13
		Population Base =	171,851
Covers 68,247 UPRNs of a possible 70,999 Leaves 2,752 unallocated UPRNs = 3.9%			

* It is not unusual to add more births than deaths at this stage of the process. In general, we find a greater time lag between when a baby is born and registered with a GP (which is the responsibility of individuals), as compared with a death being registered and being removed from a GP register (which is the responsibility of the coroner system and GP).

Table 2.3: Population count audit trail for a case study

The population count of children 0–16 is less than the 2008 Child Benefit count by only 727. The counts of males aged 65+ and females aged 60+ are 338 and 135 less respectively than state pension counts at August 2008. Hence, these two comparators suggest that the administrative count may slightly understate the population in these two age bands, assuming that the pension and benefit counts to be accurate and contemporaneous. The number of single occupancy households is higher than the Census 2001 count, but it is not implausibly different given the timing differences between snapshots. The vacant UPRN rate of 3.9% is 1.1% higher than the 2.8% given for March 2008 for the number of vacant dwellings and second homes as a percentage of total number of dwellings on the Valuation List. However,

this difference can be explained by timing and definitional differences, for example when records are added after a property is built differ on the LLPG and the Valuation List.

It is assumed that any UPRN with more than nine people in residence is potentially unusual and could indicate an error. Only 152 or 0.2% of the allocated UPRNs are affected by this, and all were checked for possible explanations. Approximately 40 of the people affected are in UPRNs known to be hostels and a further 319 in addresses that are obviously care homes. The highest occupancies of any UPRN, 28 to 61, are in these properties. The remaining cases are distributed across normal residential addresses with occupancy predominantly in the lower ranges of 10 to 15 (see Figure 2.6). This very small number and the fact that many are genuinely multiple occupancy properties again indicate that the results are capturing legitimate household structures. This could be further refined and validated by obtaining the maximum capacities of known multiple occupancy addresses (e.g. hostels).

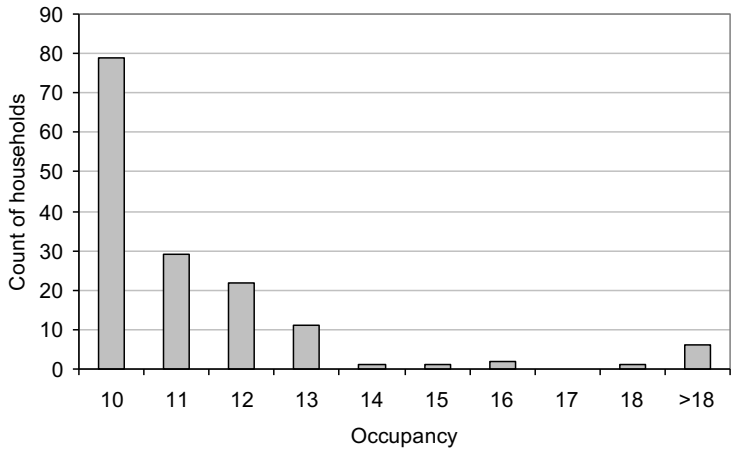

Figure 2.6: Distribution of high UPRN occupancy levels resulting from the case study

Numerous other checks are possible, including for example the number of households in which there are children but no adults. Few in number, these cases can arise where the child occurs on a database but not the parent or guardian, e.g. an adult who is unregistered with a GP or is not the person responsible for paying council tax, etc. Based on the experience of other case studies, such checks provide confidence that the results are reasonable; however, it is always useful to consult local authority experts and analysts for further verification (e.g. in cases of recently demolished areas). Further comparisons may also be undertaken with alternative sources of population estimates, although clearly there is danger of circularity—i.e. using external estimates to verify an administrative count which is in turn is being used to validate an external estimate.

The external estimates available are the ONS MYEs or GLA figures, if the authority is situated in the London area. It is possible to envisage a number of different checks against these sources, for example comparison by age band, or at sub-authority level, such as ward or Super Output Area level (note that a comparison at a household level is not an option using GLA or ONS sources). We illustrate our findings with a comparison by 5-year age band as shown in Table 2.4. In constructing the age bands using administrative data, it is necessary to take into account a relatively small number of confirmed records for which there is no date of birth, no gender, or both. Since it is possible to establish that many of the 'age-unknowns' fall into the adult age range, it is relatively straightforward to devise an arguably reasonable distribution of these among the relevant age groups to correct for this.

As Table 2.4 shows, the administrative population count at 30th September 2008 is higher than the original ONS MYE 2008 count of 168,853 by 2,998 persons. In May 2010, the ONS revised its MYEs for 2002 to 2008 to reflect improvements to methods and data sources on migration. The revised 2008 figures, only published in rounded form, have been included in column four of Table 2.4. Interestingly, the new count comes to 171,600, which is now only 251 less than the administrative count. However, it is worth drawing attention to the fact that the administrative count was produced and disseminated within 3-months of the snapshot date, as compared with the ONS revised count which took 2 years longer to produce an almost identical total figure.

Age group	Administrative population at 30/9/2008	ONS* 2008 MYE (old)	ONS** 2008 MYE (revised)	GLA*** 2008 (revised)
0-4	15,059	15,735	15,800	15,742
5-9	12,438	11,554	11,600	11,465
10-14	11,993	11,879	11,900	11,382
15-19	11,276	11,380	11,500	11,472
20-24	13,078	12,255	12,700	10,152
25-29	12,614	12,861	13,800	12,835
30-34	12,204	12,192	12,700	13,934
35-39	14,007	13,067	13,300	13,790
40-44	13,698	13,470	13,600	13,460
45-49	10,827	11,081	11,200	11,529
50-54	8,433	8,749	8,800	9,247
55-59	8,129	7,553	7,600	8,099
60-64	6,658	6,767	6,800	7,329
65-69	5,029	4,878	4,900	5,255
70-74	4,702	4,503	4,500	4,746
75-79	4,707	4,281	4,300	4,473
80-84	3,685	3,418	3,400	3,694
85+	3,316	3,230	3,200	3,371
Total	171,851	168,853	171,600	171,976

* Source: Office for National Statistics © Crown Copyright 2009 (experimental statistics)
** Source: Office for National Statistics © Crown Copyright 2010 (experimental statistics)
*** Source: GLA 2010

Table 2.4: Comparison of case study population age breakdown from different sources

The GLA publishes population projections for London boroughs. Unlike ONS it uses housing units in its methodology, taking into account expected future housing development in an area (Hollis and Chamberlain, 2009). The GLA 2008 low and high variants give counts of 167,475 and 172,400 respectively for Barking and Dagenham, with the higher variant designed to cope with higher anticipated migration assumptions. As is seen, the administrative count is within these margins, but closer to the higher variant. The same was true when we compared the administrative count with GLA 2009 estimates, namely that the administrative count lay between the low and high variants. The GLA's revised 2008 figure of 171,976, shown in column five in Table 2.4, is only 125 higher than the administrative count, but again took 2 years to be published. There are both similarities and differences between the counts for separate age bands for each source. The administrative count is lower than ONS for ages 0 to 4, although it is not completely clear why this should be so since both GP and birth registrations are considered reliable sources. Higher administrative counts are found in the 5–9, 20–25, 35–39 and 55–59 age groups and we have generally found this to be the case in other areas we have used this methodology, especially in London (e.g. see Mayhew and Harper, 2010b). Reasons

for this are necessarily speculative to a degree and are probably methodological in origin rather than just timing differences. For example, other sources include a baseline based on the 2001 Census and thus are possibly distorted by low response rates and imperfect imputation at the time, and secondly, failing to account properly for migration[8].

Figure 2.7 is a chart summarising the differences between the administrative count and the three other 2008 sources by 5-year age band. In general, the administrative count is relatively higher in age bands up to 25, lower between 25 and 35 than either ONS or GLA; but at older ages the differences tend to be narrower. Any estimates in the age range 20 to 40 from whatever source must be considered less robust than in other age bands because this population tends to be hardest to count. Since the administrative data approach uses current data sources in general, it is arguably a more accurate reflection of the population dependent on or using local and other services. However, each methodology is clearly different, and so has to be taken on its own merits.

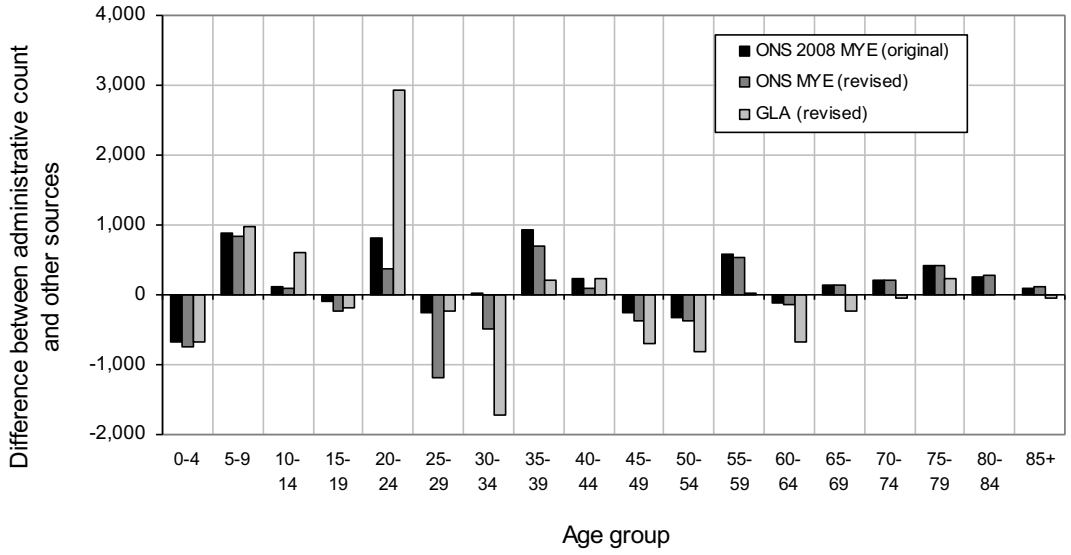

Figure 2.7: Chart showing the differences in estimates by age group between the administrative count and ONS and GLA

The above comparisons demonstrate that each source is relatively close to each other with differences of less than 2% at the aggregate level, although the earlier availability of the administrative count makes it much more attractive from a user perspective. Larger differences became apparent when comparisons are made at ward level. We found that, based on all 17 wards in the case study, the percentage difference between the administrative count and ONS ranged from −12.9% to +8.2% with a root mean square deviation of 547

[8] Undercounts in the MYEs have led them to be declared 'unfit for purpose' (House of Commons Treasury Committee 2008, p3) for many areas.

persons (average ward population is around 10,000). The same comparison using GLA 2008 (revised) figures at ward level gave slightly more extreme results, with percentage differences ranging from −17.9% to +8.1% and a root means square deviation of 621 persons.

Based on the 109 Lower Super Output Areas (LSOAs), the percentage differences between the administrative count and ONS were considerably higher, ranging from −37.7% to +15.2% with a root mean square deviation of 138 persons (the average LSOA population in this local authority is around 1,600). Clearly, these results are based on one London borough and may not be generalisable; however, they suggest that even if population figures at local authority level are comparable from the three sources, the gaps at more disaggregate geographies are greater and potentially much more of a problem, depending on the type of intended application (see Harper and Mayhew, 2012b for more discussion of this point).

In reaching these conclusions, it has been necessary to discard those administrative records that did not conform to the methodology. Table 2.5 contains a brief enumeration of the rejected categories (rows 1, 2, 3, and 5) for the case study as defined and set out in Figure 2.1 and Table 2.2. In general, we observe that the quantity of rejects is reassuringly small in relation to the confirmed population count, but as previously noted their number tends to rise with the number of data sets being used. In this regard, every case tends to be different and so it is not easy to draw general conclusions as it depends on the quality and number of data sets.

The question arises as to which count is the most reliable. Since the administrative methodology relies on current actual data rather than synthetically adjusted counts from a census base that is over 10 years old, it is arguably more likely to be accurate. It is based on the current dwelling stock and households as well as current data that have been systematically validated and combined. In broad terms, administrative counts are better at capturing recent arrivals in an area and so tend to be higher in areas where there is greater population turnover. Is it always the case that the administrative count will be close to conventional estimates?

It may be argued that this particular London borough is more straightforward than others in the sense of not having a particularly complex population and thus is unable to provide a strong enough test for the methodology. A much tougher challenge was the London Borough of Tower Hamlets, also in east London. This has a large student population, is undergoing massive re-generation, and has many second homes among the many new developments. These factors contributed to Tower Hamlets having the highest property vacancy rate we have

observed so far in any location at 7%. In addition, and partly as a result of these factors, we also found that 13% of the confirmed population was not registered with a GP, but are people that were identified from other data sets. On this basis, we found that Tower Hamlets had an administrative population count that was 6.5% higher than the comparable ONS MYE as compared with only 1.8% in Barking and Dagenham.

Reject category	Definition	Comment	Case Study Quantity
1	Population on GP register without a UPRN and not on other data sets	Caused by poor addressing or when records are for patients living outside the local authority area	0.9% of GP Register data set
2	UPRNs without any confirmed current residents	Useful as check on reasonableness of population count where it can be checked against independent evidence;	5.7% of LLPG
3	Population on other data sets without a UPRN and not on GP Register	Caused by poor addressing or when records are for patients living outside local authority area	1.4% of other data sets
5	Population who are recorded on both the GP Register and other data sets without a UPRN	Caused by poor addressing or when records are for patients living outside local authority area	Potentially 59 records in total

Table 2.5: Enumeration of rejected records for case study

2.9 Conclusions

This paper has made the case for utilising and linking local administrative data to count local populations. The method is current, has a turn-around of up to 3 months from the time the data are obtained, and can be carried out as frequently as desired. It 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 has tried to take this further and demonstrates innovatively how the problems associated with the onus being on the citizen to self-report and self-return a census survey can be bypassed. It represents a contribution to the debate of what should replace or improve the UK national census after 2011, but also addresses the strategic gap in good population intelligence at local level, which is stifling planning and stewardship of the considerable

resources that are allocated centrally through grants to finance local services. Since we believe it will be some years before there is a more credible national system for counting, we consider that there is a strong business case for this methodology to fill the gap but acknowledge that it is also capable of further refinement and development.

Although the case study gave an administrative count that is similar to other estimates at a local authority level, this has not necessarily been the case in other local authorities and the example of Tower Hamlets was mentioned. Generally, we find that in London the differences between the administrative population count and official counts have been greater than in areas that are in less flux, even though in all cases the data sets used and methodology were the same. Nevertheless, it will always be difficult for any system to capture 100% of a population, because it depends in part on how a 'population' is defined.

More transient populations such as tourists and short-term (e.g. <3 month stays) migrants could theoretically be included with access to appropriate data; similarly, data can be appended for those serving in the armed forces and prison populations or living in institutions. A more sophisticated set of population accounts would subdivide a population into, for example: the usual resident population (i.e. whose main home is in the area), the day-time only population, with further subdivisions based on length of stay to distinguish short term visitors from migrants. However, to do this rigorously might require a politically unpopular system of population registration to underpin it.

One important sub-group is the student population because it inflates local populations in term time and deflates them out of term time. We take current residents as at a snapshot date so that if students are on databases at this date, they are included, but we would only be able to identify them as students if they lived in designated halls of residence. Access to HESA (Higher Education Statistics Agency) data would provide domicile and study addresses, which would improve identification and separate enumeration of students. In future, we support the idea that published figures will need to differentiate between a term time and out of term population for an area and look forward to working with HESA to provide the necessary data.

The paper has explained the crucial role of the GP register for population estimation purposes but it is not a panacea and a would-be user of the GP register needs to contend with the following issues. Comparison between the GP register and official population data sources for different ages generally show that there are more people on the GP register than in official population figures (especially in urban areas). However, for people in their 20s, particularly

young adult males, there can be fewer because they have not bothered to register. Foreign nationals such as diplomats or others who exclusively use private healthcare may also be absent from the GP Register, although the numbers involved are small and tend to be localised (e.g. in London boroughs such as Kensington and Chelsea).

The reverse is that there are people on other data sets (e.g. young male adults) that are not on the GP register but can be confirmed through other sources. Our methodology enumerates these, but it cannot identify people who are not on any of the common data sets (e.g. illegal immigrants). An easily overlooked group that are alive and living in an area but may not yet be registered with a GP are newborns. Several hundred may be involved, which is why we use the public health births register to fill the gap. Similarly, people may not be removed from a register if they have died, but generally we find this to be much less of an issue (see also footnote (a) Table 2.3). We have already mentioned that in areas of high turnover and influxes such as Tower Hamlets, a relatively large percentage of the population is not registered but confirmed using other sources.

In theory, any additional data set could potentially improve population counts within the framework of our methodology, including some commercial data sets. Each data set needs to be included on its merits (e.g. the range of information captured such as date of birth and current address, population size and geographical coverage). These criteria would rule out many commercial data sets, but some such as loyalty card customer data may capture some people not on public data sets (e.g. new arrivals from abroad). The most useful data would therefore be sources that had the potential to fill gaps and were known to be of high quality; however, the most important barrier to obtaining access to such data sets for statistical purposes is their commercial confidentiality.

We have also considered, and to some extent have tested, the use of life style and other surveys. Assuming it is possible to access the addresses of respondents and that the survey is current, it is possible to compare demographic details such as number, age, and sex against corresponding administrative data. To date, however, we have found such surveys to be more useful as a means to extend the range of socio-economic variables in the output database to include, for example, attitudinal variables rather than for counting people as such. In practice, this entails imputing the characteristics for other similar households based on respondents to a limited survey of perhaps only a few thousand households. However, such uses raise methodological issues that go beyond the scope of this paper.

There are several more strategic issues to consider in terms of the wider adoption of this approach. Implementing the methodology at a national level has not yet been attempted but can be considered as a matter of carrying out population estimations for each of the local authorities in England and Wales[9], and then combining them. This would require consistency in the input datasets used in terms of snapshot date, coverage and quality, and an assumption that the methodology is a 'one size fits all'. The present assumption is that local authorities could do this for themselves, initially with outside technical assistance, but with data improvements and access to the necessary algorithms, the processes could become more automated and enable scale economies; this is something that would be best done in stages involving geographically contiguous authorities to enable more efficient data pooling. We believe this to be more of an administrative issue than a technical one because it goes to the heart of local authority co-operation in the area of shared population intelligence and resources.

It is important to note that the approach uses person-identifiable data in the initial stages, but that the final database is anonymised for statistical use. The use of data here has been approved under the 1998 Data Protection Act, but there remain multiple local interpretations by different data owners over the user of personal data for statistical purposes. This issue would need to be addressed if the aim were national coverage and would require government leadership, more clarity and less dithering. The normal arrangement is to create 'safe havens' that enable personal data to be linked and anonymised and packaged for statistical purposes in wholly non-person-identifiable formats.

If a national model was to be based upon the input of each individual local authority, and as an estimate, if an administrative data population count costs on average £100 k per authority, the total cost for the 348 authorities in England and Wales would be £34.8 m. However, this cost would fall in time following data quality improvements. This compares with the decennial census which costs £500 m over a 10-year cycle. In theory, it would therefore be possible to provide annual counts rather than decennial for the same or less money; however, this view needs to be tested further as there may be unforeseen costs in scaling up our approach (e.g. see Office for National Statistics, 2003b).

[9] We have ascertained that similar data sources are available in Scotland and so the same data sets and methods could be deployed there.

A completely different business model would be to utilise the considerable data resources available to central government, especially those available through the tax and benefit system. Hitherto, that route has not been possible under present legislation. However, this could change, as the Government considers the future of the census (Hope, 2010). The data sharing provisions of the Statistics and Registration Service Act 2008 is a potential model. Such a model would dictate a central rather than local led solution to population estimation, but it would also carry with it significant technical challenges and upfront costs not to mention haggling between departments over data ownership.

In our judgement, it will be several years for this to be achieved if it happens at all. Thus, it seems likely to us that local data sources will continue to play an important role in this regard simply because it allows local authorities to be in control of the data that they need for local planning rather than relying on central government to produce timely accurate data that meets local (as well as national) needs. In conclusion, this paper has demonstrated that administrative data are a viable and cost effective alternative to the current census method of counting populations. This topic continues into a second paper in which we consider how administrative population counts can be used in routine applications and combined with other data sets in potentially innovative and previously unchartered ways.

Acknowledgements This paper is based on research for the ESRC UPTAP (Understanding Population Trends and Processes) programme. We acknowledge contributions from Sam Waples of Mayhew Harper Associates Ltd and thank Richard Verrall of Cass Business School and John Eversley of ppre CIC for their comments and support.

3 Applications of Population Counts Based on Administrative Data at Local Level

Preface: Content in this chapter consists of an exact reproduction of the article published in the Journal of Applied Spatial Analysis and Policy in 2012. Only minor edits have been made to make numbering consistent throughout the thesis. As such there may be some dated references or statements. Developments in population data science since the time of publication of this paper are described in Chapter 6.

3.1 Introduction

On the face of it, why we need to count populations seems a question hardly worth debating. After all, the first UK census was in 1841 and it has continued every 10 years except for 1941[10]. Debate today is much more about how accurate and timely the counts need to be. The answer depends entirely on one's point of view and the purpose for which the information is needed. How many lifeboats should an ocean liner be equipped with needs a precise answer for obvious safety reasons; for retailers, used to uncertainty, population is arguably secondary as compared with market share and profits. For public service organisations, identifying the 'right' population depends on, for example, whether it is people who will contribute or require resources, whether it is a day-time, night–time, temporary or long-term population.

What is clear today is that there is a demand for ever timelier and more detailed population data to satisfy a growing thirst for population intelligence. We have reached a point in which most service organisations know in detail who their customers are and where they live, but are much less sure about how many others there are like them in the wider population and what other services they use or need. However, part of the thrust for more detailed and accurate population statistics is political in origin—for example to strengthen local democracy, encourage joint working across public sector boundaries, and generally encourage greater information sharing. This is partly in the name of greater efficiency and effectiveness but also, ultimately, to serve wider social objectives such as reducing health inequalities, promoting social cohesion and protecting the vulnerable and so forth.

The House of Commons Treasury Committee (2008) was damning in its assessment of the UK system for counting the population, which it described as 'unfit for purpose'. It usefully set out what it considered the three main purposes of population statistics. They were to:

[10] With the exception of 1966 where a trial interim 5-year census was carried out using a short form for every household and a long form for a sample of households

1. Allocate resources based on the distribution of the central Government grants to countries of the United Kingdom, to local authorities, health care providers, police and other services at local level
2. Provide denominators to construct ratios such as the number of crimes committed per head of population, unemployment rates and so on, and hence to evaluate policy at a national level (but also at a local level)
3. Plan, deliver and evaluate services at a local level taking into account need and demand.

This is not to refute that there are many other purposes besides. However, in any of these cases, there can be no doubt that an inaccurate population figure could skew resources, by giving a false picture about an area. In a companion paper in this volume (Harper and Mayhew, 2012a), we have set out an alternative methodology for counting populations using routinely collected administrative data which we define as data not primarily collected for statistical purposes (Vale, 2006). Although the greater use of administrative data has been talked about for years (e.g. see Ericksen and Kadane, 1986; Brackstone, 1987; Steffey and Bradburn, 1994; Penneck, 2007; Keohane, 2008), there has been remarkably little progress in the UK in implementing the necessary changes to statistical systems, or in exploiting the potential of these alternative sources for a combination of reasons. In addition, very little of this debate has percolated down to the local level which arguably is the level of government that stands to benefit most from more accurate and detailed population counts.

In this paper, we show how administrative data collected at a local level can be used to overcome the significant weaknesses under the current arrangements identified by the Treasury Committee. Section 3.2 briefly considers the limitations of presently available population statistics in each of the three core purposes identified; Section 3.3 considers how administrative data could be structured for statistical purposes to overcome these limitations; Section 4 provides a worked example using actual data that covers aspects of the main purposes of population statistics. Our focus is on the local level (local authority or below) although a national perspective is introduced where relevant; for example, it is sometimes necessary to understand the process from a national perspective (e.g. in the case of resource allocation) in order to understand the local ramifications. The wider adoption of our approach at other levels of government is critically appraised in a concluding section which considers what has been achieved and future directions for development.

3.2 Limitations of official population statistics

3.2.1 Resource allocation

Resource allocation may be defined as the process by which the public sector allocates resources to activities and areas based on specified objectives in circumstances in which the process cannot be entrusted to market forces (Barr, 2004). Resource allocation formulae typically need to capture, directly or indirectly, the need or demand for a service per unit of population, in which population is the scaling factor which must be combined with the unit cost of a service or the total quantum of resource to be distributed to territorial units, administrative areas or service delivery organisations. The particular formula used will depend on the population served and the underlying policy objectives which are integral to achieving wider social objectives - for example, to reduce health inequalities, to combat crime or raise educational standards (Marmot, 2010).

In spite of their increasing complexity and the number of other variables built into such formulae, good population estimates are crucial elements of the process for distributing resources fairly or equitably—whether it is school places, police on the beat, or health budgets. The NHS was one of the first public sector organisations to use a formulaic approach to resource allocation based on a weighted population for distributing health care budgets down to regional level and below (Resource Allocation Working Party, 1976; for review of history see Bevan, 2009). However, within this broad canvas of applications and approaches, a general distinction can be drawn between area- based funding models, and those which allocate resources to organisations that deliver specific services such as schools[11].

Area-based funding differs because it allocates block grants to geographically bounded administrative units such as local authorities which in turn use the funding to deliver a range of services. The Local Government Finance Settlement is a good example of the use of population estimates, but also of how other factors such as deprivation and specific local circumstances are taken into account. There is in addition an emerging trend towards funding models which allocate resources direct to delivery organisations (e.g. schools, primary care). Through the need to use more detailed and timelier population data, these processes have progressively exposed weaknesses and anomalies in population statistics. Such data limitations may partly

[11] A further category would be allocations at a household level such as educational vouchers, benefits or budgets for personal care all of which are eligibility based.

explain why formulaic approaches to sub-local authority resource allocation processes are uncommon, with much more regard being given to local judgement and politics.

The problems with population data are essentially of two kinds. The first is its inflexibility; for example, geographical units are often unsuited to the applications that depend on it, and specific variables such as age bands are in fixed formats, making it difficult to identify the demand for services based on non-standard age groups. However, more fundamental is the poor quality of the data itself which can be traced to a combination of low response rates during the previous census in 2001, especially in inner city areas, and subsequent population fluxes through migration (Simpson, 2007; Simpson and Brown, 2008). The consequential undercounting of population has meant that some areas have effectively experienced nearly a decade of underfunding since the last census in 2001 (e.g. see Mitchell et al., 2002; Dorling, 2007; Local Government Association, 2007). Lawrence et al. (2007), working in Brent, a suburb of London, found for example that population undercounting potentially equated to a loss in revenue of an estimated £40 m per year for the primary care trust (Lawrence et al., 2007).

3.2.2 Use as denominators

The second class of applications, ratios or related indicators based on rates, has more to do with providing a societal barometer or dashboard of indicators for economic management, policy evaluation or other applications. Ratios or rates are used in numerous contexts (employment, health, crime, education etc.), and have been used increasingly by governments and agencies for setting local targets for deliverers of related services or holding local authorities to account. They are usually expressed as percentages or rates per thousand of the population that are exposed to a particular outcome or risk such as unemployment or disease, either incidence (new cases) or prevalence (all cases). Population denominators are also needed for calculating life expectancy which is widely used to measure health inequalities.

In public health applications, attention is properly directed toward the ascertainment of accurate numerators (e.g. the number of MMR vaccinations in children, women who are breast screened). Unlike numerators which tend to rely on administrative counts through case control and reporting systems, denominators are arguably as great a source of inaccuracy. Issues arising include statistical imprecision of population counts, appropriate choice of administrative boundary, breaks in time series, a lack of contemporaneousness, or an inability to measure the population at risk due to lack of specificity in the data (e.g. in terms of age, sex, ethnicity, housing).

In epidemiological studies, good reporting systems are obviously essential for counting the numerators, but good denominators are needed also for measuring vulnerable sub-groups such as ethnic minorities or recent arrivals to the country (e.g. see Roderick and Connelly, 1992; Hayward et al., 2010). The NHS in England, for example, recommends different levels of medical provision based on TB incidence rates and so accurate data are crucial in order to calibrate appropriate levels of medical need in an area to combat this socially corrosive disease (NHS, 2007). However, alternatives to population denominators can be considered when population information is unavailable or unreliable, for example the use of satellite imagery, although clearly this suggestion would not be appropriate in the TB case or many other applications of a similar nature (Viel and Tran, 2009).

Newcastle City Council in the UK, for example, argued for the use of residential properties as the main denominator when creating neighbourhood rates (e.g. crime rates per 1,000 properties, rather than persons). The number of residential properties is available from the council's business and residential property gazetteer. Among the advantages claimed is the high quality of the data, that it is regularly updated and reflects changes on the ground (e.g. new builds, demolitions, and conversions), and that residential properties are identifiable separately from business properties. In addition, the council has control over the data so the denominator matches the period of the data for the numerator[12]. However, this may not be useful where the subject of interest is people rather than properties or households. Our approach also uses property data, but a key difference is that we link property data to administrative data so that we can construct population as well as property ratios in the denominators.

There is also potential for unwelcome interactions between a false numerator and an inaccurate denominator to produce perverse results. A national indicator used by the Government in England until recently provides local measures of the rate of hospital admissions for alcohol-related harm for every 100,000 members of the population. It uses the concept of 'attributable fraction' which is assigned to patients entering hospital based on how much of their condition may be related to alcohol consumption. The calculation is a function of relative risk estimates and population drinking estimates, and therefore relies on the accuracy of population estimates of alcohol consumption and the availability and quality of the relative risk estimates reported in the epidemiological literature (Jones et al., 2008). As a consequence,

[12] http://www.newcastle.gov.uk/core.nsf/a/nnispop [date accessed: November 2010]

it is likely that the method overstates the alcohol harm in some areas and understates the harm in others.

In summary, the danger of using misleading ratios is potentially exacerbated where ratios are used as management targets and result in resources being redirected. The lesson of the last 10 years is that management ratios need to be defined and used with caution including where there is scope for error in both numerator and denominator. It is noteworthy that the new Coalition Government has rescinded the use of targets as a means of control over public funded services and organisations and so the consequences of uncritical applications of ratios are less than previously, although the reasons are primarily political and not data driven. Nevertheless, ratios remain one of the few means of comparing one area or organisational unit with another and so the more that can be done to improve the data on which they are based the better the outcomes are likely to be.

3.2.3 Use in delivery of local services

The third class of applications concerns the design and delivery of local services. Arguably, this is the most challenging of applications as it is much harder to fudge the data. Local authorities in the UK are responsible for supplying local public services such as schools, libraries, public leisure facilities, collecting Council Tax taxes, maintaining electoral registers and managing local public facilities and infrastructure. They are expected to work in partnership with the police, emergency services and health care providers. Each has its own information systems which capture many features of local areas, including the built environment, frequently employing GIS (Geographic Information Systems); however, these are not linked together into a unified system and it is not unusual for different departments of local authorities to use unharmonised data including different population data sources.

In most cases, these systems capture data only on users and not on the population as a whole i.e. people that do not use the service as well as those that do; however, complete information about a population is normally required to identify gaps, undertake needs assessments, or to identify hard to reach groups such as older people living alone. Because management of public services is predominantly carried out at local level, population statistics must be capable of supporting this role. 'With local government in a key 'place-shaping' leadership role, it is vital that every opportunity is taken to refine and improve the available information used to gauge crucial decisions' (Keohane, 2008, p.4). Similar arguments can be set out for health and police services.

For previously stated reasons, population statistics are a long way off meeting this requirement. Local authorities need far more flexible information than is available so that they can answer questions such as: what are the population and deprivation levels for a given housing estate? How many single parents live in social housing and are on benefits? How many nurseries are there within pram pushing distance of households with young children? Who needs to have face to face contact with local services? Are there vulnerable groups that need more personalised services and how many are there (e.g. older people, single parent households, and ethnic groups)? In the following sections, we consider the necessary changes to the system of collecting population counts needed to meet these challenges.

3.2.4 Geography as a barrier

We have seen that the drive for more detailed local information puts heavier reliance on local population estimates. At a national level the percentage error in a population estimate is believed to be small, but broken down into small spatial units at the level at which service providers require accurate information, errors are magnified (Harper and Mayhew, 2012a). Pre-determined spatial units, ranging from electoral wards down to Census output areas, may not correspond to the areas that users are interested in, which may be housing estates, brown field development sites, town centres etc. In this section, we shall argue that users need greater control over the data to link and merge different sources and also to have greater control over geography (i.e. the spatial building blocks on which decision- making is founded).

A good example of the problems caused by geographical inflexibilities of population data is the Sure Start programme for young children. Sure Start is the previous Government's programme (1997 to 2010) to improve the development prospects for young children by coordinating and streamlining services for this age group. In evaluating the impact of Sure Start to see if it had met its objectives, it was found necessary to adjust or apportion data that did not match pre-determined boundaries (Frost and Harper, 2007). As Harper (2002) noted, apportionment techniques are the only option for users that require non-standard breakdowns of data which are far less accurate as a result. It is therefore arguable whether it will ever be possible to conduct robust evaluations of government initiatives as long as this arrangement persists.

History tells us that statistical boundaries of administrative areas are subject to alteration making it impossible to create an accurate picture of change through time. So one can argue a key requirement for any new system of population statistics is that it must be flexible so that

the population of any area may be determined swiftly, accurately and simply. However, even if the estimates were accurate and boundaries are unchanged, there would still be problems unless data are collected in ways that can be flexed to deal with boundary change. Having a single stable geography to suit all needs is arguably unrealistic. Moreover, it is noteworthy that attempts to change definitions and initiate new collections tend to become bureaucratic and unwieldy for a range of reasons.

The inflexibility of administrative geography is also associated with analytical problems of measurement and interpretation. There are two main effects which we now discuss in order to press home the case for change. The first of these is known as the 'ecological fallacy' (e.g. see Greenland and Robins, 1994; Openshaw, 1984a). This is based on an error in the interpretation of statistical data, in which inferences about the nature of specific individuals are based solely upon aggregate statistics collected for the group in a geographical defined area to which those individuals belong—in other words projecting on to the individual, generalizations that apply to a population. In extreme cases, this may have unfortunate consequences in terms of false attribution of causality as well as association, for example, a high crime area may have a high number of single parent households, but single parents are not the cause of crime.

The second class of effects is known as the modifiable areal unit problem (or MAUP). This is a source of statistical bias that affects statistical hypotheses by causing the correlation, or association, between two variables to vary widely (first noted by Gehlke and Biehl, 1934; see also Openshaw and Taylor, 1981). It arises when point-based measures such as the number of people at an address are aggregated into districts or zones so that summary values (e.g., totals, rates, proportions) are heavily influenced by choice of boundary. As (Openshaw, 1984b, p3) laments, "the areal units (zonal objects) used in many geographical studies are arbitrary, modifiable, and subject to the whims and fancies of whoever is doing, or did, the aggregating."

The problem is that we do not know to what extent either the ecological fallacy or MAUP bias decision-making processes unless we have more flexible data. The obvious solution is to create data that can be aggregated into spatial units of any shape or size through using geo-referenced point data at a household or person level. This approach would give users flexibility to choose their own boundaries, add new data from their own sources by person or household linking; it would have the effect of improving accuracy and timeliness and overcome the necessity for apportionment; the potential for ecological fallacy could be minimised by using individual level data (Tranmer and Steel, 1998; Mayhew, 2002); and potential MAUP problems could be investigated using thorough sensitivity analysis.

3.2.5 Use of administrative data in practice

We have argued that knowing the demographic and household characteristics of local populations is a key requirement for policy purposes. Better data help to improve local services and decision making in various ways by improvements to quality, fairness, and value for money. Different sub-groups of the population have different needs and risks; for example young children, older people and single-parent households (Coleman and Schofield, 1986). The current arrangements for collecting information about the population may have served the UK well in the past, but with today's more exacting applications we therefore fully concur with the Treasury Sub- Committee that they leave much to be desired.

As Freedman et al., (2008) argue, local level analysis is critical for local level decisions and policy in which service providers require information not only about costs and volumes of the services delivered, but also about need (how much input is required based on individual requirements), and risk (e.g. what is the probability of an adverse event such as a fall leading to injury with and without the provision of social services). Since the alternatives set out in this paper are based on locally owned data sources, the cost of compiling the data would be easily afforded by the main users. The barriers to producing better population data are, it is argued, not technical but organisational and bureaucratic.

A key requirement is to have data sharing arrangements between data owners in which data are processed in a secure environment or 'safe haven' by a small unit comprising skilled analysts who would be legally bound by data confidentiality. Their role is to geo-reference and link data at a household or individual level, and tabulate and anonymise for statistical purposes. A linked set of administrative data covering local key data sets would provide a platform for more responsive analytical services using better quality data, and more timely population intelligence to support council and local health services. When combined, the data can add greatly to what is known about a population's characteristics. For example, the Annual School Census contains much more detailed ethnic categories than those used by many authorities based on the decennial census.

There are also cost savings because of less reliance on external data sources, product licences etc., less duplication of analysts across organisations through scale economies, and enablement of data sharing with partners e.g. at district level and among health providers. The wider benefits of better data management across partner organisations are more efficient

services through joined up working and, ultimately, better outcomes for local people. Figure 3.1 sets out how a system based on administrative sources would dovetail with present arrangements in which locally available data would be processed and used to support local decision-making as well as providing information at higher levels of government.

In this scheme, data provided by the local community (box C) to service providers (Q) are stored in administrative systems (A). Such systems currently provide data to government departments by way of (B) and in turn these data are processed by departments of state (S) and used to create and evaluate policy and allocate funds accordingly. Statistics are fed back to local areas principally in the form of geographically aggregated data in appropriate administrative units (e.g. local authorities and below). The new feature of our approach is that (A) would be used to generate local statistics and intelligence through (C); (C) in turn would be used to create 'neighbourhood knowledge' that would be fed back to local services (to enable them to perform better) and to the local community (to enable them to participate in decisions about local service as appropriate).

Integral to any new system would be locally available data which, as well as supplying central government with its information requirements, would be exploited and used directly at source. In the next section, we compare the structure of current official population statistics with what can be made available using local administrative data. We do not go into detail here about sources of local administrative data, which are numerous, but instead concentrate on data structures and differences with present arrangements (however, see companion paper for examples).

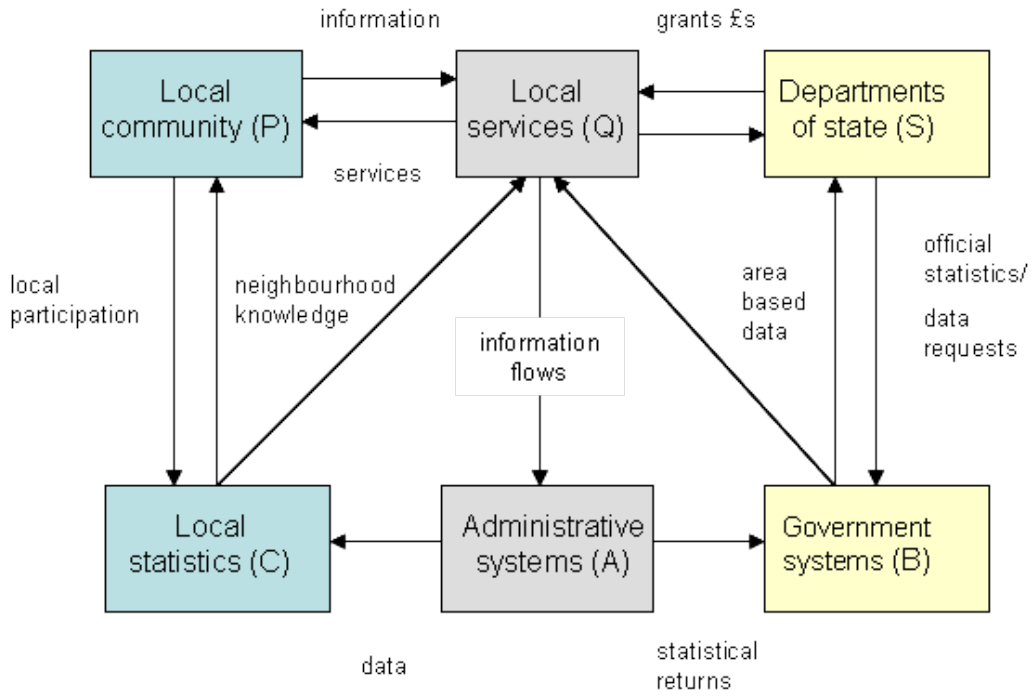

Figure 3.1: The flow of administrative data at national and local level

3.3 Data structures using administrative data sources

The main characteristic of data structures typically provided in official population figures is that they are based on territorial units, whereas in the method described here they are provided at an individual or household level. The main units used by statisticians are Electoral Wards, Census Output Areas, Super Output Areas and Postal Districts, Postal Sectors or individual postcodes. Hence, whereas a typical local authority may have thirty or so wards and each unit postcode (the smallest administrative unit) could have up to eighty addresses[13], the administrative alternative would have one row per individual resident up to the population of an area, perhaps 250,000 records or more. The information contained about each individual would be coded in such a way as to facilitate statistical analysis but would not be cluttered with the myriad of other administrative information held on source databases.

The broad structure of current official population data published by the Office for National Statistics (ONS) is shown in Table 3.1. Variables of interest are specified by the content of the forms used in the Census and published as pre-determined tables. Each variable is an aggregate, such as the number of people aged 85+. Other official data include data sets based on national administrative sources, but these are not fully integrated with ONS data or necessarily harmonised geographically, or by reporting period, or available contemporaneously. Examples include hospital admissions data, National Insurance

[13] http://www.royalmail.com/portal/rm/content1?catId=400044&mediaId=9200078#3400054

registrations and social security data. Some administrative data measure stocks at a point in time and others flows (e.g. migration); time periods are variable (e.g. financial year, calendar year, school year); and publication dates are variable.

The data structure based on administrative data methods described here is shown in Table 3.2 and this is the basis for the approach we have adopted in the application example in Section 3.4 below. Each row is an individual and each column a variable of interest. The nearest equivalent under present population data arrangements would be the census Sample of Anonymised Records (SARs), but as well as being out of date, these do not contain information that would routinely enable other data to be linked. Typically in the administrative data method there would be an area code which could represent an existing administrative unit or one designed by the user that the individual resides in. This would be followed by demographic information and include for example exact age and gender; all records in addition would be anonymised.

Subsequent columns would contain variables derived from a range of administrative sources. The data held in administrative databases essentially comprise four types: (a) categorical or fixed variables such as gender or ethnicity or date of birth; (b) event variables such as the date of a visit or transaction or a birth or death; (c) flow variables comprising the date a service or payments began and ended; (d) the quantum of service provide (e.g. hours of care, meals provided, childcare sessions, day attendance at a care centre and the costs thereof). Not all of this detail is strictly needed for statistical purposes but much will depend on the requirements of the user, with different users having access to different fields as deemed appropriate by local policy and legal requirements.

Neither does the level of detail have to be onerous from a data collection or data definition perspective. Often all that is needed is binary information, for example whether a person lives in social housing or not, categorical data such as household type, and numerical data such as the total number living in a household. Beyond these basic measures a host of other variables may be added from a range of sources such as whether a person lives in a household on benefits (a proxy for low income) and Council Tax band (a proxy for housing wealth). While these variables are not identical to census outputs, they offer valuable individual and household level socio-economic

area	area code ID	population	age group	variable 1	variable 2	etc.
1						
2	<e.g. OA>	<number of units>	<number of units>	<number of units>	<number of units>	<number of units>
3						
.						
.						
n						

Table 3.1: Typical structure of currently available official population data (OA = Output Area)

information that is both detailed and precise. Exact comparisons with the range of information available in the census are impossible, since it will depend on how many administrative data sets and therefore variables are included. However, there are some obvious examples such as religion, place of work which are variables included in the census but not in any comprehensive local administrative data source. Conversely, administrative data are far more comprehensive in areas such as such as benefit status, education, crime, housing, health care and so on.

The x and y columns in Table 3.2 are geographical references so that populations or combinations of variables can be analysed and mapped geographically. These are ascertained by linking a person's address to the Local Land and Property Gazetteer (or equivalent) and extracting the Easting (x) and Northing (y). Usually access to x and y co-ordinates are limited to those that work in a GIS environment and who are usually the custodians of the data. Other users of the data may have more restricted access depending on the context and sensitivity of particular variables or pieces of information (e.g. certain crime or health data). The important point is that the master database is flexible and users' access can be designed and tailored appropriately to their needs.

person	area code ID	age	gender	*x*	*y*	variable 1	variable 2	etc.
1								
2	<e.g. OA>							
3								
.	.							
.	.							
n								

Table 3.2: Typical structure of databases using administrative data sources (OA = Output Area)

The final data set contains additional variables that are constructed from the base data either for individuals or households, such as the number of co-residents in a household. For instance, for some purposes households are a more appropriate unit of analysis than individuals, in which case household level variables are devised by counting or summarising variables by the

property reference number assigned to each individual. For example, if five individuals on the database share the same property reference number, it is inferred that the occupancy of that property is five. To meet this need, we have developed an eight-fold household classification scheme based on individual household demography which is described in the next section. This uses descriptors such as family households with dependent children, older cohabiting households, 3-generational households and so on. These can be broken down into more detailed sub-types as required or users can even create their own classification instead.

There are many other non-core data sets that can be used to enhance the database. In local authorities these include information from service users of adult social care, libraries, educational data and data on children and families. To these can be added a range of NHS and other data sets, each of which contain much information of potential value. These include data on community health services, or hospital admissions, although special arrangements are usually needed to gain access to some of these sources depending on the application. Similar considerations apply to more sensitive information such as crime data or personal medical records; however, a discussion of the details pertaining to data access and related issues in these cases is outside the scope of this paper.

The other main source of local data, which should be mentioned in passing, is survey information i.e. information obtained from specially commissioned surveys. The advantage of surveys is that they collect precisely those data that cannot be obtained from administrative sources. Examples include qualitative and attitudinal data of one kind or another such as willingness to give up smoking or optimism about the future or satisfaction with local services. For example, a health and well-being survey commissioned by one health authority sought information on a person's self-evaluated state of health, drinking habits, income, cohabitation arrangements and so forth. Although usually based on a small sample of residents, such data can be linked to administrative data and used to impute and infer social and other characteristics of whole populations; however, a description of the methods and assumptions involved are also outside the scope of this paper.

We may generalise these statements by adding that virtually any data set could be appended provided they can be linked accurately to individual records by means of a shared identifier. For instance, some commercial data sets such as loyalty card customer data which may provide valuable information on shopping habits and expenditure patterns. In general, the most useful data would therefore be sources that had the potential to fill gaps and were known to be of high quality; however, the most important barrier to obtaining access to such

data sets for statistical purposes is likely to be their commercial confidentiality. Finally, although we have stressed that it is desirable that administrative data should be linkable at a person level, it is possible to contemplate versions of data sets in which linkage takes places at a household (address level) or at higher levels (e.g. output areas). We now turn to a worked example to show how administrative data can be used in actual applications, and explain why such applications would not be possible if users had to rely on existing population data.

3.4 Application example

Following the lead provided by the Treasury Committee, we use this section of the paper to demonstrate examples of applications using administrative data in the three main purposes suggested in their report: (a) to allocate resources; (b) provide denominators to construct ratios such as the number of crimes committed; and (c) plan, deliver and evaluate services at a local level taking into account need and demand. Whilst we cannot provide an example of resource allocation based on the whole country, we can demonstrate a case study that exemplifies each of these purposes at a local level (i.e. typically populations up to 300 k). The principles involved are no different from those at a regional or national level, although of course the analysis will be more disaggregate as a result and therefore more relevant to local decision-makers.

The key point is that data structures enabled by the use of administrative data allow common methodological approaches, regardless of geographical scale and across sub-populations whether at a household or some other spatial level for any given purpose. In addition, through their greater flexibility they help to minimise the danger of MAUP issues or problems of false correlation. We will illustrate the methodology by referring to a health-related study conducted in the London Borough of Tower Hamlets concerning the take-up of free NHS eye tests among older people. Eye testing is simply one of a host of public services that requires people to attend a location to receive a service and so the principles are general even if the details of each service differ (e.g. whether the service is discretionary such as an eye test or compulsory such as education for 5 to 16 year olds). Such distinctions do not affect what follows although it may influence the technical procedures of how resources are allocated.

3.4.1 Background to case study

Tower Hamlets is a densely populated inner London borough, located to the east of the City of London financial centre and bordering the river Thames including Canary Wharf to the south.

The borough is ethnically diverse, but also diverse in terms of income and wealth, being on average one of the most deprived boroughs in the country. Using local administrative data sources, we estimated the population in 2009 to be 234,828 people, living in 100,995 dwellings (Mayhew and Harper, 2010a). The biggest ethnic group are the Bangladeshi community who account for about 32.1% of the population, white British and other white 30.8%, and the rest a mix of Black, other Asian and mixed origins. Figure 3.2, for example, is a density map of the Bangladeshi population by Lower Super Output Area (LSOA), which has been constructed from administrative sources and is provided as one of many possible illustrations of the descriptive detail attainable[14]. However, it also turns out that the Bangladeshi population takes up eye tests more than other groups and why this is so is also of keen interest to health commissioners.

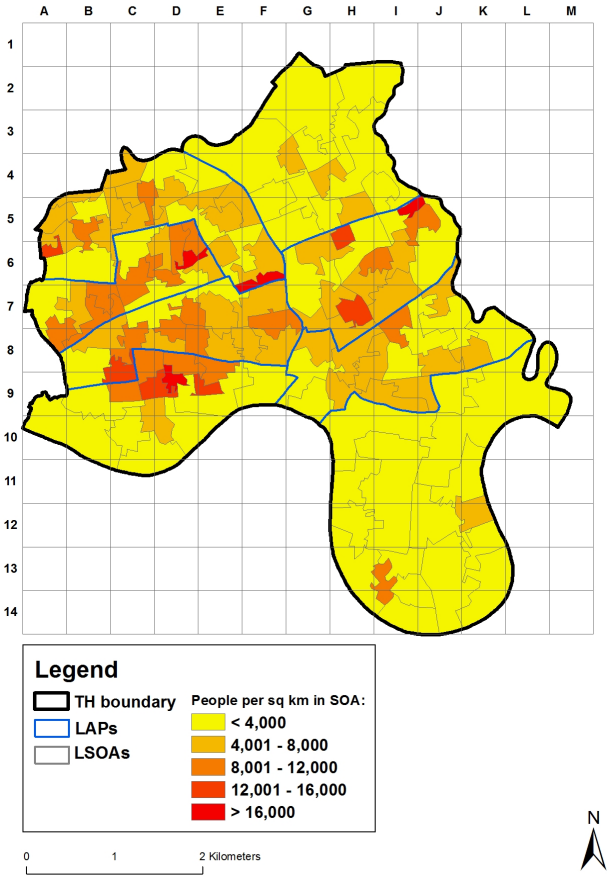

Figure 3.2: Density of Bangladeshi population in Tower Hamlets by Lower Super Output Area (LSOA) (Contains Ordnance Survey data © Crown copyright and database right 2010, and data sourced from London Borough of Tower Hamlets) Note: LAPs are Local Area Partnerships

[14] Ethnicity is assigned to individual records in the data base using a combination of school census records, hospital admissions data and a purpose developed surname recognition algorithm (not discussed here). For this reason, the assignment is probabilistic on a scale of 0 to 1 and population counts by ethnic group are constructed on this basis. A full discussion of the methodology is outside the scope of this paper.

By way of further background, Table 3.3 shows the household structure, population and benefit status of all households in the local authority based on the same administrative data and methodology. This household structure is demographically defined with regard to the age of the occupants of households and is distilled from 81 different sub-types[15]. The table shows that income deprivation is particularly concentrated in household types A to E (see key beneath table), which account for 60% of the population and 36% of households. Around 58% of households in categories A to E receive means tested benefits as compared with an average of 32% of all households, and 55% of households in categories A to E are in social housing.

[15] This classification is not to be confused with commercial products such as MOSIAC or ACORN which also classify households into types. Note that these products rely on a combination of census data and other data sources and use multivariate techniques to produce synthetic rather than demographically based characterisations.

Household type	Frequency by household type	Population by household type	Number of households on benefits	% of all households	Social housing tenure by household type	% social housing	Category	Description
A	17,337	87,635	9,234	53.3	9,321	53.8	A	Family households with dependent children
B	6,819	19,311	3,806	55.8	3,420	50.2	B	Single adult households with dependent children
C	4,530	11,158	2,570	56.7	2,384	52.6	C	Older cohabiting households(1)
D	5,389	5,389	3,539	65.7	3,565	66.2	D	Older person living alone
E	2,530	16,696	2,032	80.3	1,466	57.9	E	Three generational households(2)
F	21,703	50,699	3,259	15	4,071	18.8	F	Cohabiting adult households no children
G	41,305	41,305	6,940	16.8	8,128	19.7	G	Single adult households
H	1,382	2,635	503	36.4	470	34	H	Other households
Total	100,995	234,828	31,883	31.6	32,825	32.5		

(1) At least one resident must be aged 65+; (2) At least one resident must be aged under 20, and one aged 65+

Table 3.3: Household structure, population, tenure and benefit status

3.4.2 Take-up rates of free eye tests under the NHS

We wished to focus on the older population and its access to free eye tests under the NHS[16]. We considered the take-up of eye tests in this group relative to their geographical access to eye testing centres. Lastly, we asked the question by how much eye test take-up would increase if the geographical access to eye tests could be improved by allowing further centres to be opened through enabling GP practices to offer their premises as locations for free sight tests. The map in Figure 3.2 is relevant because it happens that the Bangladeshi population tends to live closer to eye testing centres than other sub-groups and this appears to give rise to a higher take-up of this service. The 'LAPs' in the legend are Local Area Partnership areas which are local authority subdivisions based on aggregations of wards. For some purposes these are the fundamental local unit of resource allocation.

In the UK the National Health Service provides help with the cost of glasses based on age, health or income. Eye tests provided by optometrists are either free under the NHS or must be paid for privately. Entitlement to a free eye test is granted where a person is under 16 (under 18 if in full time education), or aged 60 or over. Free eye tests are also available to people diagnosed with diabetes or glaucoma, or who are advised that they are at risk of glaucoma, who are registered as blind or partially sighted or who are being treated in hospital for an eye condition or who are being prescribed contact lenses. Entitlement is also extended to people in receipt of certain social security benefits, or people aged 40+ whose immediate relatives have been diagnosed with diabetes or glaucoma. Older people are the largest users of this service, but also the most likely not to be tested if the service is inaccessible.

Based on NHS data, Tower Hamlets has the lowest take-up of free eye tests anywhere in London among older people. In 2007–2008, take-up among the 60+ population was 17.8% as compared with a London average of 37.5%, although by 2008–2009 this situation had improved somewhat. However, the true figure may be lower because of known problems with population estimates for this borough. The local primary care trust (the commissioners of these services) were concerned to understand why take up was so low and what could be done to improve the situation. The study ranged widely into areas of epidemiology, aspects of service provision and so forth; but one strand of enquiry concerned the level of geographical access to sight testing centres.

[16] This section draws in small part on a project undertaken in conjunction with PHAST (Public Health Action Support Team) on behalf of Tower Hamlets PCT. Care Needs Assessment: Eye Health, findings and recommendations. Lead author M. Simons, 2009. See also www.nkm.org.uk

3.4.3 Geographical access to free eye tests

Although a densely-populated borough, ease of travel is variable and the locations of eye test centres tended to be skewed towards well established commercial areas in the middle and west of the borough. Figure 3.3 is a map showing the locations of eye testing centres and the locations of all households with a person aged 60+ living there. Each household has been colour coded according to whether there are 0,1,2, or 3 or more centres within a 500 m radius (10 min walk time) of each household (across the border centres are excluded). The lightest coloured symbols have most access and the darkest symbols are homes with least

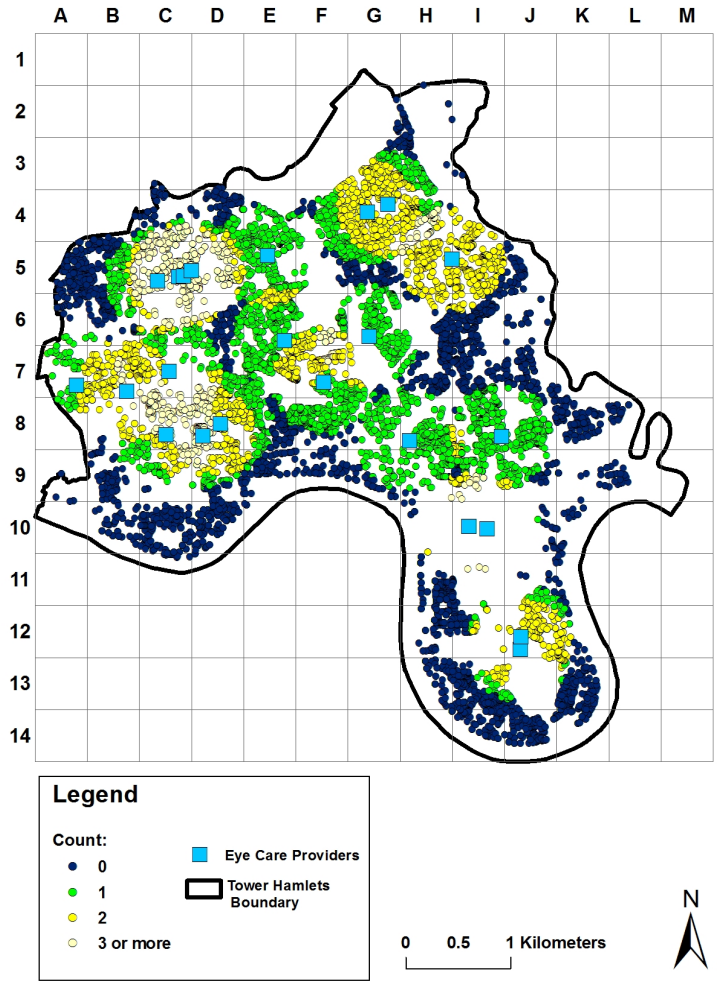

Figure 3.3: Geographical access to eye testing centres based on 10-minute walk time or 500 m. Round symbols indicate locations of households with one or more persons aged 60+ (Contains Ordnance Survey data © Crown copyright and database right 2010, and data sourced from London Borough of Tower Hamlets)

access. Those with least access are spread throughout the borough but especially along a strip bordering the River Thames to the south and in patches elsewhere (e.g. cells A5 and B5 and I6 and I7). We started by examining whether older people are more or less disadvantaged in terms of access than other sub-groups of the population. Under the National Assistance Act

1948 local authorities have a statutory duty to maintain a register of people living in their area who are visually impaired. A person does not have to register as blind or partially sighted, but if they do they may be entitled to certain benefits and services. We analysed the register and found that a person was nearly 13 times more likely to be registered if they were aged 60+.

Using the population database, we segmented the whole population into one of 16 mutually exclusive groups with each group sharing similar attributes. In this case it included whether a person is aged 60+, living alone, is not Bangladeshi, and private housing tenure. We call these risk factors because they act as markers whose influence can be quantified using regression techniques. Note that it would be possible to define other types of risk factors which could include for example a range of clinical risk factors (e.g. see Alder et al., 2005); however, our purpose here was to look at differentials in access.

Table 3.4 shows our results in which those groups with the least access are ranked first and with greatest access last. The numbers in each category are given in column one, and the levels of access and 95% confidence intervals in the final three columns. The intermediate columns indicate the presence or absence of an attribute by the symbol 'Y' and column totals give the total population, the population aged 60+, the numbers in private tenure etc. The results show that levels of access range from 20.5% living more than 500 m from an eye test centre in the best case (row 16) to 46% in row one (worst case). The Tower Hamlets average is 37.1%. This form of tabulation, known as a 'risk ladder', is only possible using linked data. The results show that confidence intervals are acceptably tight around the central estimate and that they capture access differentials succinctly.

Using logistic regression techniques, we ascertained which particular groups were the most disadvantaged. We found that a person was 1.2 times more likely to live more than 500 m from a centre if living in private tenure, 1.1 times more likely if living alone, 1.6 times more likely if not Bangladeshi (all significantly different from one at the 95% level of confidence). It followed that the more disadvantaged groups were likely to be non-Bangladeshi, those living in private tenure, and living alone.

Those aged 60+ were only 0.74 times as likely to live further than 500 m from the nearest centres (all coefficients significantly different from 1 at 95% level of confidence). This suggested that older people had better access than younger people but not as good as the Bangladeshi population which tended to be located nearer to eye testing sites. The 60+ age group with the poorest access tended therefore to be people living alone in private tenure and

not Bangladeshi; the reason why this may be important is that it provides a potential illustration of the inverse care law. This states that the availability of good medical care tends to vary inversely with the need for it in the population served (Tudor Hart, 1971) especially if it can be shown that people that live farther away from a source use the service less.

Case	Number in category	Aged 60+	Not in social housing	Living alone	Not Bangladeshi	% >500 metres from nearest eye test centre	Lower 95% CI%	Upper 95% CI%
1	32131		Y	Y	Y	46.0	45.5	46.6
2	68535		Y		Y	43.3	43.0	43.7
3	5178	Y	Y		Y	38.2	36.9	39.6
4	35345				Y	36.1	35.6	36.6
5	2451	Y	Y	Y	Y	34.8	32.9	36.7
6	6556			Y	Y	33.6	32.4	34.7
7	895			Y		32.9	29.8	36.1
8	4962	Y			Y	31.8	30.5	33.1
9	43746					30.4	30.0	30.9
10	956		Y	Y		30.2	27.3	33.2
11	1875	Y				29.3	27.2	31.4
12	26433		Y			27.6	27.1	28.2
13	4270	Y		Y	Y	27.3	26.0	28.7
14	1248	Y	Y			24.3	21.9	26.8
15	178	Y		Y		23.2	17.2	30.2
16	69	Y	Y	Y		20.5	11.7	32.0
Total	234828	20231	137001	47506	159427	37.1	36.9	37.3

Table 3.4: Table segmenting the population of Tower Hamlets by access to eye test centres according to the given risk factors

3.4.4 Impact of geographical access on take-up rates

To understand how geographical access might affect the take-up of eye tests in the 60+ age group, we analysed 14,000 administrative forms filled in by optometrists after an eye test has taken place. These forms contained a range of other useful information including the presence of certain eye conditions such as glaucoma, and so we were able identify key risk groups. Our analysis showed that males, older people and Bangladeshis were significantly more likely to be diagnosed with glaucoma than other groups, and so older people were clearly one of the high-risk groups.

We found that the ratio of people with diabetes receiving free eye tests to all those receiving eye tests was 8.1%, but this rose to 20.6% in the highest risk group (older males, and

Bangladeshis). This compares with independent estimates for Tower Hamlets as a whole of 5%; however, it is not known whether the 3% margin of difference is a mixture of self-selection or other effects (i.e. people having eye tests are more likely to have an eye condition). Overall, we found that the proportion of older people tested was twice the proportion of people aged under 16 tested, which in turn was twice the number of working age adults tested (i.e. 4:2:1).

Concentrating on the 60+ population, Figure 3.4 shows the percentage take-up of free eye tests based on their distance from the nearest eye testing centre. It shows that around 35% of those living next to an optometrist will receive an eye test in a given year, but this then falls to around 25% at 500 m and to below 10% after one kilometre. Although take-up in this age range is greater overall because needs are greater, it was noteworthy that the amount of attrition (i.e. the falloff in take-up with distance) was higher than in other age groups, and is also higher than in the Bangladeshi population. To put this in perspective, a 60+ person living nearby an eye testing centre would be tested once every 3 years on average but this would slip to 5 or more years or longer if they lived further away.

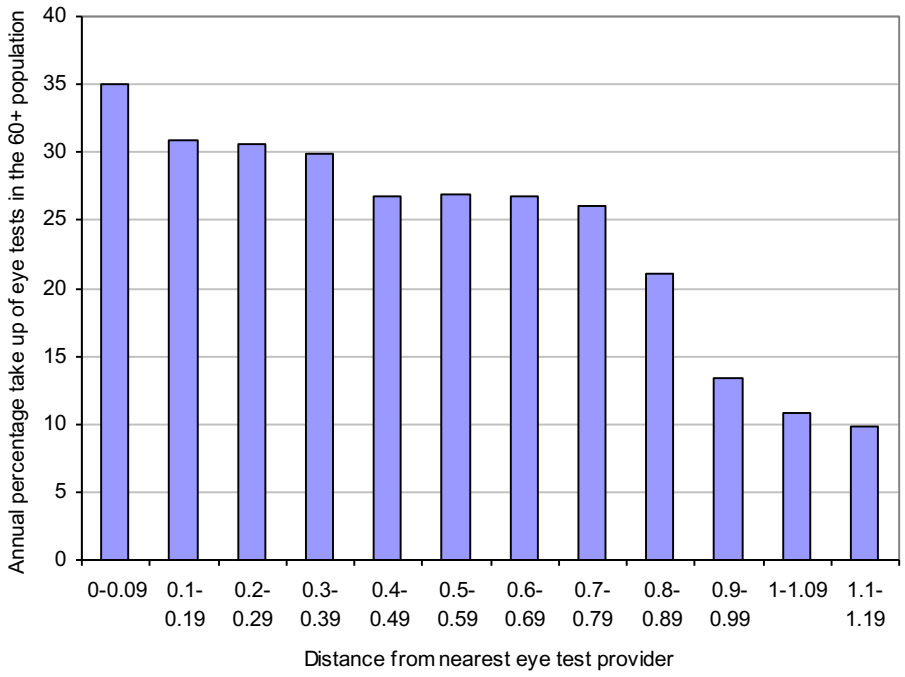

Figure 3.4: Free eye test take-up in the 60+ population based on distance from nearest eye test centre

3.4.5 Evaluation of an alternative service configuration

In answer to the question what could be done to improve access, it is the role of service commissioners to consider the best arrangements for delivering health services. One suggestion was to use local GP practices. We therefore estimated what take-up would be likely to occur in the older group if an optometrist were to perform eye tests in existing GP practices, the argument being that GPs are more numerous and more evenly spread in the borough. In other words, would re-allocating resources to more convenient locations incentivise take-up in this high-risk group? In doing so, we presumed that GP practices would be able to make space available and an optometrist would be able to travel between locations (a mobile service exists for care homes but service levels are currently low). Figure 3.5 shows the geographic effects on access were this to occur on the assumption that travel behaviour would react to distance effects in the same way. As is seen, there would be a far greater equity of access throughout the borough as a consequence, but what would be the effect on take-up?

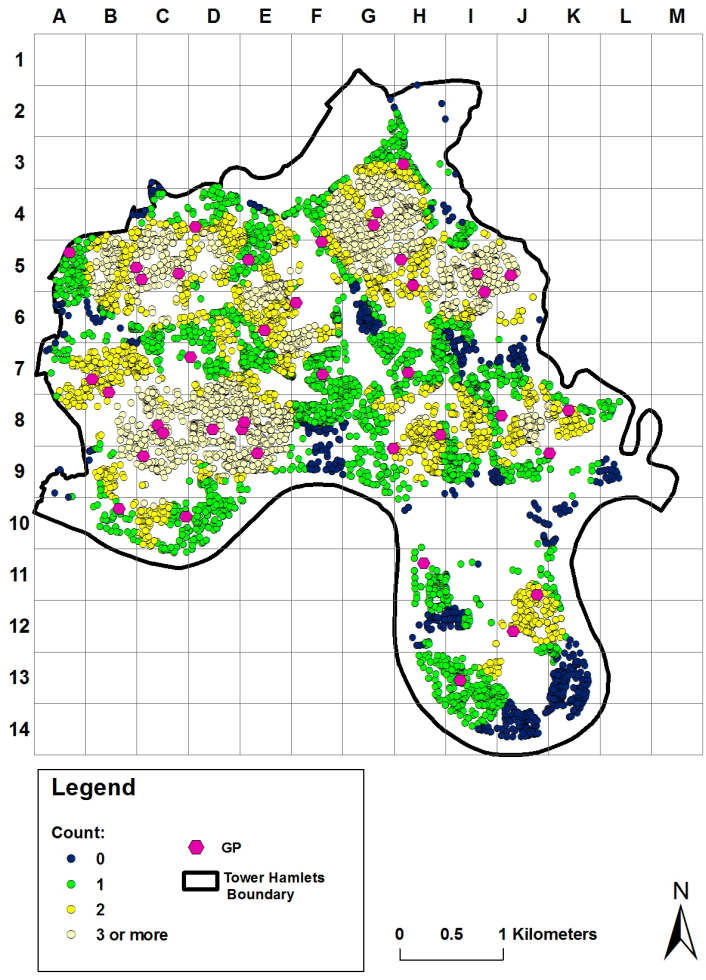

Figure 3.5: Geographical access to GP practices based on 10-minute walk time or 500 m. Round symbols indicate locations of households with one or more persons aged 60+ (Contains Ordnance Survey data © Crown copyright and database right 2010, and data sourced from London Borough of Tower Hamlets)

Figure 3.6 shows the predicted level of take-up following the hypothetical reassignment of the service to GP practice surgeries. It shows that access would be improved in the 0 to 500 m distance range and that the numbers having to travel more than 500 m would fall substantially. Overall, we found that a re-configuration would improve take-up in the borough in the 60+ age group by 8% based on this argument and the rate of overall take-up by 2%. This would have the effect of improving the borough's position within London by a few places, but it would not be sufficient to lift it up to the London average. However, this predicted effect is predicated on the assumption that there would be no other accompanying changes. One such behavioural change arising from the opportunities of co-location would be that older people would seek eye tests on routine visits for clinical check-ups at their GPs rather than having to make separate trips to different locations.

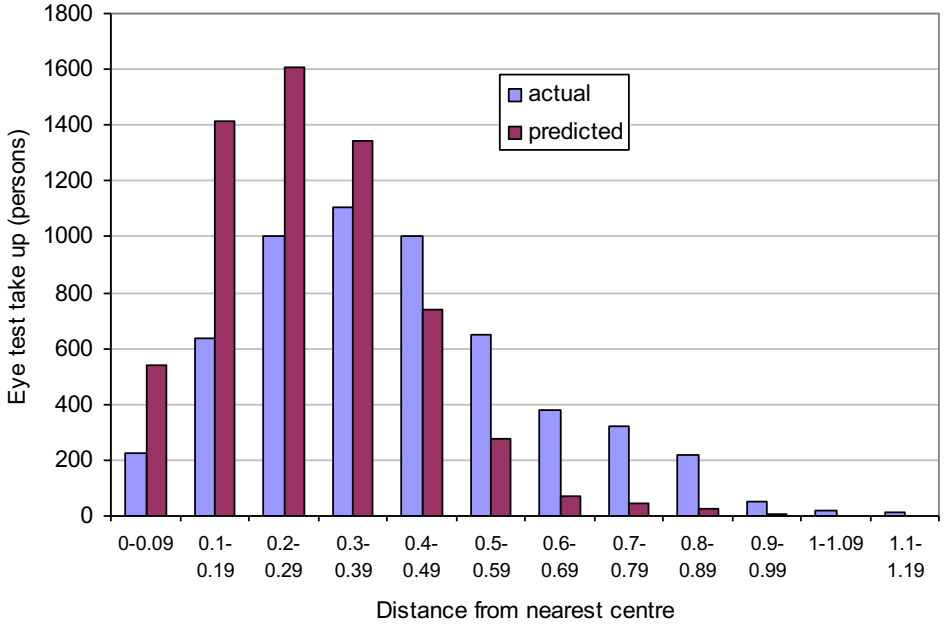

Figure 3.6: Predicted change in eye test take-up in the 60+ population following re-configuration

Thus, we can argue that the 8% improvement in take-up would be the minimum uplift attainable. Note that this analysis does not take account of the costs of re- configuration of the service or the willingness of optometrists to travel between sites. The cost implications however are likely to be fairly small relative to the benefits and since this is a geographically small borough it is maintained that optometrists would not be especially inconvenienced especially as there is already a small mobile service provided.

What has been achieved using administrative data in this example? Firstly, evidence has been provided that, based on administrative data, distance attenuates the take-up of discretionary local services such as eye tests; secondly, that take-up is affected by demographic and socio-

economic factors as well as medical need; thirdly, providing services in GP practice locations would go some way towards correcting the low take-up in vulnerable groups and estimates of the effect were provided. Similar findings might be expected from consideration of many other kinds of services; but what are the resource implications in this case?

3.4.6 Implications for resource allocation

NHS commissioners of these and other services must decide how services will be delivered and a number of funding mechanisms can be envisaged. One is that general practices would pay optometrists to visit their surgeries and so the question arises how GP budgets should be recompensed. This borough is sub-divided into eight Local Area Partnerships (LAPs), each comprising between 25,000 to 37,000 people. As previously noted, the LAPs are used for allocating resources for some services. We will illustrate what theoretical difference three simplistic funding formulae would make to the primary care budgets of each LAP for this service: (A) based on resident population; (B) based on the population served if each resident used their nearest GP; (C) based on the population served if each 60+ person sought to have an eye test at their nearest GP.

Table 3.5 shows the results. Under scenario (A) for example 15.9% of any budget would be allocated to LAP 1 and 9.7% to LAP 2; under scenario B 16.1% would be allocated to LAP 1 and 10% to LAP 2 and so on. The percentage difference in allocation between A and B varies from −0.8% to +0.9%, so overall a range of 1.7%. Under scenario C, which is based on the 60+ population, the range of variation is considerably higher, from −3.4% to +2.4%. Hence, choosing different population bases will lead to different allocation outcomes; in this case, it may be argued that this allocation would have the merit of raising take up in a vulnerable group, which, as was seen, is most likely to be deterred by having to travel.

LAP area	Total population	% of population by LAP (A)	% of population allocated to LAP based on assignment to nearest GP (B)	% of 60+ population allocated to LAP based on assignment to nearest GP (C)	Difference (B-A) %	Based on 60+ population only	Difference (C-A) %
1	37386	15.9	16.1	18.1	0.1	18.7	2.2
2	22813	9.7	10.0	9.6	0.3	8.3	-0.1
3	30048	12.8	12.2	11.9	-0.6	13.1	-0.9
4	26352	11.2	10.4	11.0	-0.8	11.0	-0.2
5	25065	10.7	10.6	13.1	-0.1	13.3	2.4
6	26406	11.2	12.1	11.5	0.9	10.7	0.3
7	30175	12.8	12.6	12.6	-0.2	13.4	-0.2
8	36583	15.6	15.9	12.1	0.3	11.5	-3.4
total	234828	100.0	100.0	100.0	0	100.0	0

Table 3.5: Alternative resource allocation scenarios

3.5 Discussion

The above example has shown how administrative data can aid local providers of a key service and enable the commissioning of better services through reconfiguration. We have not addressed in detail how resource allocation formula would work in other circumstances or for other services as there are different possibilities that would need to be worked through (for example a peripatetic service for the housebound). Much would also depend on the funding mechanisms and the budget holders who would be responsible for specific services, in this case GPs. However, assuming that primary care is the unit responsible for the delivery of a range of services, this illustration shows the extra evidence that administrative data can contribute to this category of decision-making. In the context of government plans to give clusters of GPs a much bigger role in commissioning services, the availability of robust evidence at the local level is essential.

A discussion of all possible methods of resource allocation is outside the scope of the present paper; however, there is a well-established literature on location- allocation techniques. Within the literature, a broad distinction could be drawn between methods that rely solely on the number of registrants with GP practices and those which took account of the proximity or accessibility of patients to a practice location or locations or a hybrid of both (since they would give different results). In this case we have chosen to use where people live rather than where they are registered to avoid a possible circularity of aim (i.e. people registered with a GP, because it is the only one available).

Could currently available population data have provided a similarly detailed analysis? The user would be able to associate eye tests with locations of residence and hence the geographical distribution of take-up at a population level. However, demographic data would only have been available at output area level but not necessarily in disaggregated age categories. Because data are spatially aggregated, it would not have been possible to calibrate the level of take-up attrition with distance with sufficient accuracy and indeed none may have been found; in addition MAUP issues would also have arisen so the results would have been biased by the geography used.

The second problem is that the data would not have been able to distinguish between ethnicity, housing tenure or household demography, so that calibration of the influence of individual risk factors such as these would have been ruled out and yet these were found to be significant influences on take-up. The possibility of ecological fallacy would also have arisen e.g. low take-up is the result of deprivation and not old age and deprivation. Finally, the accuracy of the base data would be questionable since it would be reliant on mid-year estimates which in turn are based on a census baseline that was over 8 years old. To conclude, it is hard to see how such a re-configuration of resources could have been evaluated or justified except through anecdotal evidence and trial and error unless administrative data had been used.

3.6 Conclusions

This paper is intended to contribute to the highly topical debate on how the use of administrative data can replace or improve the current sources of data on population especially at local levels. It has done so by pointing out the significant deficiencies and disadvantages of present arrangements and showing how administrative data could be captured, structured and used in more useful ways. A worked example has been included as evidence that the approach is both practical and achievable.

The first of the three main purposes of population data is to allocate resources to local authorities, health care commissioners and providers, police and other services, and subsequently to areas and services within each territorial entity. The current system of information on population arguably does a reasonable job down to territorial level, albeit the data are flawed through being out of date, and are based on ineffective collection methods with high levels of imputation.

At sub-local authority scale, inflexibilities and inaccuracies are magnified with the result that figures could skew decision making by creating a false evidential picture on the ground. One direct consequence is that denominators used to construct ratios, the second given main purpose of population statistics, such as the number of crimes committed, unemployment rates or new TB cases per head of population will be wrong in the most hard to count areas, with a range of possible consequences.

The approach adopted in this paper is shown to work well at a local level and has several key advantages over the alternatives, including more granularity, greater flexibility and timelier data. This is especially true as far as the given main purpose of population statistics is concerned, namely local planning and intelligence. By being able to link data at person and household level reduces possible concerns about the modifiable areal unit problem and ecological fallacy issues. Working with this level of granularity gives one much greater control over definitions, geography, time windows and analytical methods.

The three main purposes of population estimates have been stated as resource allocation of central funds; to provide denominators; and to aid the effective planning of services and delivery at the local level. The first two relate mainly to capturing accurate population counts, and at the very least, counts that are more accurate than those presently available from national statistics. The proposed methodology meets these criteria because it was originally developed in response to requests from local authorities who perceived there to be a discrepancy between official estimates of their populations and the actual population they believed they had which in turn impacted on their central government revenue allocations.

Implementing the methodology at a national level has not yet been attempted but can be considered. At first glance, it would be a matter of carrying out the administrative population count for each of the local authorities in England and Wales, and combining them into national coverage. This is certainly feasible given the universal coverage of the National Land and Property Gazetteer (NLPG), and component data sets described in a sequel paper in this journal. Consideration would need to be given to people not on standard local data sets or for various reasons treated differently in terms of administrative data. These include the armed forces, prison populations and students in higher education. However, this is no different to present arrangements under the census. Similarly, other external independent data sources could provide information on private school pupils, or for example private GP patients.

Based on our experiences of using administrative data, we believe that the bureaucratic issues are probably more of a barrier to implementation than the technical issues. Both this paper and (Harper and Mayhew, 2012a) have shown how technical issues can be resolved, such as data linking, how information may be structured, and finally how the information may be used to address each of the key areas of application outlined in the introduction. Our experience of working in different locations is that the bureaucratic impediments to a wider adoption of these techniques are mainly the result of confusion between the uses of data for personal and research purposes.

In legal terms, the advice of the Information Commissioner is that Section 33 of the 1998 Data Protection Act provides that personal data may be processed for research purposes notwithstanding the requirement of the second data protection principle providing that a number of conditions are satisfied. These are: no substantial damage or distress is likely to be caused to any data subject; personal data will not be processed in order to support decisions about particular individuals; personal data will not be disclosed (except to a researcher) in a form which identifies living individuals. The data forming the basis for the worked example described in this paper was approved for use by the local PCT and underpinned by a legally enforceable data sharing protocol.

Acknowledgements This paper is based on research for the ESRC UPTAP (Understanding Population Trends and Processes) programme. We acknowledge contributions from Sam Waples of Birkbeck College, London, and thank Richard Verrall of Cass Business School, David Lawrence of the London School of Hygiene and Tropical Medicine and John Eversley of ppre CIC for their comments and support.

4 Using Administrative Data to Count and Classify Households with Local Applications

Preface: Content in this chapter consists of an exact reproduction of the article published in the Journal of Applied Spatial Analysis and Policy in 2016. Only minor edits have been made to make numbering consistent throughout the thesis. As such there may be some dated references or statements. Developments in population data science since the time of publication of this paper are described in Chapter 6.

4.1 Introduction

4.1.1 Why analyse households?

Households are fundamental economic units of production and consumption in which goods and tasks are shared for mutual benefit. Important examples of productive household activities include cooking, cleaning, childcare, care for older people and education (Van der Heyden et al., 2003; Eurostat Statistical Books, 2009).

The Office for National Statistics (ONS), which is the UK's largest independent producer of official statistics and is responsible for the census in England and Wales, was among the first to measure and value unpaid goods and services produced by households and others have since followed in their path (Office for National Statistics, 2000; Holloway et al., 2002).

Statistics Finland, for example, found that GDP is increased by 40 % and household consumption by 60% when its value is included in the National Accounts (Varjonen and Aalto, 2006). More recently, the ONS estimated the contribution to GDP of raising children alone to be worth 23 % (Fender, 2013).

However, the use of household level information has been overlooked in many potentially useful applications, in part because of difficulties of measurement and definition but also the problem of assigning attributes such as income to households as opposed to individuals. The root of the problem, it can be argued, is the difficulty of classifying households systematically among their many variants.

The ONS defines a household as one person living alone, or a group of people (not necessarily related) living at the same address who share cooking facilities and share a living room, sitting

room or dining area. It can consist of more than one family or no families in the case of a group of unrelated people. A dwelling by contrast is a self- contained unit of accommodation with its own front door potentially containing more than one household[17].

In practice, there is a cost in meeting the strict conditions of such a tightly worded definition and experience shows that many users will prefer something that is less than ideal rather than nothing at all. Partly for these reasons, household data tends to be produced by agencies or companies for specific purposes and so, overall, the system is somewhat fragmented and incoherent.

There are other arguments why more should be done to put households at the forefront of research. Households are the basic units for transactions such as paying utility bills, property taxation and for rubbish collection and so tend to have a commercial or proprietary basis, and households are also the basis for measuring poverty in society in the UK[18].

Estimates of the demand for house building rely heavily on household forecasts, which depend in turn on preferred living arrangements (e.g., people living alone as opposed to family units with or without children). Household attributes such as age, gender, occupancy etc. are useful predictors of the need for local services but are hard to source (Bowling, 1991; Ohwaki et al., 2009; Larsson et al., 2006; Ulker, 2008).

From a health perspective, there is interest in the protective value of living in different types of households (e.g. see Marmot, 2010). For example, Vaupel, (2010) considered that only about 25 % of the variation in adult life spans is attributable to genetic differences, noting that "older people are healthier when they live in insulated housing, wear appropriate clothing, eat appetizing food and enjoy their days."

The importance of households is established in other social policy domains. Examples include childcare (Eurostat Statistical Books, 2009), accessing GPs, nursing care and hospital admissions (Van der Heyden et al., 2003), childhood immunisation (Bronte-Tinkew and Dejong, 2005; House and Keeling, 2008), exposure to smoking (King et al., 2009), and alcohol and marijuana use (Wagner et al., 2008).

[17] See ONS: https://www.gov.uk/definitions-of-general-housing-terms

[18] The standard measurement is known as 'Households Below Average Income' (see http://research.dwp.gov.uk/asd/index.php?page=hbai and www.poverty.org.uk/techinical/hbai.shtml).

At the political level, socially excluded or otherwise economically challenged households have been a recent focus of attention because of their high social costs and demands on public money (HM Government, 2011). All of the above arguments suggest that it is timely to look again at how household statistics can be improved and hence this is the primary focus of this paper.

Our particular perspective is from the standpoint of using administrative rather than official data sources. We illustrate our approach at local level where we believe the opportunities for change are greatest. Our methods do not rely on any one single data source but rather several which are combined systematically. The methods and results presented are a sequel to two previous papers by the authors in this journal (Harper and Mayhew, 2012a; 2012b).

Our approach is bottom up in that we use locally available administrative data that we link at a person and address level. This flexibility means that it can be manipulated to suit different definitions and types of household as used by other agencies in the UK or overseas including Eurostat or OECD. Because it goes further than the current system of household classification used by the Department of Communities and Local Government (DCLG) in England, it can also be tailored to local circumstances.

4.1.2 Present arrangements and the case for change

The expanding demand for household statistics reflected in the above is only partly being met by official statistical sources, which tend to be disparate, lacking in consistency, only available in certain geographies and variable in periodicity. In England, the ONS (Office for National Statistics) and DCLG are the responsible agencies for providing official national statistics on households and their equivalents elsewhere in the UK.

To date our work has focussed on the 'local authority' unit, which is equivalent to the concept of a 'municipality' as used by Eurostat. A local authority is responsible for planning and providing services such as housing, education, social care, roads, libraries and rubbish collection. It raises taxes through a levy on properties and receives grants from central government. The term is interchangeable and essentially equivalent to other frequently used terms such as 'borough', 'council', or 'district'.

Data collection begins with the decennial census, the latest of which was in 2011 with households used as the primary unit of enumeration (Baffour et al., 2013). The lowest geography for which data are available at is Output Areas with between 40 and 125 households in each. With the census as a baseline, DCLG then generally provide two-yearly projections at local authority district level and indicative figures of future numbers by household type if past demographic trends were to continue.

Producing household statistics is split into two main stages. The first is the production of ONS local authority based population projections by sex and single year of age, using assumptions about births, deaths and migration. The second stage combines this with information on household composition from Censuses to estimate the proportions of households by local authority area and household type (Department for Communities and Local Government, 2010b).

DCLG are required to provide consistent national and regional projections (Department for Communities and Local Government, 2008; 2010a) in order to allow for comparisons. The results are in effect statistical projections in which household types are fixed and inflexible and not actual counts of households[19]. However, in the view of users, the rapidly changing population in some areas reduces their accuracy and value and figures are not easy to reconcile with other data (Department for Communities and Local Government, 2008).

Nevertheless, DCLG forecasts are extensively used by government departments and local authorities in preparation of development plans (Department for Communities and Local Government, 2010a), and in the assessment of future housing need by house builders and utility providers as well as, for example, the Cambridge Centre for Housing and Planning Research, the Town and Country Planning Association, and the Joseph Rowntree Foundation (Holmans, 2012).

More detailed information about households (e.g., their income and spending patterns) is captured in surveys and used to inform social research and policy development at a national level. The UK Data Service disseminates many of the UK large- scale survey datasets that are available for households such as the Labour Force Survey, the Family Expenditure Survey, and the General Lifestyle Survey which ran from 1971 to 2012[20]. However, they are difficult to use

[19] Although DCLG household types have changed since previous projections following user consultation
[20] See: http://ukdataservice.ac.uk/get-data/key-data.aspx#/tab-uk-surveys

for other purposes and cannot be easily linked to other data at local level without the use of imputation.

In addition to the above sources, commercial geo-demographic products are available at the household level but users must pay. These provide consumer and lifestyle typologies of households rather than an enumeration of household demographics, and are reliant on census, survey and estimated data and so also use imputation to a large degree. In summary therefore, if we take all the different sources of data on households available, the central problems are a lack of coherence among the different sources, coupled with complicated and sometimes opaque methodologies.

The gap that we address in this paper is at local authority level, although our approach is both generalisable and scaleable to other geographies. Local authorities require timely and granular information on population, housing stock, housing costs, tenancy and a host of other variables at sub-local authority level to help inform and review current policy and services. The problem is that the relevant data are distributed among a range of council administrative systems that exist in silos (e.g., property registers, and local Council Tax records).

By being unable to access or link these data can create real problems for users. For example, a recent House of Commons Select Committee report said that "local government would really like more frequent data, if that was possible" (House of Commons Public Administration Select Committee, 2014). However, because support for households and families is largely provided through local authorities, their work is being severely hampered by the lack of evidence for social investment initiatives or calibrating interventions (HM Government, 2011; Harper, 2002; Voas and Williamson, 2001).

Aside from this, the demand for better local data has been growing apace with the introduction of new legislation and financial pressures for local authorities to become more efficient. Examples include the Localism Act (2011) which gives local authorities more freedoms including responsibility for their own Local Development Framework (e.g. Leeds City Council, 2011) and the Health and Social Care Act (2012), which resulted in the creation of Health and Wellbeing Boards to guide local commissioners of services across the NHS.

Given all the above, the inevitable conclusion is that the present state of household statistics is highly unsatisfactory. Available information is too aggregated, inflexible, and out of date or modelled and cannot be easily linked to other data domains such as education, social care or

the private rented sector for effective local policy and planning. It is for these reasons that this paper puts forward a different basis for collecting and maintaining household statistics using locally available administrative sources (Harper and Mayhew, 2012a; 2012b).

Our paper is also timely because it coincides with wider moves to utilise administrative sources for the future production of official population statistics. In particular the aim is to expand the role of administrative data to replace or supplement the Census as first set out as part of the 'Beyond 2011' programme and now being implemented under the 'Census Transformation Programme'. The currently recommended system is for a predominantly online census in 2021 supplemented by further use of administrative and survey data (Office for National Statistics, 2014).

4.1.3 Aims of paper

In this paper, we concentrate on local authority areas for reasons previously set out but also because they capture and are able to provide the required administrative data sets. The proof that a gap in official sources exists is evidenced by the many different studies we have been commissioned to carry out by local authorities and reference is made to some of these. Nevertheless, the approach is by no means perfect as there are some shortcomings that cannot be easily filled and these are also identified.

An important advantage of administrative data is that it is possible to add attributes to households that are not available in official data. This includes for example information on Council Tax, low income households, environmental health and education but also the usage of local services such as libraries or social care. Data are not necessarily limited to local administrative data sources; for example, survey-derived attitudinal data such as health, diet, and household spending patterns can also be considered subject to availability (examples can be provided on request).

The examples in this paper draw upon work undertaken for the six London Olympic boroughs which hosted the Olympic Games in 2012 between 2011 and 2014 (e.g. see Mayhew et al., 2011). The local authorities concerned are Barking and Dagenham, Greenwich, Hackney, Newham, Tower Hamlets and Waltham Forest with a combined population of 1.5 m and 0.6 m households. The resultant databases are in use by each local authority and the combined database used by the Greater London Authority.

The specific aims of the paper are to:

(a) Describe a system for producing flexible classifications and enumerations of household types using locally collected administrative data at address level

(b) Provide worked examples including a short case study on child poverty for informing local policy and decision-making

(c) Inform the wider statistics community by comparing our methods and results with official figures and to highlight differences as appropriate

The rest of the paper is structured as follows. The next section describes the core methodology for converting administrative data into household types and the attachment of attributes. The following section describes an application based on a case study of child poverty in the London Borough of Hackney, one of the six Olympic boroughs. The next section compares our results with official data sources and assesses any differences. A final section discusses the findings and concludes.

4.2 Creating household statistics from administrative data

4.2.1 Background to demographic counts

We use as our base the population estimations created by Mayhew Harper Associates during 2011 for the six Olympic boroughs[21] as at 27th March 2011. Full details of the methodology can be found in (Harper and Mayhew, 2012a; 2012b). This database contains the age and gender of residents for every address within the local authority areas (i.e., taxable entities on the Council Tax register and entries on the Local Land and Property Gazetteer).

Our approach implicitly assumes that households and addresses are one and the same, i.e., they appear as separate entries on address registers and correspond to taxable units for Council Tax purposes. The reason for making this point clear is that administrative sources are address based, so that the case where there is more than one household per address this would not necessarily be identifiable.

Where there is more than one address per individual, we use the most recent and delete the others, but we cannot verify second addresses if they fall outside the boundary areas. It is

[21] The content and results presented in this paper are informed by work commissioned by the six London Olympic boroughs to provide estimates of the population using administrative sources and was designed to coincide with the 2011 Census.

possible to live at more than one address e.g., when parents have joint custody of children. The 2011 Census deals with this by recording the usual residence and the second address of children in this situation and we seek to do the same.

Such limitations are not easily overcome and are features shared with other sources of household statistics. For example official statistics often struggle with the identification of HMOs (Houses in Multiple Occupation) or communal establishments. We are usually able to distinguish these from private and social tenure households and also distinguish between residential and non-residential uses. However, we cannot necessarily differentiate between married or unmarried households as the data are not sufficiently complete or accurate.

Examples of the main communal establishments are prisons, care/nursing homes and educational establishments (e.g., student halls of residence). A residual category includes hotels, hostels, boarding houses, guest-houses, hospitals, sheltered accommodation, children's homes, psychiatric homes/hospitals and defence establishments. Generally speaking using administrative sources, it is possible to identify such establishments from local registers, property gazetteers and other sources.

4.2.2 Alternative household classification systems using administrative data

Local policy makers and planners need to be able to identify differences between household types e.g., pensioner households or three-generation households to support effective policy and decision making. As we shall show, the DCLG types are restrictive and unhelpful, although it is probable that DCLG would maintain that other types of households could be recreated if there was a demand for them. However, such decisions are not in local authority control and so lead to inflexibility.

Figure 4.1 summarises the stages in the process that starts with a list of the administrative systems that provide the starting point for the creation of the person level database in which all subjects are de-identified and which ends in applications in specific policy domains. In the next sections, we explain our method of household enumeration which follows directly after the population enumeration stage as shown in the diagram.

The process aggregates person level data by age, sex or other attributes into one of eight core household types. In this standard classification, households are broken down into eight higher level categories A-H (see Figure 4.1). The next stage in the process is to link to each household

data pertaining to particular attributes of households such as household size, tenure, tax band, benefit status etc.

From this it is relatively simple to develop sub-types of households for addressing specific issues of interest. These typically fall into policy domains such as health, housing, education, the local economy etc. Such information can be used for planning local services, drawing up strategic plans including Joint Strategic Needs Assessments (JSNAs) and so on. However, the same information can also be considered for use in a multitude of other different applications.

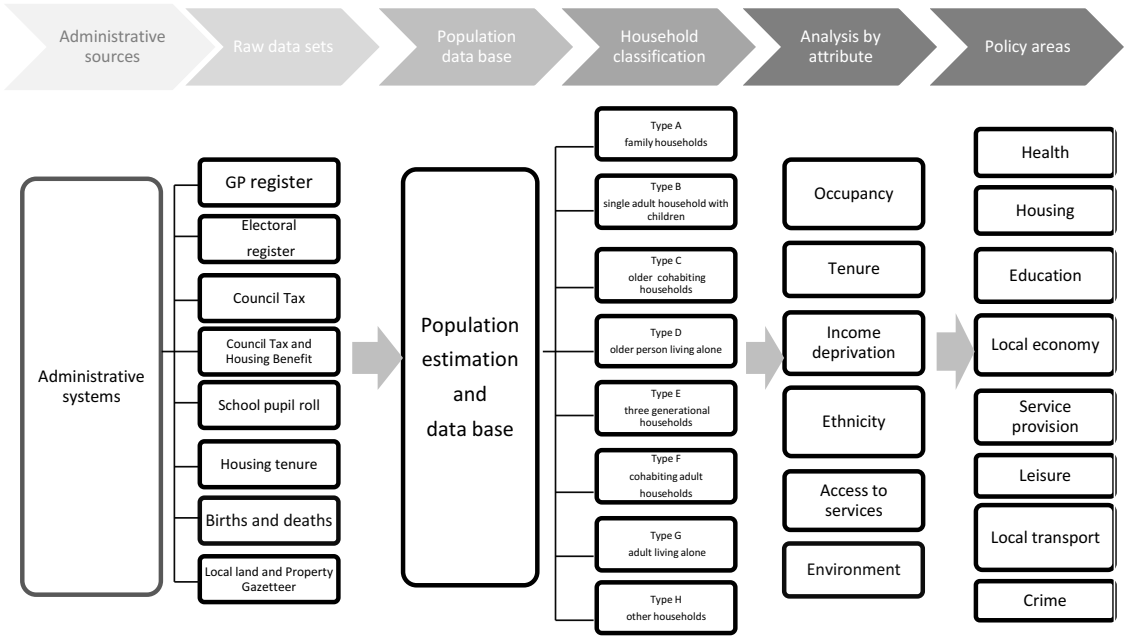

Figure 4.1: Stages in the production of person and household level data and policy domains supported

Take the example of older households, these are much more likely to frequent local shops, doctors, libraries, post offices and day centres than other households and so the whereabouts of these households can be used to inform public providers of these services as appropriate. Other examples include the identification of households suffering isolation or neglect, or houses in disrepair which may be a danger to health or an encouragement for anti-social behaviour.

In consideration of the possible applications a key point to note is that we restrict ourselves to statistical uses of the data. In other words we are not concerned with operational uses which rely on the identification of households or the names of people living in them in order to support some form of Council action. This use of the data is not covered by data protection legislation for which different legal considerations apply but it is covered for statistical uses.

4.2.3 Enumerating household types

We start with the premise that people can be sorted into household types according to their age and the number of occupants. Based on these two variables, we show that it is possible to define as many categories and sub-categories of household as we wish in a single consistent framework. Initially, we define eight household types based on the definitions in Figure 4.2 which we call the 'eight standard types'. Using age and size of households as descriptors we can divide each type into their constituent age groups as shown in Table 4.1 with examples of each.

The methodology is flexible with regard to the number of age groups to be included. To keep to description and presentation manageable we use just three here: Group 1 children (0–19), group 2 working age adults (20–64), and group 3 older adults (65+). Row one is a Type A is a family household with two children and two or more adults (the additional adults could be an older sibling, friend or relative, or someone temporarily resident at an address); and so on. Gender differences can also be included as further sub-types and these are also discussed.

Type	Age group 1	Age group 2	Age group 3	House-hold size	Description
A	**OO**	**OO**		4	Couple household with two children
B	**O**	**O**		2	Single adult household with one child
C		**O**	**O**	2	Older couple household with one person aged 65+
D			**O**	1	Older person living alone
E	**O**	**OO**	**O**	4	3-generational household with one child, couple and an older person
F		**OOO**		3	Cohabiting adult household
G		**O**		1	Adult living alone
H	**OO**		**OO**	4	Split generation household
H'	**OOOO**			4	Young household (e.g. students, teenage parent)

Table 4.1: Specific examples of households defined by size and age group (Key: O indicates a person)

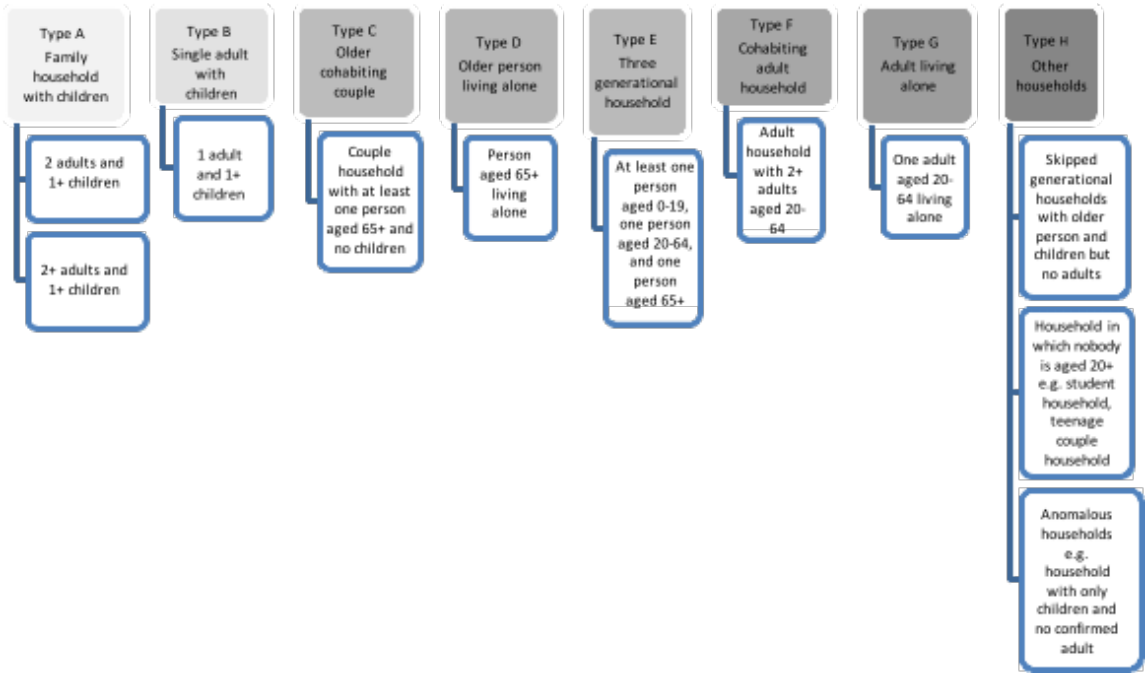

Figure 4.2: The eight standard household types from administrative data

Of the examples shown, Type H households are easily the smallest in number and tend to fall into several heterogeneous sub-categories. Occupants could be young people (possibly students), or are from a split generation (e.g., a household in which children live with their grandparents). Type H may contain what are described as 'concealed' households i.e., separate households within a single address. It can also include anomalous cases where the administrative data have identified children as living at an address but no adult; these cases may be genuine or anomalous due to missing data.

Although, as previously noted, local administrative data cannot easily determine whether a couple household is a married household, divorced or whether people are related in some other way (other than by sharing a surname which is not always reliable or sufficient), for typical uses of household level data it is rarely essential to know this. Conversely, the ability to specify both age widths and household size offers scope to study various attributes of households in far greater detail.

First, we need to be able to enumerate all possible combinations by age and size of household in order to analyse their relative occurrence in the population as well as their attributes. It can be shown that the equation for the number of possible combinations N of households with r age categories and up to n people is given by:

$$N = \frac{1}{0!} + \frac{r}{1!} + \frac{r(r+1)}{2!} + \frac{r(r+1)(r+2)}{3!} + \cdots + \frac{r(r+1)(r+2)\ldots.(r+n-1)}{n!}$$

Where n is the number of occupants per household (0, 1, 2, 3, 4...n) and r is the number of age categories (1, 2, 3, 4....r). Each term inside the brackets multiplied by r gives the number of households with 0, 1, 2, 3, 4...n people where zero indicates the 'void' case (i.e., an empty property). Table 4.2 enumerates the number of possible household types for up to 6 age categories and 6 occupants. The inclusion of void households, the first term in the equation, is retained in order to derive an empty property rate for an area.

number of age categories (*r*)	Number of people in the household						
	0	1	2	3	4	5	6
1	1	1	1	1	1	1	1
2	1	2	3	4	5	6	7
3	1	3	6	10	15	21	28
4	1	4	10	20	**35**	56	84
5	1	5	15	**35**	70	126	210
6	1	6	21	56	126	252	462

Table 4.2: Possible combinations of household demographic types based on size and age (see text for details of the highlighted cells)

For any given value of r and n the sum of the terms gives the total possible combinations of household types. For example, there are a total of 1+3+6+10+15= 35 combinations of household types with 3 age categories and up to 4 people if the void case is included. This is highlighted in row three of Table 2 which adds to 35. Note that this is the same as for the number of combinations for 4 age groups and up to 4 people in the cell below and for 5 age categories and 3 occupants, similarly highlighted.

This result in turn is the same as the number of household types with 2 people and 5 age categories (1+3+6+10+15=35) which is highlighted in column three. In general therefore it can be seen that: $\sum_{n} N_{rn} = N_{r+1,n.}$ and $\sum_{r} N_{rn} = N_{r,n+1.}$

A standard result and important simplification in combinatorial mathematics is that:

$$N = \frac{(n+r-1)!}{(r-1)!n!}$$

Where r is a row in Table 4.2 and n is a column, for which n+r≥1 and r≥1. For example, for N_{44} this is $\frac{7!}{3!4!}$ = 35 which is the same as the previous result as previously be seen the accounting framework that is the result can expand rapidly which means that the number of categories

can soon become unwieldy. Examples will be given shortly in which different sub-sets are selected and analysed in more digestible form.

The inclusion of gender to identify same sex households can also be considered although this leads inevitably to even more variants, but may be relevant in specific applications. However, this possibility simplifies if we are only concerned with the gender mix of a household and not with gender mix within an age group or level of occupancy.

For example any household can be labelled single sex (M or F, or of mixed gender, m). Occasionally there may be data gaps and the gender of one of more people at an address cannot be sourced in which case an 'unknown' category may be included. In practice the number of cases of gender 'unknown' is small and so it is convenient to combine the 'mixed and unknown categories' without much information loss.

In cases where gender is included, the number of household combinations must be scaled by a factor of 3 except for people living alone in which case the scale factor is 2.

If the aim is to consider both occupancy and gender mix then the possible combinations is further increased. For example if occupancy is three, then the possible combinations are MMM, MMF, MFF, and FFF.

4.2.4 Mapping household counts on to standard types

All possible combinations of households conveniently map on to the eight standard types A to H as previously defined in Table 4.1. Proceeding with the example above based on three age groups and occupancy levels of up to four per household, Table 4.3 shows how this mapping works (similar tables can be produced for other combinations of age and occupancy).

This example produces 35 mutually exclusive household types including the void case and is chosen simply because it is compact enough to include in a small table, albeit it is not exhaustive (i.e., it excludes cases where occupancy is greater than 4 persons). This example gives rise to three variants of Type A family households, three Type B, 9 Type C and so on.

Although it is seen that the most occurring standard type is Type H of which there are 10 variants, in practice they only account for less than 2 % of all households. If voids are excluded, typically the most numerous household types, accounting for around 96 % of the total, are

Types A, B, C, D, F and G. Type E 3- generational households also account for less than 2 % of the total and so are similar to Type H.

Case	Age group 0-19 (A)	Age group 20-64 (B)	Age group 65+ (C)	Household occupancy (A+B+C)	Standard household type
1	0	0	0	0	void
2	0	0	1	1	D
3	0	1	0	1	G
4	1	0	0	1	H
5	0	1	1	2	C
6	0	0	2	2	C
7	0	2	0	2	F
8	1	1	0	2	B
9	1	0	1	2	H
10	2	0	0	2	H
11	0	2	1	3	C
12	0	1	2	3	C
13	0	0	3	3	C
14	0	3	0	3	F
15	1	2	0	3	A
16	1	1	1	3	E
17	1	0	2	3	H
18	2	1	0	3	B
19	2	0	1	3	H
20	3	0	0	3	H
21	0	3	1	4	C
22	0	2	2	4	C
23	0	1	3	4	C
24	0	0	4	4	C
25	0	4	0	4	F
26	1	3	0	4	A
27	1	2	1	4	E
28	1	1	2	4	E
29	1	0	3	4	H
30	2	2	0	4	A
31	2	1	1	4	E
32	2	0	2	4	H
33	3	1	0	4	B
34	3	0	1	4	H
35	4	0	0	4	H

Table 4.3: Mapping household demographic combinations on to the eight standard types, A to H for the case of three age groups and up to four occupants

4.2.5 Examples of household enumeration

An administrative data-derived population count is arranged such that each person is represented as a row in a de-identified database to which other attributes can be linked relating to the individual or to the household. Users of this approach will be particularly

interested in examples that would not be reproducible using official sources but are nevertheless deemed useful.

These will depend on their availability in other datasets used. This could include information about the services accessed by individuals or households (e.g., schools attended); or it could involve commonly required attributes such as size and tenure as already suggested. An especially important example is benefit status: in the UK households may qualify for financial support to pay their rent or reduce Council Tax bills. Eligibility is based on income and savings and so we use this as a proxy for a low-income household.

The example in Table 4.4 is designed to provide boroughs with a picture of low-income households. It covers all six Olympic boroughs and is simply a summation of the population and household types split by gender, average occupancy, low income status, tenure and tax band. Tenancy figures indicate the size of the council and social rented sectors; tax band information is included on relative housing wealth by household type, with bands A to C being a proxy for relatively low value housing[22].

It confirms that the most numerous household types are Types A and G which are single working age adult households or cohabiting working age adult households.

Household type	Frequency of household type	Population	Average occupancy per household	% male only	% female only	% mixed or unknown gender	% housing units tax banded A-C	% of households on benefits	% of social housing units
A	125,856	597,670	4.7	0.8	2.0	97.2	65.4	38.9	34.2
B	57,209	160,366	2.8	6.8	27.2	66.0	74.5	56.2	43.6
C	40,596	100,702	2.5	6.0	6.9	87.1	59.6	41.8	32.4
D	42,114	42,114	1.0	36.4	63.4	0.2	76.9	58.3	48.0
E	11,893	71,027	6.0	0.6	3.2	96.2	56.1	55.9	35.0
F	108,798	274,558	2.5	13.4	10.2	76.5	62.1	21.3	25.6
G	188,610	188,610	1.0	52.7	40.2	7.1	68.8	23.5	26.7
H	8,797	19,715	2.2	27.8	32.6	39.6	57.2	41.4	30.9
Total	583,873	1,454,762	2.5	23.9	23.7	52.4	66.8	34.3	31.9

Table 4.4: Summary table showing a breakdown of households across the six Olympic Boroughs and selected key attributes

However, the greater proportion of the population lives in Type A family households and Type F cohabiting adult households. Numerically, the smallest standard types are Type E 3-generational households and Type H households. Types B, D, E and H are more likely to be

[22] In the UK, residential properties are banded by value into eight categories from A (lowest value) to H (highest value) with bands A to C being a proxy for low value housing and D to H for higher value housing. The convention is to italicise tax bands in order to distinguish them from household types.

benefit households; Types B, D and E are more likely to live in social housing; and Types B, D and G more likely to live in lower value properties.

Table 4.4 also shows that gender mix by household type is quite intuitive but it is not always possible for data reasons to identify gender, so that in 2.5 % of households gender is 'unknown'. For simplicity this has been subsumed into the 'mixed and unknown column'. From the table we can infer that females are nearly twice as likely to be the sole survivors in older type D households and are more common in single parent households (e.g. the case of a female parent or guardian and at least one female child).

Differences in the average size, occupancy and age of households can be demonstrated in various ways. Figure 4.3 is a scatter-gram showing average occupancy versus average age at output area level across all six boroughs. It shows that each household type forms a characteristic cluster in these two dimensions; only Type H does not show any clustering tendency.

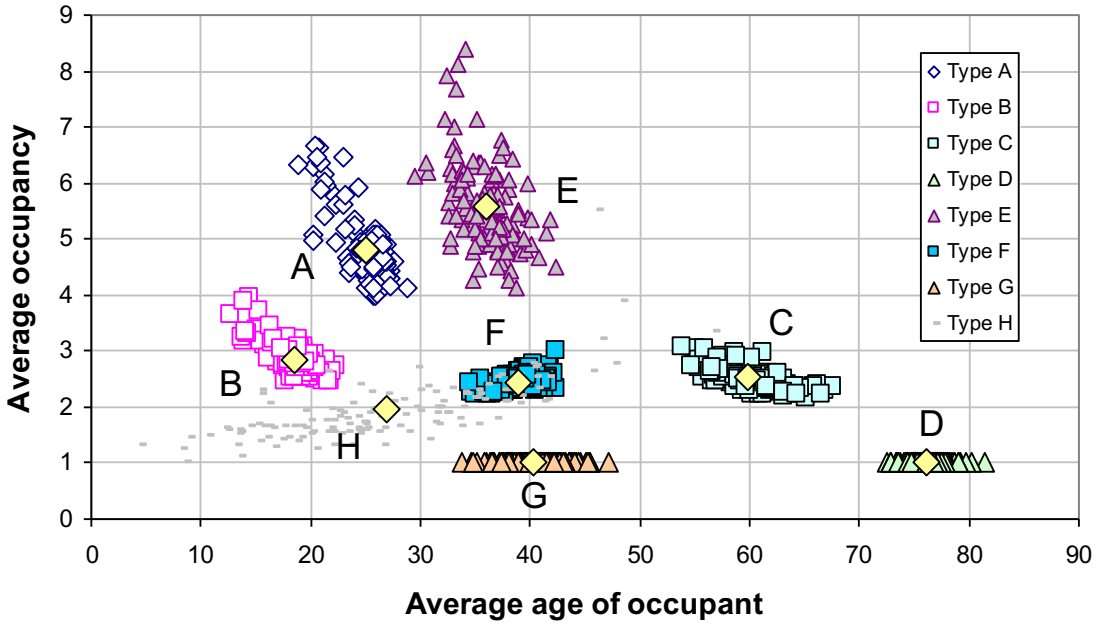

Figure 4.3: Scatter-gram of household types showing occupancy versus average age by output area

Table 4.5, on which Figure 4.3 is based, shows that Type A family households typically contain four or five persons, including children, with an average age of 25 years and occupancy of 4.7 persons; Type B single parent households are about 6 years younger and range in size from two to three persons and an occupancy of 2.8 persons; Type C older cohabiting households

range in size between two or three persons with average age of 62 years and an average occupancy of 2.5 adults.

Household type	Average age	Standard deviation	Average occupancy	Standard deviation	As % of all households
A	25.0	1.8	4.8	0.6	21.6
B	18.4	0.3	2.9	0.3	9.8
C	59.9	2.6	2.5	0.2	7.0
D	76.1	1.6	1.0	0.0	7.2
E	36.0	2.3	5.6	0.8	2.0
F	38.9	1.7	2.4	0.1	18.6
G	40.3	2.5	1.0	0.0	32.3
H	26.9	9.8	2.0	0.6	1.5

Table 4.5: Average household age and occupancy

Type D older single person households average 77 years and are the dominant type of older household at the oldest ages; Type E three-generational households have an average age of 36 years and occupancy of 5.8 persons; Type F households are cohabiting adult households with an average age of 40 years and occupancy of 2.8 persons; Type G single occupancy adult households have an average age of 42 years. Type H is the least homogenous type of all but only account for 1.5 % of all households.

4.3 A case study: child poverty in Hackney

Local authorities are interested in enumerating the number of child households for a range of purposes; for example, the concept of 'child yield' is frequently used to predict the demand for housing and school places. Such information is also extremely valuable to health providers and social services to identify vulnerable families and health needs. In this short case study, we enumerate, map and analyse children living in households by tenure and benefit status and compare access to children's centres in the borough. We choose as our case study the London Borough of Hackney, one of the six Olympic boroughs[23]. In its state of the borough report in 2013, it records that about 37 % of all children in Hackney are affected by child poverty, according to the standard national child poverty measure and is the third highest rate in London[24].

Poverty varies spatially within the borough and impacts different communities unequally. Hackney has a very diverse population with at least 14 nationalities each having over 1000

[23] http://www.hackney.gov.uk/Assets/Documents/estimating-and-profiling-the-population-of-hackney.pdf
[24] See 'State of the Borough Report 2013; Section 2 Child Poverty and Family Well-being'. http://www.hackney.gov.uk/Assets/Documents/Reduce-Child-Poverty-and-improve-Family-Well-being.pdf

members. It is also home to the Charedi population, a major Jewish orthodox sect with a population of around 18,000.

One of the things we are able to do is to identify households by ethnicity and in one case by religion. This is based on an extensive data set based on self-declared ethnicity derived from the School Pupil Census which we use to probabilistically assign ethnic status to people and households.

Hackney Council believes it is useful to build up a picture of different communities, whether defined by socio-economic criteria such as age, gender, ethnicity or different religious affiliations, in order to understand better their size and distribution and to design services that better meet their needs and expectations equally and fairly.

Working closely with the Charedi community, we used the Shomer Shabbas, a register of Charedi heads of households of Jewish orthodoxy, to estimate the population. Using the highly distinctive names therein, we estimated the probability of people with these names being Charedi, extending our search to include the whole population, not only those on the register.

In parallel, we also identified two other communities for analysis, namely Turkish and Bangladeshi households. Each community forms a distinctive group in terms of child yield, tenure and benefit status and like the Charedi are easy to identify. In comparison the Charedi community is highly clustered towards the north of the borough, but the other two communities are more widespread.

Table 4.6 enumerates the whole population and each community by number of children, tenure and benefit status as at 2011. Our definition of a 'child' is any one age 19 or under for these purposes (later we focus on the 0–4 s). As can be seen the table shows quite different experiences in each community according to each of the attributes: benefit status (a proxy for low income[25]) and social housing (a proxy for supported housing).

For households with at least one child the table incorporates two household types: Types A family households and B single adult households with children. Households that have no

[25] Hackney bases its measure of child poverty on the proportion of children living in families in receipt of out of work benefits or tax credits with reported income less than 60 % of median income. Our measure based on locally administered means tested benefits gives a very close approximation to this.

children or are empty ('void') are included for completeness. Comparing columns it can be seen that child yield in each of the communities is much higher than for the whole borough.

In the case of the Charedi community household size is especially large with 53 households having 10 children and 31 more than 10. Also we see that Charedi households are far less likely to live in social housing than either Turkish or Bangladeshi households.

All					Turkish				Bangladeshi				Charedi			
No. of children in house-hold aged 0-19	No. of house-holds	% social housing	% on benefits	% of all house-holds	No. of house-holds	% social housing	% on benefits	% of all house-holds	No. of house-holds	% social housing	% on benefits	% of all house-holds	No. of house-holds	% social housing	% on benefits	% of all house-holds
void	5,975	n.a.	n.a.	5.4	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
0	73,310	45.9	33.5	66.3	1,943	66.2	70.8	45.3	511	56.6	54.6	35.1	1,424	19.0	35.2	36.0
1	13,733	60.2	48.5	12.4	926	80.9	79.5	21.6	227	74.9	74.4	15.6	550	16.5	61.1	13.9
2	9,635	63.9	53.5	8.7	925	85.3	86.4	21.6	256	81.3	76.6	17.6	473	18.0	67.9	12.0
3	4,432	71.0	61.7	4.0	370	87.3	87.8	8.6	215	85.1	84.7	14.8	364	19.5	76.4	9.2
4	1,832	70.7	66.3	1.7	89	82.0	75.3	2.1	140	87.1	82.9	9.6	319	18.2	69.6	8.1
5	736	62.4	70.8	0.7	29	75.9	89.7	0.7	68	77.9	88.2	4.7	219	19.6	65.8	5.5
6	394	51.8	72.1	0.4	6	83.3	83.3	0.1	19	68.4	84.2	1.3	178	23.0	69.1	4.5
7	242	38.8	72.3	0.2	2	50.0	100.0	0.0	11	45.5	72.7	0.8	150	19.3	67.3	3.8
8	139	30.9	73.4	0.1	1	100.0	100.0	0.0	6	83.3	100.0	0.4	108	21.3	72.2	2.7
9	99	24.2	65.7	0.1	0	0.0	0.0	0.0	1	100.0	100.0	0.1	89	22.5	66.3	2.2
10	59	22.0	66.1	0.1	0	0.0	0.0	0.0	0	0.0	0.0	0.0	53	18.9	66.0	1.3
>10	41	22.0	17.1	0.0	0	0.0	0.0	0.0	2	50.0	50.0	0.1	31	16.1	58.1	0.8
Total/ average	110,627	48.3	37.5	100.0	4,291	75.7	77.7	100.0	1,456	72.1	71.0	100.0	3,958	18.8	56.0	100.0

Table 4.6: Summary table showing a breakdown of households in Hackney according to the number of children, housing tenure and benefit status according to three different communities as at 2011

The table shows that the percentage of households on benefits is 33.5 % in a childless household rising to 48.5 % in one-child households and steadily increasing to 73.4 % in 8-child households. This percentage varies considerably between the three communities, but in general the greater the number of children the more likelihood a household will qualify for financial assistance.

How does this compare with other low income households? A useful finding is that the risk of any household being on low income can be boiled down to a small number of risk factors. Using logistic regression it can be shown that a household is 2.6 times more likely to be on means tested benefits if there is any child aged 0–19; and 3.4 times more likely if there is an older person aged 65+ (for further information on logistic regression (Altman, 1999) see e.g., Altman 1999).

Table 4.7 summarises the five main risk factors and their influence on income poverty: Four relate to ages of occupants and one to housing tenure and together they statistically explain 87 % of the variation in benefit households. However, they also have the special property that they can be used in combination to reproduce each of the eight standard household types. For example, a Type A household must have at least one- child age 0–19 and two adults aged 20–64, but if it has only a single adult and at least one child then it is a Type B household.

Risk factor	Odds ratio	Lower CI	Upper CI
Any child 0-19	2.6	2.5	2.7
Single adult 20+	1.3	1.26	1.33
At least one person age 65+	3.4	3.30	3.60
At least one person aged 20-64	0.9	0.87	0.98
Living in social housing	4.2	4.10	4.40

Table 4.7: Odds of income deprivation by risk factor including 95 % confidence intervals (CI)

The odds in Table 4.7 are multiplicative so that for example a single adult Type B household would be 2.6×1.3×4.2=14.2 times more likely to be on benefits than a household with none of these risk factors. Extending this further, a Type C older household must have at least one person aged 65+, but if that person lives alone then it is a Type D one person older household. In contrast, a Type E 3-generational household must have at least one child, an older person and a working age adult.

4.3.1 Access to children's centres in Hackney

It is generally accepted that having access to affordable, good quality childcare has a bearing on parental decisions when they are in the process of returning to or entering work which can help lift families out of poverty. For many years there has been a national programme of Sure Start children's centres to target those in greatest need of support and for which responsibility for their running has since been devolved to local authorities[26].

In this section we evaluate to what extent childcare and other needs are being met in Hackney based on the existing network of centres. Under present rules children aged 0-4 years old who are resident in the borough are eligible to attend one of 22 such centres located in the borough. However, their attendance at these centres is subject to strict criteria including evidence of residence in Hackney and also proof of household income.

Plainly it is important that the centres should be accessible to those in greatest need and so we mapped all households likely to qualify on these criteria and also the centres. We would expect a typical average catchment radius of 0.5 km for this number of centres and size of local authority; we term this radius 'pram pushing distance' and it is equivalent to a 6 to 10 min walk time.

We split all households with children aged 0-4 years old into groups identifiable by being in one of the three communities above and also whether on means tested benefits or not. We then mapped the results and tabulated how many households meeting these criteria had access to none, 1, 2 or 3+ centres according to households in each community: Turkish, Bangladeshi or Charedi.

Based on the 15 k households with children aged under 5, 8.2 k were low income households. Of these, 26 % of all households had no access to one or more children's centres within pram pushing distance whereas 74 % did. We also found that benefit households had slightly better access than non-benefit households which is what one would expect, albeit by only an unexpectedly small margin (only 68 versus 66 %).

Figure 4.4 is a map of children's centres and of all households in Hackney that meet both the benefit criterion and have at least one child under 5 years old. Each household is colour-coded according whether there are 0, 1, 2 or 3+ centres within 500 m. It can be seen that children's

[26] https://www.gov.uk/sure-start-childrens-centres-local-authorities-duties

centres are widespread throughout the borough, but also that many households fall outside the pram pushing criterion (see dark blue symbols).

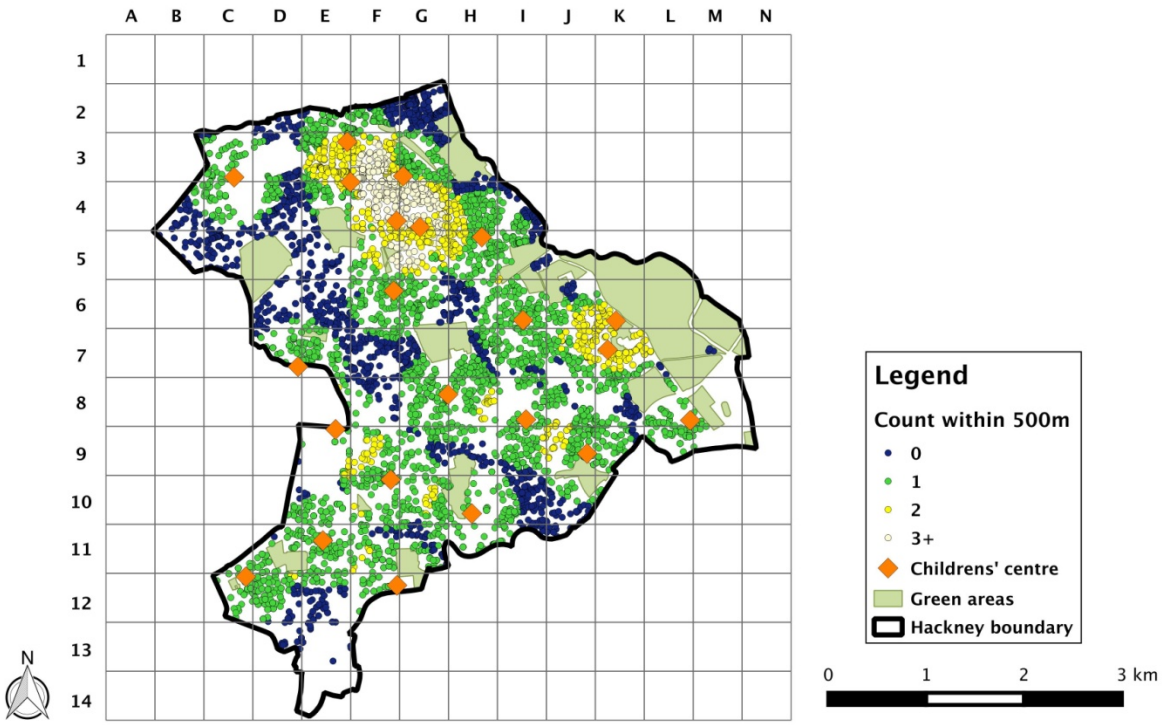

Figure 4.4: Map showing the locations of households on benefits with children aged<5 that are outside pram pushing distance from the nearest children's centre

The clear impression is that the map shows several large gaps in the network – but also areas with access to 3 or more centres. The area of greatest choice is in the north of the borough between cells E3 and G5. This area is strongly identified with the Charedi community but because the Turkish and Bangladeshi communities do not experience the same degree of co-location, their access is much more variable by comparison.

Further analysis shows that 46 % of Charedi households have a choice of two or more centres within pram pushing distance as compared with only 13 % of Turkish and Bangladeshi households. Clearly local authorities do not set out to create unequal access to public services but we would argue the quality of the data they use often mean that decisions are too broad brush relative to the objectives they seek to achieve. Our main conclusion therefore is that children's centres are widely dispersed in this borough but their planning could have benefited from better fine tuning. Although it is not possible to reverse the clock by reconfiguring existing centres, it cannot be ruled out that some centres may be forced to close due to budget constraints and so this is also another possible use of the data.

This kind of analysis can be used to ensure services are located equitably but other factors are important too. For example, although the Charedi community appears to be very well served geographically, it tends to make parallel arrangements for its own children's needs based on their religious beliefs. Hence, the issues are even more complex but it is precisely for dealing with these issues that our methodology is well suited.

Access to children's centres is based on residence and so there is a further question of boundary effects when we are dealing with other services that are located just outside the local area. We did not cover these cases here because of the strict residence eligibility conditions governing this service, but they can be easily addressed by working with services in neighbouring areas or by analysing several boroughs together especially for services with trans-boundary catchment areas.

4.4 Administrative counts versus official household statistics

Previous sections have sought to explain the use of local administrative data to enumerate and classify households starting with the raw administrative data. How to present and use the information in digestible form for different purposes was set out in a simple accounting framework using examples and tested using a case study. A key question is how much confidence can users have in administrative approaches to the enumeration of households?

In this section we compare our findings with official figures produced by the ONS, DCLG and GLA (Greater London Authority). Whilst we do not expect to find an exact correspondence, it is useful nonetheless to identify reasons for any differences. From our experience of reconciling the administrative approach with official sources, we expect to meet two potential problems. One is the different basis used to count populations (i.e., Census versus administrative sources); the second is translating administrative data into exact copies of officially used households definitions.

One obvious and insurmountable difference is that Census data are for a point in time and updated only 10-yearly, whereas administrative data are constantly being updated. For this reason, official household statistics covering intervening years will tend to be based on a complicated mix of fact and imputation in which potential errors are impacted by timing differences in population counts and the assumptions used regarding household formation.

As well as DCLG, the GLA (Greater London Authority) also produces its own household projections for London boroughs using housing development trajectories based on the Strategic Housing Land Availability Assessment (SHLAA). The GLA use the same household definitions as DCLG but a key difference is that they use their own population estimates as a basis. However, the availability of GLA data affords the opportunity to benchmark our household counts with both sources. The version available at the time of this analysis is the 2009 round SHLAA based projections[27].

Our results for the year 2011 are shown in Tables 4.8 and 4.9 for each of the six Olympic boroughs. Table 4.8 compares household counts for 2011 based on administrative sources with comparative figures created by DCLG in 2010 (the version available at the time of the analysis), an updated DCLG version made available in 2013[28], Census 2011, and the GLA 2009 round SHLAA based projections, all without communal establishments.

[27] The most recent GLA data available at the time of writing is based on 2013 SHLAA data

[28] As well as subsequent interim results since 2013, an update incorporating full Census 2011 information has been postponed by DCLG until late 2015

	Households ('000s)					Difference % in households			
Local authority	Admin 2011 (A)	GLA 2011 [a] (B)	DCLG 2011 [b] (C)	DCLG 2011 [c] (D)	Census 2011 [d] (E)	Admin - GLA	Admin-DCLG (C)	Admin- DCLG (D)	Admin - Census
Barking and Dagenham	70.5	72.3	69.3	70.1	69.7	-2.5	1.7	0.6	1.1
Greenwich	101.6	106.7	99.5	101.4	101	-4.8	2.1	0.2	0.6
Hackney	103	98.1	92.1	102.1	101.7	5	11.8	0.9	1.3
Newham	104	103.2	91.8	102.3	101.5	0.8	13.3	1.7	2.5
Tower Hamlets	101.2	100.6	98	102.1	101.3	0.6	3.3	-0.9	-0.1
Waltham Forest	97.9	95.1	92	97.4	96.9	2.9	6.4	0.5	1
Total	578.2	576	542.7	575.4	572.1	0.4	6.6	0.5	1.1

Table 4.8: Comparison based on total number of households by local authority using administrative, DCLG, GLA and ONS Census data, and the % difference of each compared to the administrative data counts

The results show that for the whole region administrative household counts based on administrative data are 0.4 % higher than GLA at 578 k but 6.6 % higher than the 2010 version of the DCLG figures (source b in table 4.8). The addition of the more recent DCLG 2013 figures (source c in table 4.8) substantially reduces this difference down to 0.5 %. This demonstrates that the DCLG figures have fallen more in line with the administrative data results, and that the Census figures are also very close – as is to be expected because the 2013 DCLG figures draw on the Census results.

It also demonstrates the change from using the Mid-Year estimates as the original population base for DCLG 2010 figures that were known to have undercounts for these areas, to the Census 2011 population estimates in the DCLG 2013 version. This is hence demonstration of the inconsistencies than can occur and which are reinforced even within the same sources over time.

Within the Olympic boroughs the differences vary by local authority but two that particularly stand out are between the administrative and original DCLG household counts for Hackney and Newham. Again, it is noteworthy that these are greatly reduced in the post-Census 2013 DCLG figures.

Another comparable measure is estimates of the number of vacant dwellings. For the administrative data method we define the vacant dwelling rate as the percentage of the total number of residential addresses on the LLPG (Local Land and Property Gazetteer) that are not occupied, having first removed all communal establishments from the LLPG for consistency. We use this particular gazetteer for our population estimations because it is provided and used by the local authorities themselves.

	total available dwellings ('000s)	vacant properties ('000s)					vacant properties %				
Local authority	admin 2011	admin 2011 (A)	GLA 2011 [a] (B)	DCLG 2011 [b] (C)	DCLG 2011 [c] (D)	Census 2011 [d] (E)	admin 2011 (A)	GLA 2011 [a] (B)	DCLG 2011 [b] (C)	DCLG 2011 [c] (D)	Census 2011 [d] (E)
Barking and Dagenham	72.9	2.4	0.6	3.6	2.8	3.2	3.3	0.8	4.9	3.8	4.4
Greenwich	112.8	11.2	6.1	13.3	11.4	11.8	9.9	5.4	11.8	10.1	10.5
Hackney	107.9	4.9	9.8	15.8	5.8	6.2	4.5	9.1	14.6	5.4	5.7
Newham	108.9	4.9	5.7	17.1	6.6	7.4	4.5	5.2	15.7	6.1	6.8
Tower Hamlets	115.7	14.5	15.1	17.7	13.6	14.4	12.5	13.1	15.3	11.8	12.4
Waltham Forest	103.1	5.2	8	11.1	5.7	6.2	5	7.8	10.8	5.5	6.0
Total	621.3	43	45.3	78.6	45.9	49.2	6.9	7.3	12.7	7.4	7.9

Table 4.9: Comparison of count and % of vacant dwellings by local authority using administrative, DCLG, GLA and ONS Census data

Table 4.9 shows the differences in vacant property rates between the five sources. These range in value from 0.8 % in Barking and Dagenham using GLA definitions to as high as 15.7 % in Newham using the original DCLG definitions. One reason why rates between the local authorities vary to this extent is due to the exceptionally active regeneration in the Docklands area of east London, affecting mainly Tower Hamlets where there are large numbers of new apartments which were unoccupied at the time.

Another reason is traceable to the lower ONS population counts on which the original 2010 DCLG household counts are based. Without these, in Table 4.9, across the whole area the total vacancy rate only varies by 1 % across the sources, if the original DCLG estimates are excluded. For example, our work in the six Olympic boroughs produced an administrative-based population count of 1.46 m, which is 0.8 % higher than the GLA's but nearly 11 % higher than the equivalent count published by the ONS at that time.

However, this is not an artefact of when data were produced but a systemic problem which can be traced back in time. The London Borough of Newham is a particularly good example of this. At the time of our work in March 2011, the published ONS population for Newham was 240 k compared with our own figure of 299 k. Following revisions to their methodology, the ONS released new figures in November 2011 in which Newham's population had increased from 240 to 272 k.

The final Census 2011 population estimate for Newham is 308 k, a figure created from Census surveys and a number of subsequent adjustments. This may be compared with figures published by the GLA which increased its own estimate for Newham from 268 to 296 k in June 2011, a figure that was partly informed by our own work. The discrepancies between ONS, GLA and administrative sources and also within ONS sources are illustrative of how figures can quickly get out of kilter in areas of high in-migration and regeneration, as the case in Newham but also in neighbouring boroughs.

4.5 Comparison of household types using official figures

The household typology based on the government's own published methodology is shown in Table 4.10 (taken from Department for Communities and Local Government, 2010b). On the face of it, there is no reason why the previously identified discrepancy between sources should impact unduly on our own household typology as long as definitions are comparable even if

the total quantum of households differs. As can be seen in Figure 4.2, the scope of our own typology is richer in detail due to the accounting framework we have created.

As illustration of the differences we focus again on the London Borough of Hackney. This is because it has one of the largest differences in household counts based on our figures and 2010 DCLG's and so provides a rigorous test. We acknowledge that any generalisations concluded from one local authority may not necessarily apply to other types of local authority area or nationally (see check list at Appendix 4.A which provides a summary of quality issues relevant to our research).

Household type	Description
One person households	Male Female
One family and no others [a]	Couple [b]: No dependent [c] children Couple: 1 dependent child Couple: 2 dependent children Couple: 3+ dependent children Lone parent: 1 dependent child Lone parent: 2 dependent children Lone parent: 3+ dependent children
A couple and one or more other adults [d]	No dependent children 1 dependent child 2 dependent children 3+ dependent children
Lone parent and one or more other adults	1 dependent child 2 dependent children 3+ dependent children
Other households [e]	See notes

[a] Households with dependent children and no non-dependent children
[b] 'Couple households' are either married or cohabiting
[c] A dependent child is a person in a household aged 0 to 15 (whether or not in a family) or a person aged 16 to 18 who is a full-time student in a family with parent(s)
[d] In these categories, the other adults may include another couple and/or another lone parent and/ or a non-dependent child
[e] The 'Other households' category above is an aggregation of five categories from the original Census table C1092 supplied by ONS

Table 4.10: The government household typology scheme

For comparison purposes, we recreated DCLG household types using administrative sources. DCLG definitions are quite demanding in terms of their specificity, so reproducing these figures is likely to provide a robust test. The first issue to consider is the definition of a dependent child. For DCLG purposes, this is a person in a household aged 0–15 (whether or not in a family) or a person aged 16 to 18 who is a full-time student (in a family with parents). As was

seen, our primary classification uses age 19 and under and so the first task was to alter our definition of a younger person.

A practical problem was to identify children in full time education (FTE). Our principal data source was the school pupil census, in which persons are flagged if they are aged 16–18 and attended a school in the borough or a neighbouring borough. They are also flagged if they are on Connexions[29] data in which young people are flagged as 'FTE' if they are in full time education. In practice, it could not be determined how many of those not registered as FTE attended private schools or other state schools in neighbouring boroughs.

Households ('000s)					Difference % in households		
Household type	Admin 2011 (A)	DCLG 2011 [a] (B)	DCLG 2011 [b] (C)	Census 2011 [c] (D)	Difference (A-B)	Difference (A-C)	Difference (A-D)
Family household (couple)	20.5	18.2	24.8	28.3	2.30	-4.30	-7.80
Family household (lone parent)	8.6	7.7	8.3	10.8	0.90	0.30	-2.20
Family household with other adults & dependent children	10.0	8.2	9.5	4.2	1.80	0.50	5.80
One person household	47.2	43.5	36.5	35.6	3.70	10.70	11.60
Other households	16.7	14.5	23	22.8	2.20	-6.30	-6.10
Total	103.0	92.1	102.1	101.7	10.90	0.90	1.30

[a] copyright © DCLG, 2010 [b] copyright © DCLG, 2013 [c] copyright © ONS, 2012

Table 4.11: Household type counts in London Borough of Hackney using administrative, DCLG, and ONS Census data, and the % difference of each compared to the administrative data counts

Using administrative data we were able to re-create the 'one person household' category, identified by DCLG as persons aged 16 or over living on their own: however, anyone aged less than 16 living on their own (an extremely rare case and probably anomalous) are considered to be in the 'other category'[30]. The DCLG category 'one family and no others' was also able to be re-created. This includes mixed sex couple households aged 19 and over with or without dependent children or households with only one adult aged 19 and over with dependent children.

The DCLG categories of 'couple' or 'lone parent' with one or more other adults' were combined for simplicity into 'family households with other adults and dependent children' where these households have dependent children. Note that in evaluating the 2010 DCLG

[29] Connexions was a UK governmental information, advice, guidance and support service for young people aged 13 to 19 (up to 25 for young people with learning difficulties and/or disabilities), created in 2000 following the Learning and Skills Act. It is no longer a coherent National Service.

[30] In reality, children under 16 will not be living on their own, but in these cases the resident adults were unable to be captured using administrative data. These cases are small in number.

classification system, the definition of a 'couple' only includes mixed sex, married or cohabiting couples, as defined by the ONS. In other words, DCLG household types tend to be narrower in scope by their omission of same sex households, and also omit three-generational households.

The 2011 Census household types can be grouped in the same way as DCLG categories for comparison purposes, although now same sex couples are considered a 'family' household. Notwithstanding these definitional subtleties, a crude comparison of the results using DCLG 2010 (column B) and 2013 figures (column C), Census 2011 (column D) and administrative sources (column A) is shown in Table 4.11. As can be seen, the totals are generally quite close although the administrative total is marginally higher than the official sources at that time.

Two key findings are that administrative sources enumerate far fewer 'couple family' households and 'other households' than do 2013 DCLG and the Census, but more one-person households and family households with more than two adults. Such differences are due mainly to definitional issues particularly the boundary line between children and adults; however, there are discrepancies between DCLG and the Census, again exposing a lack of consistency between sources.

4.6 Reasons for differences between sources

Other factors have come to light following a recent discussion of the results from the 2012-based projections of households in England as compared with the Census. This review concluded that differing definitions of 'couples' was one of the key issues, but it is also maintained there had been a too slow a reaction to changes in migration (BSPS, May 2015).

Another source, The UNECE (UNECE, 2011), criticises census estimates of single parent households, noting that there can be large differences between one- person households, defined on the basis of a 'housekeeping unit', or a 'dwelling'. This could be another reason for the discrepancy in census counts that use the former definition, and the administrative counts that use the latter which is also reflected in Table 4.11.

Other factors can be speculated for the differences seen such as the effects of the recession and housing crash at the time of the Census which may have suppressed household counts (e.g. Department for Communities and Local Government/Royal Statistical Society meeting, 2013; McDonald and Williams, 2014). All of the above suggests that definitional changes as well as differences in methodology continue to be a serious problem for DCLG.

We understand DCLG is now reviewing the methodology with the aim to make it 'simpler, and more transparent'. It is especially telling, that even after all this work, the most recent DCLG figures are still labelled as provisional because they 'do not yet fully incorporate Census 2011 data' (BSPS meeting, 2015). Overall, the picture therefore remains complex and unsatisfactory.

McDonald and Williams (2014) suggest that it is time for local authorities to consider their own situation carefully and their statistical needs, a view with which we would strongly concur. The root of the problem is not only the lack of coherence in how household statistics are produced, but also a lack of granularity and flexibility over definitions and therefore outputs.

Considering all the difficulties it is re-assuring that our analysis gives us greater confidence that administrative sources are more likely to provide a long-term solution to these issues than continually tweaking present arrangements. We believe this is a strong argument for adopting the approach described in this paper, which gives users more control, greater flexibility and better timeliness.

4.7 Discussion

This paper has identified a gap in the availability, quality and functionality of household statistics at local level. As well as a confusing diversity in sources, household statistics in general suffer from over-aggregation, a lack of flexibility and coherence, coupled with in inability to link to other data except at output area level. Using administrative data, we are able to provide a current enumeration and typology of households with flexible definitions and geography that supports linkage.

We observed that there were a considerable number of administrative data assets available to local authorities for these purposes. This availability creates the conditions for local authorities to develop their own local systems for meeting local needs provided this work is carried out in a data-secure environment, if not alone then in consortia. This would fit with the grain of locally devolved powers and the roles and responsibilities of newly created Health and Wellbeing Boards under the 2012 Health and Social Care Act.

Our approach has been refined in numerous studies in which accuracy, timeliness and specificity of detail were important considerations. This includes the six Olympic borough study referred to in this paper but also in other parts of London and England. So far we have

not attempted to create household projections from administrative data, but we have started to consider household turnover and changes in household type between administrative snapshots.

The current momentum in the UK is to increase the use of administrative data in population analysis and censuses as confirmed by the 2014 ONS announcement (Office for National Statistics, 2014). Administrative data present different problems for users as compared with censuses especially concerning their coverage and reliability (e.g. see Administrative Data Taskforce, 2012; Zhang, 2011). Our view is that their potential is not fully realisable unless they are jointly analysed within a systematic rule based framework. These issues and how to address them are further discussed in (Harper and Mayhew, 2012a; 2012b).

It is useful nevertheless to refresh ourselves on how these issues may propagate through the process of producing household statistics, and if the outputs are sensitive to these. Our administrative data based population and household estimation relies on the linkage of multiple administrative data sources, making it necessary to assess uncertainty and error with great care. This is because no one source captures the whole population, and so rules and assumptions are needed to deal with conflicting information, duplicates, and over or under coverage.

In the absence of comparable methodologies, it is difficult to know if more efficient methods can be devised, but a checklist of quality control issues is contained at Appendix 4.A which may be cross-referred with the analysis and approach taken in this paper. Our methodology is not based on a sampling procedure and as such does not support the use of confidence intervals[31]. We use external comparators as much as possible to back up our results, and the fact that official statistics have fallen into line and validated our results especially after the 2011 Census helps to vindicate this.

There still remain several unsolved issues. For example, boundary issues arise where a person has two addresses but in two separate authorities and so some double counting is possible. In theory, these issues are reconcilable at national level but not if they have addresses in different countries; in other words, no system is perfect. We also mentioned in passing the difficulties of identifying the marital status of households. In a perfect world, this would

[31] Note that DCLG are in a similar position, stating that their 'projection methodology does not enable calculations of probability, standard errors or confidence intervals

require us to link records on marriage and divorce which could pose formidable technical and other challenges.

In terms of transferability, equivalent administrative datasets to those found in the UK would not necessarily be available in exactly the same form elsewhere, although for countries with registration systems instead of censuses such as Finland it should be possible to recreate our typology without difficulty. For countries still using censuses it should also be possible to recreate our typology but probably not at the same level of geographical granularity or with the same flexibility.

To conclude, this paper proposes a different basis for classifying and maintaining household statistics using locally available information sources. The key advantages are that data can be produced on a timely basis in any geography and can be linked to other attributes. This of course requires not only a good knowledge of local data sources but also how to access and exploit them. In our case, access to the data forming the basis for this paper was approved for use by the then local PCTs (Primary Care Trusts) and local councils and underpinned by legally enforceable data sharing protocols including non-disclosure to third parties.

4.8 Appendix 4.A - Measuring statistical quality of population estimation and household counts from administrative data method

ONS provide a checklist of quality measures and indicators for use when measuring and reporting on the quality of statistical outputs. They also record the dimension of quality being measured in each case, using the six European Statistical Service (ESS) Dimensions of Quality developed by Eurostat[32]. In the following table our methodology is compared against each of these standards.

Relevance: The degree to which the statistical product meets user needs for both coverage and content. The administrative data population estimation is an estimate of a local authority's population at a snapshot in time derived from the use and linkage of core local authority administrative datasets and a rule-based system applied to establish who are current residents.

Each person in this count is a separate database entity and assigned to a property address. The summary of the demographic profile of each property address is the basis for the household counts and classification typologies.

[32] Based on ONS Guidelines for Measuring Statistical Quality http://www.ons.gov.uk/ons/guide-method/ method-quality/quality/guidelines-for-measuring-statistical-quality/index.html

The method was developed in response to a need from local authorities to have an alternative source of population statistics that did not rely on survey methods and that was more timely and quicker to produce.

The outputs are used by local authorities who have requested the service to quantify their current population and to inform commissioning, service planning and policy.

The output is created bespoke for a local authority, and is therefore highly relevant to their local context.

Accuracy: The closeness between an estimated result and the (unknown) true value.

The methodology does not enable calculations of probability, standard errors or confidence intervals and therefore these cannot be calculated.

Outputs are compared to available benchmarks to ensure results are sensible.

The methodology uses current data on actual existing residents rather than projections or survey adjustments and imputation. It accounts for no one dataset having complete population coverage by joining separate datasets together to maximize coverage, and for inflation and duplication.

The methodology is inevitably dependent on the accuracy of the input data and is vulnerable to the known issues associated with administrative datasets (noted elsewhere). It is also dependent on the accuracy of the assumptions used in the rule-based system.

No rounding is used in the output.

The outputs are representative of that snapshot in time and may date quickly if the population is in a high state of change.

Timeliness and Punctuality: Timeliness refers to the lapse of time between publication and the period to which the data refer.

The population estimation is carried out at a snapshot in time and results are available typically 2 to 3 months after that time. This is possible because existing data assets are used and no new information is required to be gathered.

Accessibility and Clarity: Accessibility is the ease with which users are able to access the data. It also relates to the format(s) in which the data are available and the availability of supporting information.

Outputs are handed over to the relevant local authority staff with full training, description and metadata. These are available only to nominated and approved staff at the individual and household level due to the potentially identifying nature of the data. Staff can provide aggregate outputs for others if required.

This output is in database format for ease of use in statistical systems.

Outputs are owned and managed by the local authority only and are not otherwise publicly available unless in published form.

Comparability: The degree to which data can be compared over time and domain.

Over time, where a local authority has commissioned population estimations more than once, outputs are comparable in that the same datasets (for that snapshot in time), methodology and assumptions are used each time. Input datasets are kept consistent as much as possible but are vulnerable to known change issues associated with administrative datasets (noted elsewhere).

Geographically, outputs between the local authorities that have outputs available are comparable in the same way as over time, mentioned previously. It has not been quantified if the datasets and assumptions are biased by different types of local area, therefore caution should be used in this respect.

Outputs are not available for every local authority area in England and Wales and are therefore not nationally comparable.

Coherence: The degree to which data that are derived from different sources or methods, but which refer to the same phenomenon, are similar.

Outputs are compared to available benchmarks to ensure results are sensible. Population estimations from other sources may differ due to different methodologies, assumptions and definitions used and care should be taken when making comparisons.

---

5 Impact of Asthma on Educational Attainment in a Socioeconomically Deprived Population: A Study Linking Health, Education and Social Care Datasets

Preface: Content in this chapter consists of an exact reproduction of the article published in the journal PLOS ONE in 2012. Only minor edits have been made to make numbering consistent throughout the thesis. As such there may be some dated references or statements. Developments in population data science since the time of publication of this paper are described in Chapter 6.

Funding: This study was supported by Asthma UK (Health Charity) project 05/048. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

5.1 Introduction

Asthma is the commonest long-term disorder affecting children in the UK and most economically-developed countries (Asher et al., 2006). Health status and education of children are closely linked (Bush and Saglani, 2010). In a recent review on asthma in children (Kleinert, 2007), no mention was made of asthma's impact on educational performance, an important omission given the increasing recognition for a move to assessing impact of long-term conditions on patient/parental-centred outcomes. There is debate about the possible effect of asthma on children's educational performance, with studies producing conflicting results, some finding an adverse effect (Eagan et al., 2004; Ellison-Loschmann et al., 2007; Austin et al., 1998) some no effect (Milton et al., 2004; Anderson et al., 1983; Silverstein et al., 2001; Fowler et al., 1992), and others a beneficial effect (O'Neil et al., 1985; Gutstadt, 1989). Taras (Taras and Potts-Datema, 2005) proposed asthma was related to poor exam performance and called for evaluations focusing on populations at increased risk.

Populations at particular risk are children of low socioeconomic status and children of south Asian and Black ethnic minority origin: both experience increased asthma morbidity and poor educational attainment (Netuveli et al., 2005; Office for National Statistics, 2009). However, no study has examined the impact of asthma on school examination performance in children from large ethnically diverse socio-economically deprived populations. Results of such a study would help guide policy by identifying and then targeting potentially modifiable factors that relate to poor educational attainment.

We therefore tested the hypothesis that asthma worsens educational attainment in children from socio-economically deprived, multiethnic populations. Tower Hamlets, in east London, is the UK's third most deprived borough (Communities and Local Government). Its population is largely White or Bangladeshi, providing an ideal setting for this study. The study received Local Research Ethics Committee approval.

5.2 Methods

5.2.1 Study participants

Fifteen general practices in Tower Hamlets were approached and 14 participated in this cross-sectional study covering 1st July 2001 to 30th June 2005. These practices provided care for almost 50% of the borough's children. Inclusion criteria use of the EMIS computer system (one of the main software suppliers to UK general practices) (Egton Medical Information Systems Limited, 2007), and a list of over 5,000 patients (to maximise our dataset). We undertook MIQUEST (Morbidity Information QUery and Export SynTax) searches on all patients aged under 20 on 30 June 2005. For confidentiality, clinical data were collected separately from administrative data. We included all who had sat at least one national Key Stage 1–3 attainment test between 2002 and 2005 (Table 5.1), identified on the Annual School Census (ASC), which includes all children attending Tower Hamlets state schools. National Key Stage tests assess children in England, Wales and Northern Ireland against the content of the National Curriculum, and so provide a standardised comparison of academic performance. Details of the children included are given in Table 5.2.

Key stage	School year	Approx. pupil age	Topics assessed	Maximum mark attainable
Key stage 1	Year 2	7 years	English, maths, science	30
Key stage 2	year 6	11 years	English, maths, science	100
Key stage 3	Year 9	14 years	English, maths, science	150

Table 5.1: Key stage tests for the UK's National Curriculum

5.2.2 Outcome variable (Table 5.2)

The outcome variable was level of individual attainment at Key Stages 1, 2, and 3. Children who sat more than one Key Stage test contributed more than one observation to the data. For Key Stage 1 the mean score of reading, writing and mathematics and for Key Stages 2 and 3 the mean score of English, mathematics and science was calculated for each pupil for 2002 to 2005.

5.2.3 Predictor variables (Table 5.2)

Our primary predictor was asthma status coded in three groups: 1) no diagnosis of asthma; 2) diagnosis of asthma with one or more bronchodilator prescriptions in the relevant year ('active asthma'), and 3) diagnosis of asthma but no bronchodilator prescription in the year before July 1st of the year in which the Key Stage test was sat ('inactive asthma'). Data on asthma status were obtained from practice records. We used the H33 Read code allocated by the practice to identify children with asthma. In two practices where coding of asthma was poor, any child receiving repeat prescriptions of long acting bronchodilators and/or inhaled corticosteroids was allocated an asthma code, as were those with six or more prescriptions of short acting bronchodilators over a four-year period.

Asthma severity was assessed by using British Thoracic Society (BTS) medication step as a proxy. We assigned each child a BTS step by examining their prescriptions in the final year of the study (British Thoracic Society, 2007).

Asthma control was assessed by estimating numbers of short-acting bronchodilator devices prescribed per child per year as a proxy. Whilst MIQUEST searches do not provide the actual number of inhalers prescribed on a given prescription, we examined the individual prescriptions of 50 randomly selected asthmatic children in each practice to find the average number of short-acting bronchodilator inhalers prescribed per prescription and thus estimated the number of devices prescribed per child.

Ethnicity was obtained from ASC data and was categorised into three groups: 1) White or 'ethnicity not recorded', 2) Bangladeshi, and 3) 'other ethnicity'. Coding of Tower Hamlets minority ethnic group schoolchildren is almost 100% complete; we assumed the few children with 'ethnicity not recorded' (1.6% of total) were likely to be White, and checked this in sensitivity analyses.

Social adversity was measured using socio-economic information at the household level from the Housing Department of the London Borough of Tower Hamlets: the Council Tax band of the child's residence (defined by value of dwelling on 1st April 1991), whether it was of 'social housing' tenure, if anyone at that address was in receipt of Council Tax or Housing Benefit and if it was a 'one adult' household. The ASC provided information on eligibility for free school meals, special educational needs, exclusion from school, name and type of school attended

(mixed or single sex). Practice records provided data on presence of a smoker in household, number of smokers, and possible co- morbidities including allergic rhinitis, eczema, glue ear, diabetes, chronic tonsillitis and mental health problems.

5.2.4 Data linkage

We extracted an 'Administrative File' from practice computers consisting of names and addresses of children and a discrete Patient ID number derived from a randomly-generated practice number concatenated with the patient's EMIS number. A single merged Administrative File from all the participating practices formed the foundation of the Master Dataset (in Microsoft Access) onto which socioeconomic, educational, and finally clinical data were merged. The procedure by which we linked data was as follows:

i) Assigning socio and demographic data using matching by address. We assigned each address on the Master Dataset a Unique Property Reference Number (UPRN) from the Local Land and Property Gazetteer. This overcame the problem of differently formatted versions of the same address in different datasets. Using the UPRN, we then linked socio-economic information at the household level to the Master Dataset, including data from Council Tax Banding and Council Tax Benefit datasets. Using address, we also linked data from general practices on which households had smokers.

ii) Assigning school attainment data. We ascribed each pupil a unique pupil number (UPN, derived from the Annual School Census) and added these to the Master Dataset using matching by name, date of birth, UPRN, and postcode. This was an iterative process requiring detailed checks at each stage. Matching difficulties arose, for example, where first names were shortened or incorrectly spelt, surnames given with middle names, dates of birth incorrectly entered or postcodes changed. Un-matched records were subjected to successive relaxations of the criteria and subsequent resultant matches were scrutinised for errors. Once the matching process was complete, we identified possible twins using duplicate surnames and dates of birth. We checked these records manually to ensure we had assigned the correct ASC data to the correct twin.

School attainment data	variable	n	%/SD/range
number of pupils with results only for	KS 1	4,074	33.6%
	KS 2	3,720	30.7%
	KS 3	3,360	27.7%
number of pupils that took two KS tests	KS 1 & 2	16	0.1%
	KS 2 & 3	966	8.0%
attainment (English): mean (SD)	KS 1	13.7	0.0
	KS 2	52.5	0.2
	KS 3	34.9	0.2
attainment (Maths): mean (SD)	KS 1	15.2	0.0
	KS 2	60.7	0.2
	KS 3	66.9	0.2
attainment (overall): mean (SD)	KS 1	14.2	0.0
	KS 2	56	0.2
	KS 3	60.8	0.2
special educational needs (at KS)	KS 1	807	0.2
	KS 2	1,117	0.2
	KS 3	971	0.2
Clinical data			
asthma diagnosis (ever)		2,206	18.2%
asthma treatment	inactive	1,125	9.3%
	active	946	7.8%
bronchodilators: 50th (10th–90th) centiles	3 months [a]	1.8	1.3 to 3.6
	12 months [b]	3.4	1.4 to 8.0
BTS Step in 2005	1	767	6.7%
	2	450	3.9%
	3+	153	1.3%
comorbidities	any atopic disease	3,483	28.7%
	allergic rhinitis	1,266	10.4%
	eczema	2,592	21.4%
	mental health problems	521	4.3%
	diabetes	19	0.2%
smoking household	yes	6,679	55.0%
	1 smoker	4,474	67.0%
	2 smokers	1,572	23.5%
	3+ smokers	633	9.5%
Socio demographic data			
sex	males:	6,091	50.2%
ethnicity	White/NA	3,216	26.5%
	Bangladeshi	7,126	58.7%
	other	1,794	14.8%
Council Tax band*	A: <=£40 k	29	0.3%
	B: >£40 k to <=£52 k	2,533	21.6%
	C: >£52 k to <=£68 k	5,856	49.9%
	D: >£68 k to <=£88 k	1,392	11.9%
	E: >£88 k to <=£120 k	1,314	11.2%
	F: >£120 k to <=£160 k	544	4.6%
	G: >£160 k to <=£320 k	67	0.6%
	H: >£320 k	0	0.0%
in receipt of benefits		8,646	71.2%
living in social housing		8,188	67.5%
one adult household		1,370	11.3%
free school meals		7,607	62.7%

KS 1 sat aged 7, KS 2 sat aged 11, KS 3 sat aged 14.

SD = standard deviation.

range = 10th to 90th percentile.

BTS = British Thoracic Society.

active asthma = bronchodilator in period 12 months before Key Stage test.

inactive asthma = no bronchodilator in period 12 months before Key Stage test.

[a] =in period 3 m prior to KS test.

[b] = in period 12 m prior to KS test.

*Council Tax band based on property valuation on 1st April 1991.

NA = not available.

Table 5.2: Characteristics of 12,136 pupils that sat Key Stage tests in 2002 to 2005.

All children on the Master Dataset were found their corresponding match on the 2002 to 2005 Annual School Census data – none remained unmatched. If a child did not exist on Census data, this indicated they did not attend a Tower Hamlets state school during this period, and no pupil or Key Stage data could be assigned to them. They were therefore excluded from the analysis.

iii) Pseudonymising the database for confidentiality. Until this point the Master Dataset contained patient identifying administrative details. To pseudonymise the dataset we then removed these to leave only the discrete Patient ID number, age and sex.

iv) Assigning general practice clinical and prescribing data. Our clinical and prescribing data from MIQUEST searches included the discrete Patient ID number but no other patient details. Using this discrete ID number, we then matched and attached the clinical and prescribing data to the pseudo anonymised database, thus retaining confidentiality.

5.2.5 Statistical methods

Sample size. We anticipated 20,000 children would have analysable data, 20% of whom would have asthma. Allowing for clustering of attainment by school, and assuming 100 children per school and an intra-cluster correlation coefficient (ICC) (for the difference in attainment between children with and without asthma) of 0.05, gives, working back, a simple sample size of 3,361 children using the formula $N_{simple} = N_{cluster}/(1+(100-1)*ICC)$. With 80% power at the 5% significance level, this would allow us to detect a difference in standardised overall attainment of 0.12 standard deviations between 672 children with asthma and 2689 children without.

Analysis. Analysis was carried out in Stata 10.1 (StataCorp, 2007) and involved regression of attainment on the predictor variables and potential confounders. To maximise power we combined all years and all Key Stages. Attainment could be affected by whether a child is old or young for their year. We therefore included a variable which reflected this: age was

standardised within each academic year to have a mean of 0 and SD of 1. Assuming that Key Stage tests varied in difficulty between years, attainment was also standardised within each academic year. Standardised attainment was regressed, in a linear model, on standardised age, and several socio-demographic and clinical variables pertaining to asthma and co-morbid conditions. We used data reduction techniques to decide on the variables to consider for this multiple regression, including removing binary variables which would not be discriminatory, and removing socio-demographic variables for which P.0.1 in bivariate regression models. The final selection of variables was: asthma control in the 12 months prior to the Key Stage test, ethnicity, standardised age, sex, living in a smoking household, living in a property in Council Tax bands A, B or C, living in social housing, in receipt of Council Tax/housing benefit, in receipt of free school meals, having special educational needs, diagnoses of allergic rhinitis, eczema, and mental health problems. In primary analysis asthma control was noted in the 12 months prior to Key Stage tests. We fitted a multiple linear regression model with robust standard errors using White's sandwich estimator, implemented by specifying school as the clustering variable (White, 1980). This overcame the underestimation of the standard errors due to the tendency for children within schools to be more alike than children between schools (intra-cluster correlation).

The percentage change in test scores for a unit change in a regression coefficient are given in Appendix 5.A column 2. These were obtained using the following multiplication factors (KS1: 12.3, KS2: 16, KS3: 12.5), which were calculated from SD (KS test)x100/max possible score for KS test).

We conducted six separate sensitivity analyses to investigate the effect of varying assumptions on our model estimates (Appendix 5.A). These were: defining inactive asthma as having no bronchodilator for the last three (instead of twelve) months, excluding a more recent Key Stage test result where a pupil had sat two tests to remove the interdependence of test results for the same child, using three different ways of grouping ethnicity and not imputing an asthma diagnosis in 79 children registered at two practices.

Finally, for the subset of children for whom BTS step could be ascertained in the final year of the study, we fitted the same model as in the primary analysis except that we first stratified our data by active or inactive asthma, and included an indicator variable for BTS Step 2, 3, 4 or 5 versus 1 and a variable giving the estimated average number of bronchodilator prescriptions in the 12 months prior to the Key Stage test.

Barking and Havering Local Research Ethics Committee gave full ethical approval on 6th October 2005 for the study methodology (REC reference number: 05/Q0602/83). The Research Ethics Committee did not require written consent to be given by the patients next of kin, carers or guardians on the behalf of the minors/children for their information to be stored in the hospital database and used for research. A STROBE checklist is provided as a supplement.

5.3 Results

Our search identified 30,841 children aged 0–19 years old registered with the 14 study practices. Of these, 20,683, were identified on the Tower Hamlets Schools Census Data, i.e. they had attended a Tower Hamlets school during the study period, the remainder (10,158) being too young to attend school. Of the 20,683, we retained those who were aged 5–14 years old and had sat at least one Key Stage test. This produced a database that contained observations on 12,136 children, from 97 schools, who had sat at least one of Key Stage tests 1, 2 or 3 (Table 2). 2,206 (18.2%) children had a diagnosis of asthma. The majority of children (1,370) that had BTS step recorded had mild asthma (767 children) (step 1). A third were at step 2 and the remainder at step 3 or above. Allergic rhinitis was recorded in 10.4%, eczema in 21.4% and atopy in 28.7%. Mental health problems (largely behavioural and developmental) were diagnosed in 4.3% of children.

58.7% of children were Bangladeshi, with 26.5% White or 'ethnicity not recorded' (1.6% of all children) and 14.8% 'other ethnicity'. Most children lived in socioeconomically deprived circumstances. Council Tax bands A, B or C (the lowest valued dwellings) were recorded against 71.8% of addresses. Almost three quarters of children were from families receiving housing and/or Council Tax benefit. Just over two thirds of children lived in social housing and 11.3% were from households with one resident adult. A majority (55.0%) of children lived in smoking households.

The estimated median number of inhaler devices prescribed in the three month period pre Key Stage test was 1.8 (10th–90th% spread 1.3–3.6). In the 12 months prior to the Key Stage test, the median number of devices prescribed was 3.4 (1.4–8.0). 21.5% of children with asthma had a prescription for a bronchodilator in the three months prior (recent asthma control) to the school examination, rising to 45.1% in the period 12 months prior to their examination.

	simple regression			multiple regression [c]		
independent variable	**β**	**95% CI**	**p-value**	**β**	**95% CI**	**p-value**
[a] inactive asthma vs. no asthma	0.047	(-0.011 to 0.105)	0.11	0.023	(-0.025 to 0.071)	0.35
[b] active asthma vs. no asthma	0.053	(-0.007 to 0.112)	0.08	0.066	(0.013 to 0.119)	0.02
Bangladeshi vs. White	-0.07	(-0.155 to 0.016)	0.11	-0.082	(-0.157 to -0.007)	0.03
other ethnicity vs. white	0.052	(-0.021 to 0.125)	0.16	0.008	(-0.054 to 0.071)	0.79
standardised age	0.086	(0.064 to 0.108)	<0.001	0.053	(0.034 to 0.071)	<0.001
girls vs. boys	0.108	(0.067 to 0.149)	<0.001	0.003	(-0.042 to 0.048)	0.89
smoking household (Yes vs. No)	-0.132	(-0.172 to -0.093)	<0.001	-0.072	(-0.107 to -0.037)	<0.001
Council Tax band (A–C vs. D–H)	-0.099	(-0.146 to -0.052)	<0.001	-0.059	(-0.094 to -0.025)	0.001
living in social housing (Yes vs. No)	-0.144	(-0.185 to -0.104)	<0.001	-0.047	(-0.082 to -0.011)	0.01
in receipt of benefits (Yes vs. No)	-0.236	(-0.293 to -0.179)	<0.001	-0.115	(-0.163 to 0.066)	<0.001
free school meals (Yes vs. No)	-0.192	(-0.242 to -0.142)	<0.001	-0.051	(-0.093 to -0.008)	<0.001
special educational needs (Yes vs. No)	-0.943	(-1.013 to -0.872)	<0.001	-0.912	(-0.981 to -0.842)	0.02
allergic rhinitis diagnosis (Yes vs. No)	0.056	(0.005 to 0.106)	0.03	0.022	(-0.022 to 0.067)	0.32
eczema diagnosis (Yes vs. No)	0.041	(-0.003 to 0.085)	0.07	0.018	(-0.025 to 0.061)	0.41
mental health problems (Yes vs. No)	-0.281	(-0.407 to -0.155)	<0.001	-0.154	(-0.256 to -0.052)	0.003
constant term				0.533	(0.421 to 0.645)	<0.001

β = regression coefficient.

[a] No prescription for a bronchodilator in the 12 months before the Key Stage test, but asthma diagnosis ever.

[b] At least one prescription for a bronchodilator in the 12 months before the Key Stage test and asthma diagnosis ever.

[c] model includes all variables listed.

The intra-school correlation coefficient in the multiple regression model is 0.06, 95% CI (0.04 to 0.08)

Table 5.3: Coefficients, 95% confidence intervals and P-values for the effect of socio demographic and clinical variables on standardised attainment scores in Key Stage tests 1, 2 and 3, from multiple regression model allowing for clustering

5.3.1 Primary analysis

Analysis of attainment for Key Stages 1, 2 and 3 combined is presented in Table 5.3. For ease of interpretation, we have presented these results in the abstract as percentage changes in examination scores for Key Stage 2 – the most commonly sat examination in the dataset. Regression coefficients are presented here (for percentage changes in scores for all Key Stages

see Table 5.3). As the outcome is standardised overall attainment, the interpretation of a particular regression coefficient is, for all other variables fixed, the proportion of a standard deviation change in attainment given a one unit change in the predictor variable of interest. As age was also standardised, a one unit change in age is equivalent to a one SD change in age (approx. 2.8 years). 22% of the variability in standardised overall attainment was explained by the model.

Asthma status. A weak positive association was found between overall school examination attainment and having active asthma (asthma treated with a bronchodilator during the past 12 months): β=0.066 (95% CI 0.013 to 0.119). No association was found for inactive asthma: β=0.023 (95% CI -0.025 to 0.071). For these groups (active and inactive asthma) combined, a weak positive association was found: β= 0.04 95% CI (0.00 to 0.08).

Ethnicity. Bangladeshi children did significantly worse in the Key Stage tests than White children β=-0.082 (95% CI -0.157 to -0.007), though there was no evidence of a difference between children of 'other' ethnicity and White children: β= 0.008 (95% CI -0.054 to 0.071).

Social adversity. Statistically significant associations were observed between attainment and several socio-demographic variables. Children from smoking households: β=-0.072 (95% CI -0.107 to -0.037), living in social housing: β=-0.047 (95% CI -0.082 to -0.011), in receipt of housing/Council Tax benefit: β=-0.059 (95% CI20.094 to -0.025), and receiving free school meals: β=20.051 (95% CI -0.093 to -0.008) did significantly worse in the tests than children from less deprived households. Children identified as having special educational needs: β=-0.912 (95% CI -0.981 to -0.842) or with mental health problems: β=-0.154 (95% CI -0.256 to -0.052) also performed significantly worse. Girls did not score significantly higher than boys: β=0.003 (95% CI -0.042 to 0.048). Children that were older for their year did better than younger children: β= 0.053 (95% CI 0.034 to 0.071). There was no association between eczema: β=-0.018 (95% CI -0.025 to 0.061) or allergic rhinitis: β=-0.022 (95% CI -0.022 to 0.067) and overall attainment.

Asthma severity and attainment. We examined whether children with more severe asthma might have poorer examination scores. Table 5.4 gives the adjusted effects of asthma severity (BTS step) and other predictor variables on standardised attainment from two separate multiple regression models (active asthma: 882 children, inactive asthma: 462 children). We found no association between BTS step and attainment in children. Furthermore,

bronchodilator use in those with active asthma was not associated with attainment in the Key Stage tests.

	asthma	β	95% CI for β		P-value
BTS Step 2,3,4 or 5 vs. 1	inactive	-0.098	-0.259	0.062	0.23
	active	-0.003	-0.097	0.091	0.95
Bronchodilator use (last 12 months)	inactive	NA	NA	NA	NA
	active	0.000	-0.017	0.017	0.98
Bangladeshi vs. White	inactive	-0.110	-0.291	0.071	0.23
	active	0.044	-0.082	0.169	0.49
other ethnicity vs. white	inactive	0.022	-0.125	0.169	0.76
	active	0.125	-0.019	0.269	0.09
standardised age	inactive	0.095	0.019	0.171	0.01
	active	0.079	0.022	0.136	0.007
girls vs. boys	inactive	0.057	-0.081	0.194	0.42
	active	-0.086	-0.177	0.006	0.07
smoking household (Yes vs. No)	inactive	-0.127	-0.253	-0.001	0.05
	active	-0.020	-0.136	0.095	0.73
Council Tax band (A–C vs. D–H)	inactive	0.038	-0.091	0.168	0.56
	active	0.052	-0.101	0.205	0.5
living in social housing (Yes vs. No)	inactive	-0.029	-0.206	0.149	0.75
	active	-0.045	-0.195	0.105	0.55
in receipt of benefits (Yes vs. No)	inactive	-0.003	-0.172	0.165	0.97
	active	-0.097	-0.245	0.051	0.2
free school meals (Yes vs. No)	inactive	-0.190	-0.330	-0.049	0.009
	active	-0.076	-0.185	0.043	0.207
special educational needs (Yes vs. No)	inactive	-0.880	-1.067	-0.689	<0.001
	active	-1.024	-1.210	-0.838	<0.001
allergic rhinitis diagnosis (Yes vs. No)	inactive	-0.038	-0.185	0.108	0.6
	active	-0.004	-0.130	0.122	0.95
eczema diagnosis (Yes vs. No)	inactive	-0.076	-0.228	0.075	0.32
	active	0.063	-0.042	0.168	0.24
mental health problems (Yes vs. No)	inactive	-0.143	-0.562	0.276	0.5
	active	0.197	-0.036	0.429	0.1
constant term	inactive	0.593	0.239	0.947	0.001
	active	0.572	0.337	0.806	<0.001

Table 5.4: Table 5.3: Coefficients, 95% confidence intervals and P-values for the effect of socio demographic and clinical variables on standardised attainment scores in Key Stage tests 1, 2 and 3, from multiple regression model allowing for clustering

5.4 Discussion

5.4.1 Summary

We found no evidence for an adverse effect of asthma or asthma severity on examination performance in this large cross sectional study of children from a highly socio-economically deprived, multiethnic area. There was a very small positive association between attainment

and having active asthma in the 12 months prior to the Key Stage test versus having no asthma. With respect to asthma these findings are reassuring for parents and teachers alike.

Examination performance was poorer in Bangladeshi children, and those experiencing social adversity (eligibility for free school meals, living in social housing, 'one parent' households and households with a smoker), those with mental health problems and special educational needs, suggesting that these should be the foci of policies to improve educational attainment.

5.4.2 Strengths and weaknesses

Strengths include use of large unbiased datasets in a locality with characteristics ideally suited to testing the study hypothesis: high asthma prevalence, socioeconomic deprivation and ethnic diversity with two groups (White and Bangladeshi) predominating with accurate and near complete coding of ethnicity. Excellent participation by local general practices allowed us to capture data on half the children in the borough sitting school examinations. A sophisticated methodological approach allowed us to link general practice, housing and educational authority databases, enabling us to address a wide range of clinical and socio-demographic factors, and importantly, to explore relationships between clinical factors and social outcomes. We believe this is the first time these disparate data sources have been merged. Our ability to match Unique Property Reference Numbers and Unique Pupil Numbers to general practice administrative data enabled us to identify households of those living in poor socio-economic circumstances, households with smokers, and to identify ethnicity, those receiving free school meals and children with special educational needs, as well as providing examination results. This methodology provides a useful tool for further epidemiological research.

Our finding that aspects of social adversity were associated with poor educational outcomes is in keeping with previous work, see Richards (Richards and Wadsworth, 2004), and is a good indication of the validity of our approach and the linked dataset generated. These latter variables are more plausible influences on school exam performance than asthma. Data on absenteeism may have helped to elucidate our findings as school absence in children with asthma has been shown to vary with ethnicity and housing conditions (Parcel et al., 1979; Free et al., 2010). It was disappointing that some markers of asthma severity including peak flow recording and asthma consultations were poorly recorded in general practice. Relatively few children with very severe asthma meant it is difficult to generalise our findings to that population. Extrapolating BTS step for a single year relied on the assumption that severity for

each child was stable over the study duration. We did not use mediating and/or moderating models in our analysis, and it is possible that at some level of social disadvantage having asthma might have an adverse impact. Possible explanations for our finding of a very small positive association between asthma and attainment in the 12 months prior to a Key Stage test are speculative but include a chance effect, a selection effect (for example, asthma occurring in brighter children), or a behavioural effect (children with asthma studying or performing in tests in a different way).

5.4.3 Comparison with other studies

Previous studies fall into three categories: (i) retrospective studies measuring educational level reached in adults with asthma (Eagan et al., 2004; Ellison-Loschmann et al., 2007), (ii) case-control studies investigating results of examinations in children with and without asthma (Anderson et al., 1983; Silverstein et al., 2001), (iii) comparison of examination results of children with asthma and general population data (O'Neil et al., 1985; Gutstadt, 1989). Two retrospective studies have linked a diagnosis of asthma with lower levels of education (Eagan et al., 2004; Ellison-Loschmann et al., 2007) and two studies found a positive association with above average scores (O'Neil et al., 1985; Gutstadt, 1989). Our study is consistent with the majority of other studies in not finding an association between poor attainment and asthma (Anderson et al., 1983; Silverstein et al., 2001; Fowler et al., 1992). Our results could reflect well-controlled asthma (Clark et al., 1984). One study (Austin et al., 1998) found that children with severe asthma performed worse in examinations - an association we did not find, even examining children at BTS step 3.

5.4.4 Clinical and policy relevance

Our work suggests the important drivers of poor performance relate not to asthma, but to ethnicity, social adversity, mental health problems and special educational needs. These should be the foci of policies to improve educational attainment. Pakistani and Bangladeshi children are amongst the lowest achieving groups in the UK (West et al., 1992). Reasons for their poorer performance could be related to language difficulties and absenteeism. Two American studies have shown that black children miss more school than white (Taylor and Newacheck, 1992; Diette et al., 2000). Social adversity was significantly associated with lower test scores. Biological and psychosocial mechanisms might explain this relationship (Richards and Wadsworth, 2004). While children from lower socio-economic backgrounds experience more morbidity (Spencer, 2000) including asthma (Mielck et al., 1996; Baker et al., 1998;

Duran-Tauleria and Rona, 1999; Kitch et al., 2000) this finding was independent of asthma diagnoses and ethnicity. It is plausible that children with special educational needs and those with a history of mental health problems fared worse in tests, especially as behavioural difficulties and developmental problems made up a substantial proportion of the latter.

Our classification of asthma severity by medication step might suggest that most children have mild asthma. This interpretation warrants caution. East London has the highest rates of hospital admission for respiratory illness in children in London, much of which is likely to be asthma related. Future work should explore possible under-treatment in this locality.

5.4.5 Conclusion

Our results provide the first large scale assessment of the relationship between asthma, and educational attainment in an ethnically diverse and socioeconomically deprived population. Our results are reassuring with respect to effects of asthma on educational attainment, but provide reason for policymakers to prioritise social adversity and mental health problems as drivers of poor exam performance. Our data linkage methodology has the potential to be applied in other areas as a means to inform health care policy by quantifying links between demography and clinical and social outcomes.

5.4.6 Acknowledgments

We are grateful to the partners and staff of the general practices for participating in the study. We thank the London Borough of Tower Hamlets for providing ASC and LLPG data, and the East London Common Information Service for the General Practice Register. CG is guarantor and is responsible for the integrity of the work.

5.5 Appendix 5.A - Sensitivity analyses

The appendix table presents regression coefficients and their standard errors for six separate models to assess the sensitivity of the regression coefficients for the asthma control variables

to different specifications of ethnic groups, time window defining active asthma (prescription of a bronchodilator within either three or twelve months prior to Key Stage test), having results for one or two Key Stage tests and removing the children for whom a diagnosis of asthma had been imputed from their prescription history. The first column shows the primary model presented in Table 5.2. The estimates are not particularly sensitive to different model specifications and the overall percentage of the variation in attainment explained was 22% in every case. There was no association with short-term control of asthma and overall attainment. To satisfy the linear regression assumption of independent errors, the second test sat by children with results for two Key Stage tests was dropped from the data set (1,002 observations). The regression coefficient comparing attainment in children with treated asthma to children with no asthma diagnosis was 0.069 and significant at the 5% level. The three models specifying different ethnicity groupings (A, B and C) led to very similar conclusions for the association between treated and untreated asthma and overall attainment. In conclusion, we found some evidence of a small (approx. 0.07 SDs) positive association between having asthma and at least one bronchodilator prescription in the 12 month period prior to a Key Stage test, and overall attainment in the test. There were 79 children for whom a diagnosis of asthma was imputed on the basis of their prescribing history alone. Analysing them as children without asthma instead made virtually no difference to the results.

Appendix 5.A. Sensitivity of effects on standardised attainment to a variety of model specifications. The intra-school correlation coefficient in each of these models is estimated to be 0.06 95% CI (0.04 to 0.08).

sensitivity analysis	12m control		3m control		1st test only		ethnicity A		ethnicity B		ethnicity C		no imputed asthma	
R-squared = 0.22 for each model	13,117 obs		13,117 obs		12,115 obs		12,890 obs		13,117 obs		13,117 obs		13,038 obs	
variable	β	se(β)	β	se(β)	β	se(β)	β	se(β)	β	se(β)	β	se(β)	β	se(β)
inactive asthma vs. no asthma	0.023	0.024	0.047	0.022	0.033	0.024	0.018	0.024	0.023	0.024	0.031	0.025	0.023	0.024
active asthma vs. no asthma	0.066	0.027	0.019	0.039	0.069	0.028	0.065	0.027	0.066	0.027	0.069	0.027	0.070	0.027
ethnicity 2 vs. 1	-0.082	0.038	-0.081	0.038	-0.091	0.038	-0.092	0.039	-0.086	0.039	-0.060	0.034	-0.083	0.038
ethnicity 3 vs. 1	0.008	0.031	0.008	0.031	0.007	0.033	0.000	0.034	-0.003	0.033	NA	NA	0.006	0.031
standardised age	0.053	0.009	0.052	0.009	0.060	0.008	0.051	0.009	0.053	0.009	0.053	0.009	0.053	0.009
girls vs. boys	0.003	0.023	0.003	0.023	0.014	0.025	0.001	0.023	0.003	0.023	0.004	0.022	0.057	0.069
smoking households (Yes vs. No)	-0.072	0.018	-0.072	0.018	-0.070	0.018	-0.073	0.018	-0.072	0.018	-0.078	0.018	-0.070	0.018
council tax bands A-C vs. D-H	-0.593	0.017	-0.059	0.018	-0.058	0.018	-0.060	0.018	-0.059	0.017	-0.062	0.018	-0.059	0.017
living in social housing (Yes vs. No)	-0.047	0.018	-0.046	0.018	-0.045	0.019	-0.048	0.018	-0.047	0.018	-0.042	0.018	-0.047	0.018
in receipt of benefits (Yes vs. No)	-0.115	0.024	-0.114	0.024	-0.126	0.026	-0.111	0.024	-0.114	0.024	-0.125	0.025	-0.114	0.025
free school meals (Yes vs. No)	-0.051	0.022	-0.051	0.022	-0.052	0.021	-0.050	0.022	-0.051	0.022	-0.052	0.022	-0.050	0.022
special education (Yes vs. No)	-0.912	0.035	-0.911	0.035	-0.918	0.036	-0.912	0.035	-0.912	0.035	-0.910	0.035	-0.914	0.035
allergic rhinitis (Yes vs. No)	0.022	0.022	0.024	0.022	0.030	0.024	0.023	0.023	0.023	0.022	0.024	0.022	0.023	0.022
eczema (Yes vs. No)	0.018	0.022	0.018	0.022	0.013	0.022	0.015	0.022	0.018	0.022	0.022	0.022	0.019	0.022
mental health (Yes vs. No)	-0.154	0.051	-0.155	0.051	-0.170	0.052	-0.150	0.052	-0.155	0.051	-0.150	0.051	-0.143	0.211
constant	0.533	0.056	0.533	0.057	0.528	0.058	0.545	0.058	0.537	0.057	0.537	0.057	0.533	0.057

Key

β = regression slope, se(β) = standard error of regression slope
12m control = bronchodilator prescription in 12 months prior to Key Stage test
3m control = bronchodilator prescription in 3 months prior to Key Stage test
In the first three (models) columns, ethnicity 2 vs. 1 refers to Bangladeshi vs. White/unknown, and ethnicity 3 vs. 1 refers to 'other ethnic groups' vs. White/unknown
In the ethnicity A model, ethnicity 2 vs. 1 refers to Bangladeshi vs. White, and ethnicity 3 vs. 1 refers to 'other ethnic groups' vs. White (unknown excluded)
In the ethnicity B model, ethnicity 2 vs. 1 refers to Bangladeshi vs. White, and ethnicity 3 vs. 1 refers to 'other ethnic groups'/unknown vs. White
In the ethnicity C model, ethnicity 2 vs. 1 refers to Bangladeshi/other vs. White/unknown.
NA = not applicable
No imputed asthma: despite evidence of asthma in the medication history in the absence of a Read code, a diagnosis was not imputed.

6 Conclusions

In this thesis, it is demonstrated how routinely collected administrative data can be linked together and rules applied to them to create an estimated count of local populations. This is at the level of individual persons and households with related socio-economic and service use variables, providing a detailed profile of the population. As demonstrated by my papers and the large number of consultancy assignments undertaken over a long period, linked administrative data coupled with systematic approaches for analysing the linked data adds significant value to the datasets, which were originally collected for other purposes.

This approach has been compared to the methods, results and resources of the traditional census of population in England and Wales, and is suggested as a feasible alternative to this with lower costs, faster turnarounds, and more granular and flexible outputs. This is in the context of a time when budgets are being cut and government departments are having to do more with less, as well as an increasing awareness of and move towards exploiting routinely collected administrative data, and big data generally.

This approach is found to be especially suited to the needs of local authority evidence-based policy and service planning, where the trade-off between statistics of national statistics standard and statistics that are produced rapidly at higher granularity is more acceptable.

The four papers in chapters 2 to 5 cover a natural progression from how demand for the method and outputs came about, a description of the core methodology, and to a range of innovative applications. Since each of these chapters is a self-contained paper, including its own set of conclusions, these conclusions will not be repeated in this chapter. Instead, the importance and impact and contribution of the work will be discussed, in terms of the different audiences it has reached and the ways in which the research has been used in or informed further work carried out by myself, co-authors, and others.

6.1 Impact of the research

6.1.1 Academic penetration

As well as the four published papers, the work has been presented at numerous academic conferences. Here are some examples:

- Mayhew, L & Harper, G 'Rethinking Welfare Measurement using a Whole Systems Approach', Perspectives of Improving Economic Welfare Measurement in a Changing Europe', 34th CEIES and Eurostat Seminar, 11th September 2007, Helsinki
- Harper, G & Mayhew, L 'Population Estimates Using Local Administrative Records', All Change – How can we get Better Population Statistics to Plan Local Services?, British Society for Population Studies Conference, 19th May 2008, Royal Statistical Society
- Harper, G and Mayhew, L 'Using Administrative Data to Estimate Population Size and Structure', LARIA (Local Area Research and Intelligence Association) Annual Conference – Doing More with Less, 4th April 2011, University of York
- Harper, G and Mayhew, L 'Using Administrative Data to Count Local Populations and Profile Households', British Society for Population Studies Annual Conference, 8th September 2011, University of York
- Harper, G 'Using Administrative Data for Local Purposes in Great Britain', Life After the Census: Using Administrative Data to Analyse Society, 9th May 2012 University of Ulster. The Northern Ireland Longitudinal Study Research Forum
- Mayhew, L 'Re-thinking Households – Using Administrative Data to Count and Classify Households', British Society for Population Studies Annual Conference 2012, 12th August 2012, University of Nottingham
- Harper, G 'Exploiting Administrative Data in the UK as an Alternative to the Census', Population Geography in a Post-Census World, 32nd International Geography Congress, 30th August 2012, University of Cologne
- Harper, G and Mayhew, L 'Re-thinking Households – Using Administrative Data to Count and Classify Households with some Geographical Applications' IGU Conference, June 2013, University of Leeds
- Harper, G and Mayhew, L 'Re-thinking Households – Using Administrative Data to Count and classify Households with some geographical Applications' Royal Statistical Society Annual Conference, 4th September 2013, Northumbria University

- Harper, G 'Exploiting Administrative Data for Population Estimation and Profiling – Experiences and Applications 'Royal Geographical Society - Institute of British Geographers Annual Conference, 3rd September 2015, University of Exeter

As can be seen, the research was relevant for population, demographic, statistical and geographic conferences, and was well received by the audiences. These presentations instigated discussion and further work, particularly with ONS as to how the census in England and Wales could transform into an administrative data census. This is discussed later in section 6.1.3 of this chapter.

As at 17th February 2017, paper 1 (chapter 2) in Applied Spatial Analysis and Policy was downloaded 845 times and cited 7 times. Paper 2 (chapter 3) in the same journal was downloaded 667 times and cited 7 times. Paper 3 (chapter 4) in the same journal was downloaded 554 times. Paper 4 (chapter 5) in PLOS ONE was viewed 3,769 times, and saved 30 times.

The work continues to inform my current and ongoing research interests. My experience of linking and analysing administrative data led to employment on the City, University of London School of Health Sciences 'Timing of birth and its outcome' project. My role has been to quality assure the linkage of national birth registration and hospital episode delivery data and assist in analysing the linked file.

Further funding has been given to extend the research in paper 4 to analyse if the introduction of Low Emission Zones reduces childhood asthma. My contribution is to link clinical, socio-economic and air quality data to enable this analysis.

The research on counting households using administrative data in paper 3 has led to an invitation to contribute to a meeting about households in administrative data, a collaboration between the RSS Social Statistics Section and the Administrative Data Research Centre for Scotland.

6.1.2 Commercial implementation

Myself and Professor Mayhew have been developing this system for the exploitation of administrative data and estimating and profiling populations in a commercial capacity since

2000 under the name 'Neighbourhood Knowledge Management' (*nkm*)[33]. To date over 60 projects enabled by the methodology have been completed for more than 20 local authorities, healthcare organisations and the third sector in a wide range of commissions. These include education, public health, housing, social care, crime, service design, economic evaluation, transport planning, equality impact assessments, investigations into chronic disease and Joint Strategic Needs Assessments (JSNA) for Primary Care Trusts.

The results of these commissions have been endorsed by the users themselves and in the media. As the first commissioner of the administrative data population estimation methodology in 2005, the London Borough of Brent have utilised the results in many ways. One of the most unusual was using estimates of the number and composition of households affected by a tornado weather event that hit the area in 2006 and caused damage to several residential streets. The intelligence provided by the administrative data population estimation helped the emergency services to account for potentially missing persons who might have been injured or killed. More generally it has been used by Brent as evidence of official statistics population under-counts and subsequent under-funding based on the local government Formula Grant. It has featured in House of Commons debates and Treasury and other Select Committee Reports. Hansard records that Sarah Teather MP referred to *nkm's* work as evidence of a serious under-count in official statistics and under-funding in a House of Commons debate (Teather, 2007). A letter from a councillor was published in The Guardian referring to the same evidence (Moher, 2012). Another early success was helping Brent to secure £60m in inner city funding for a project in South Kilburn.

There have been many other published examples of our work in the media in the national and local press. An article in The Guardian referred to the London Borough of Hackney's administrative data population estimation to describe the extent of the problem of population under-counting and under-funding in all boroughs in East London (Hill, 2012). London Borough of Waltham Forest referred to their administrative data population estimation as "The most up to date and accurate source of population data that we have. The ethnic breakdown is also unique and provides a far broader breakdown of ethnicities than those in the census" (London Borough of Waltham Forest, 2012). An article in the Jewish Chronicle referred to our work in estimating the size of the Cheredi population, a Jewish orthodox sect based mainly in north London.

[33] www.nkm.org.uk

The Department of Health valued the methodology by commissioning an online public health intelligence and local analysis toolkit based on the *nkm* approach. Their 2012 strategy paper referred to *nkm* work in the London Borough of Tower Hamlets as an example of best practice to shape commissioning and delivery of services (Department of Health, 2012).

One of the largest commissions was to apply the administrative data population estimation methodology in 2011 to the six Olympic boroughs (London Borough of Hackney, 2012) to identify their needs preparing for the 2012 Olympics and in the post-Olympic legacy period. The research was timed to coincide with the 2011 Census to enable comparisons to be made between these two population estimation methodologies, as well as Greater London Authority (GLA) estimates. The main findings were that estimates using the administrative data methodology were 3.5% higher than GLA estimates and 9.5% higher than ONS estimates for the same year, and that the latter fell into line with the administrative data estimates after further quality assurance and adjustments. The study was used to inform the six local authorities of their population size, to guide housing policy for the next five years and to monitor health and wellbeing under the Joint Strategic Needs Assessment of the Primary Care Trusts. Subsequent published official 2011 population estimates for the Olympic boroughs confirmed the accuracy of our then estimates.

Awareness of the *nkm* approach was also raised with commercial associations such as the British Market Research Associations and retailers such as Tesco. It was recognised that the data would be a perfect accompaniment to their own data which included surveys and retail data from different stores, and no other source provided all the above benefits. This was because, although geodemographic products such as MOSAIC or Acorn were available, these gave only high level household and postcode level typologies derived using imputed data from surveys, including the out-of-date Census itself. However, it was quickly established that the use of public data in such commercial applications would be in breach of local authority data sharing protocols and was never pursued. However, it has been acceptable to use commercially sourced surveys for linking to administrative data where the survey was local authority commissioned.

We have presented in council chambers and departments of local authorities including housing, social care and education directorates to both senior and elected officials in England, Scotland and Northern Ireland. As part of our work to engage with clients, we devised legally binding data sharing protocols, secure means of transmitting data, and encryption protocols where necessary. The work in this respect preceded the now more common modalities of

establishing data safe-havens which are gradually becoming more common in the public and academic sectors.

The methodology and its outputs have provided commercial users with more accurate, detailed and relevant data and evidence to make better policy decisions, design services better, and save money, and ultimately benefit local communities and service users.

All the commissions predominantly came about from recommendations and word of mouth between local authorities and from participation at relevant conferences and user groups. It became apparent that there was a significant demand for these innovative methods and outputs and *nkm* become established as the only providers of and experts in such services.

Currently, *nkm* continues to build on this foundation but with some constraints. For example, the implementation of the Health and Social Care Act in 2012 resulted in the abolishment of Primary Care Trusts and led to the end of local access to the GP Patient Register with identifiable variables, responsibility for which was transferred to the Health and Social Care Information Centre (now NHS Digital). This dataset is crucial to the administrative data population estimation methodology and without it cannot be carried out in its existing form. An attempt to carry out the administrative data population estimation methodology for the London Borough of Havering in 2013 using only local authority datasets without the GP Register counted less than half of the population. As a result, and until the renewal of access to the GP register is clarified, no further population estimations have been pursued since then by *nkm*, although we continue to advise the ONS as required, which does use the GP Register in their proposed Administrative Data Census (ADC).

Instead, applications of the *nkm* administrative data and linkage expertise has diversified and evolved and continues to be in demand. A major new direction has been helping local authorities manage their housing policies more effectively. For example, the linkage of local administrative datasets to identify the Private Rented Sector (PRS) and determine any association with levels of anti-social behaviour has been especially effective. Depending on the strength of the evidence found, *nkm* provides local authorities with a robust case for introducing private landlord licensing schemes locally. Examples of reports and the subsequent introduction of licensing schemes can be found on many local authority web sites. The robustness of our approach was recognised in a judicial review brought against Enfield Council regarding the introduction of landlord licensing in December 2014. The presiding judge Justice McKenna found that "… the judgement does not find fault with the licensing scheme or

challenge the evidence-based report (i.e. *nkm*) upon which the decision to implement this scheme was taken – which Justice Ouseley has previously described as 'detailed and careful'".

Another change of direction occurred in 2016 when *nkm* were invited to analyse and link together the contents of a database belonging to children's charity which distributed grants to children in crisis in all parts of the U.K. The *nkm* analysis showed that previous allocations of grants could be significantly improved upon using better local measures of need. Not only was the work implemented, it received national coverage in local media and was commended by the Joseph Rowntree Foundation.

Interest has been shown again recently by previous local authority clients for an administrative data population estimation refresh. *nkm* are exploring new options to access the required data. Many other users are dissatisfied with the restricted data access legal gateway since the Health and Social Care Act 2012, and this could change again in the future as a result. It is a continually evolving landscape.

6.1.3 Informing national statistics and trailblazing administrative data strategy

As well as the methodology being recognised and implemented academically and commercially, the body of work coincided with and in fact preceded a general move towards exploiting administrative data for enhancing or replacing the census survey of population in the UK to provide national population statistics, and for research.

National statistics

National statistics bodies in England and Wales (ONS), Scotland (National Records of Scotland - NROS) and Northern Ireland (Northern Ireland Statistics and Research Agency– NISRA) began a debate on whether the decennial census survey of population should be discontinued in its current format, and if so what should replace it.

This was initiated by the census of population results being declared as "unfit for purpose" by a House of Commons Treasury Select Committee in 2008 (House of Commons Treasury Committee, 2008) and the Minister for the Cabinet Office in the same year announcing that the 2011 Census "would be the last".

In reaction to this ONS set up the Beyond 2011[34] programme in 2010 to explore alternatives to running a traditional ten-year census in 2021 in England and Wales.

By this stage *nkm* already had five years' experience of designing and carrying out administrative data population estimation projects for local authority areas, with an established methodology covering all aspects of implementation and delivery including data linkage and rules to define current residents from the linked datasets. In 2012, the House of Commons Science and Technology Committee report 'The Census and Social Science' (House of Commons Science and Technology Committee, 2012) referred to the *nkm* administrative data methodology as "confirmation that there is a credible alternative to the census for the purposes of local government".

Based on *nkm's* extensive experience, in 2012 we were invited to present our methods to the Northern Ireland Longitudinal Research Forum 'Life after the Census' conference. In 2013 NROS alongside the General Register office for Scotland (GRO) and NHS Greater Glasgow and Clyde and Glasgow City Council consulted with *nkm* to explore administrative data alternatives to the census to devise a more accurate formula for funding health boards in Scotland. This was initiated by concerns of serious under-counting of the population in Glasgow by the 2011 Census.

Then the ONS Beyond 2011 team consulted with *nkm* in 2014 to get a better understanding of the methodology, and to share their progress to date and to consider potential collaboration and transferring knowledge. By that point, the six alternatives for the 2021 Census had been narrowed down to two options: an online census or an administrative data census plus annual surveys. The National Statistician recommended the online census option while making increased use of administrative data and surveys to enhance the statistics from the 2021 Census.

By 2015, the Minister for the Cabinet Office announced that "our ambition is that censuses after 2021 will be conducted using other sources of data and providing more timely statistical information". ONS set up the Census Transformation Programme (CTP)[35] to develop Administrative Data Census (ADC) estimates for comparison to the 2021 online Census results and improve population statistics through increased use of administrative data and surveys.

[34] https://www.ons.gov.uk/census/censustransformationprogramme/beyond2011censustransformationprogramme
[35] https://www.ons.gov.uk/census/censustransformationprogramme

Since then the ONS CTP has continued to explore and assess data linkage and population estimation methods using administrative data. Since 2015, annual administrative data research outputs have been given. Their aim is to give their final recommendation about the future provision of population statistics in 2023.

The 2016 CTP conference[36] stated the need for radical change in national statistics for evidence-based decision making and because users want more targeted granular data and faster, and value for money. To successfully implement an ADC will require easy and rapid access to data, high quality efficient linkage using identifiable data, established methods and outputs of sufficient quality. *nkm* had already recognised these needs and created methods for an administrative data population estimation, and provided them in their services to local authority clients.

Appendix 6.A sets out a timeline of key development points for *nkm* and ONS administrative data population estimation methodologies. It illustrates that the *nkm* methodology set out in this thesis was a trailblazer for using and linking routinely collected administrative data, for the purpose of estimating populations. The ONS or indeed other official statistics bodies in the UK, were not providing similar services at that time.

Comparison to official population statistics is a key thread to the whole narrative of the research in this thesis. To begin with, the administrative data population estimation methodology was created in direct response to concerns about the inadequacies of the Census 2001 results and outputs for some local authorities. Indeed, the results from *nkm*'s administrative data population estimations for the six Olympic boroughs were used by ONS to quality assure the Census 2011 results for these areas. Then, as the methodology and applications developed and were commissioned by more users, the results and outputs have been compared to those of the census to assess suitability as an alternative source of population statistics, and the pros and cons of each.

In reflection, further assessment can be made on the similarities and differences between the two options, and the impact this research has made. ONS has used administrative data for various purposes since 2002, as can be seen Appendix 6.A. This includes to improve migration and mid-year population estimate statistics. It is only since 2015 that they have been working towards a full 'Administrative Data Census' methodology as part of the CTP.

[36]https://www.ons.gov.uk/census/censustransformationprogramme/progressanddevelopment/2016researchconference

From the ONS CTP annual research outputs since 2015, specific methodological similarities to the *nkm* method are evident. They both share the general concept of linking together multiple sources to improve the coverage of outputs and applying rules and assumptions to determine who is a current resident. This is called the 'minimum confirmed population' in *nkm*, and a 'Statistical Population Dataset' (SPD) by ONS.

In addition, similarly to the existing *nkm* method, ONS are now applying associative matching techniques (finding new links between data by drawing upon the strength of confirmed matches for other individuals at the same address), 'superseding' assumption rules (it is more likely a match is correct if the individual is present on multiple datasets rather than one, and some datasets are more reliable than others), the assignment of records on the SPD to the most likely address using activity data (the latest address that an individual had interaction with a service is most likely to be the correct address), and the use of school census data to help enumerate children. They have stated their intention to use council tax, electoral register, birth and death registration data in their next outputs, and using the address and Unique Property Reference Number (UPRN) as the population building block. These are datasets and methods that *nkm* have always used in their methodology. ONS reference Harper and Mayhew's paper 1 in their outputs[37].

In general, most of the issues in 'The ONS CTP Annual Assessment of ONS's progress towards an ADC post-2021' report[38] were considered by *nkm* many years previously, as can be seen in the papers in this thesis and other outputs such as the UPTAP report. ONS now also state that there is a trade-off present in employing an administrative data population estimation methodology instead of a survey of population, something the research in this thesis has always alluded to.

nkm set out a methodology for counting and classifying households using administrative data in paper 3 of this thesis. ONS released outputs in February 2017[39] on their initial test of using administrative data for this, and again reference Harper and Mayhew's paper 3 in their report. This uses the same concept as *nkm*'s household estimation methodology. Households are

[37]https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/sizeofthepopulation/researchoutputsestimatingthesizeofthepopulationinenglandandwales2016release

[38]https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusannualassessments

[39]https://www.ons.gov.uk/census/censustransformationprogramme/administrativedatacensusproject/administrativedatacensusresearchoutputs/householdsandfamilies/occupiedaddresshouseholdestimatesfromadministrativedata2011and2015

represented as UPRNs with at least one confirmed resident from the administrative data population estimation.

However, comparisons must also consider important differences. Methodologically, ONS have access to additional datasets, such as the Department for Work and Pensions CIS (Customer Information System) and HESA (Higher Educational Statistics Agency) data, and use different rules to create their SPD. Unlike the *nkm* method, they are intending to use a population coverage survey to adjust for over- and under-coverage and develop confidence intervals around the SPD. To date ONS have been limited to anonymised linkage of datasets, but *nkm* had access to identifiers to enable more accurate linkage. The May 2016 Cabinet Office consultation paper 'Better Use of Data in Government'[40] supports changes in legislation to improve ONS' access to identifiable data for official statistics purposes.

ONS have a range of stakeholders and user needs to meet and have much greater resources to employ in researching methodologies and quality. They are working at a national scale and their outputs need to be comparable between all local authorities in England and Wales (and indeed all UK countries), be of national statistics standard, and be longitudinally robust. The CTP programme is aiming to assess if administrative data can fully achieve the same outputs of the census survey of population in terms of quality and the full range of population and household characteristics. *nkm* captured a range of attributes if they were available in the input datasets that were linked.

In contrast, the research and methods in this thesis were instigated in direct response to local authority's concerns and needs, and overall are most suitable for this group of users. Although the research is at a local scale, it is proof of success of the concept and methods, and as such is pioneering work and important evidence for the ONS to draw upon.

It is easier to make radical changes at local level rather than national level, and it will be some time yet before there is a national Administrative Data Census.

Administrative data for research

In parallel, since the research was published there has been a general increase in the understanding of the value and use of linked routinely collected administrative data. The ESRC

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

funded Administrative Data Research Network (ADRN)[41] was established in the UK in 2013 after the Administrative Data Taskforce report[42] recommended to set up an independent organisation that would help social and economic researchers gain access to linked, de-identified administrative data in a safe and lawful way. This is so that the information the government collects can be used to benefit research. The Digital Economy Bill[43], announced in May 2016, seeks to further increase access to de-identified administrative data for research.

The ADRN states that amongst the benefits of using administrative data for research is that it is more economically efficient, and has greater impact on policy because the research is achieved more efficiently and provides a more up-to-date evidence base. Their aim is to use administrative data for "better knowledge, better society".

Again, the research in this thesis had already recognised these benefits of using administrative data for policy and setting precedent of legally and safely accessing these datasets and methods for linking and analysing them. The recent ADRN conference in 2016[44] offered little in new perspectives or innovations, confirming that this research was ahead of its time.

The 2016 International Population Data Linkage conference[45] proposed the new academic field of 'Population Data Science', of which it can be argued this research has contributed to and is a part of. Data science can be considered the maximisation of the value of data for decision-making. Similarly, it is proposed there is a "statistical scientific paradigm shift"[46] from single-source to multi-source official statistics. This is reflected in 2017's New Techniques and Technologies for Statistics (NTTS) conference[47], where data integration and new methods for statistics are being discussed. This shows that these issues are relevant to Europe and internationally. This research has made an important contribution to that paradigm shift.

[41] https://www.adrn.ac.uk/
[42] https://www.adrn.ac.uk/media/1376/improving-access-for-research-and-policy.pdf
[43] https://www.publications.parliament.uk/pa/bills/cbill/2016-2017/0045/cbill_2016-20170045_en_1.htm
[44] http://www.adrn.ac.uk/media/1249/adrnconf16programme.pdf
[45] https://www.ipdlnconference2016.org/
[46] *Li-Chun Zhang* https://www.statslife.org.uk/events/eventdetail/842/-/the-potential-of-non-survey-data-in-ons-social-statistics
[47] http://ec.europa.eu/eurostat/cros/content/ntts-2017_en

6.1.4 Research Excellence Framework (REF) recognition and award

The body of research around the methodology was part of a group of submitted case studies deemed to be internationally excellent by City, University of London's internal panel preparing for the Research Excellence Framework (REF) in 2012[48,49].

The REF submission summarised the high level of impact the work has had. The award states that "there was impressive evidence of outstanding impact from this unit across a range of areas and the impact template was consistent with a balance of outstanding and very considerable impact. The case studies concerning demographic change and the cost of social care, estimating populations and mortality and life expectancy estimates were assessed as outstanding." This further highlights the tangible impact the research has had.

The research was also short-listed in 2015 for City, University of London's Vice-Chancellor's Award in recognition of outstanding research impact. It achieved the LARIA (Local Area Research and Intelligence Association) Excellence in Research Commendation 2011.

6.2 Critical reflection

The methodology in the research relies heavily on the linkage between the administrative datasets. However, this critical aspect was not discussed in a lot of detail in any of the papers in this thesis, yet data linkage is now a stand-alone academic research area in its' own right.

I designed bespoke linkage algorithms to match addresses and people across the datasets. There is no consistent unique person identifier on government data in England and Wales. Instead, person identifiers are compared to establish matches. *nkm* had the advantage of access to identifiable data, something that is now rarely possible. This also allows for quality of the linkage to be checked post hoc and clerically. ONS have only worked with anonymised linkage, and are seeking for non-anonymised linkage because it is more accurate. Change in legislation is required to support access to identifiable government data for official statistics. ONS state that high quality matching will be essential in the production of accurate population estimates. Future work could be to disseminate my valuable experience of this.

[48] http://impact.ref.ac.uk/CaseStudies/CaseStudy.aspx?Id=44377
[49] http://www.city.ac.uk/news/spotlight-on-research/better-population-data-striving-for-greater-accuracy

The quality and accuracy of the population estimation results were assessed as much as possible within the scope of the research and the resources available, mainly against sensible benchmarks. It is difficult to do this more thoroughly without a gold standard to compare it against. ONS are using traditional census results to compare their administrative data census against. However, this was not an option for this research because it finds the census to not be gold standard, and the census is not available at household level breakdowns for comparisons at this scale. The ONS Census Transformation Programme are looking to find a way to assign confidence intervals that does not rely on comparisons to the traditional census, as this will not always be available. Future work could consider how to improve on assessing quality and confidence in results. This could be assessed in terms of bias, not only in linkage, but in the datasets used and the rules applied.

The methodology in its current form could not be replicated exactly today. This is due to recent changes in data content and access. Further work could find ways to future-proof the methodology in these regards. ONS are also aware of how these issues could affect the feasibility of an Administrative Data Census.

The methodology has not been attempted at national level. Future work could assess the suitability of this. Conceptually this could be the combination of administrative data population estimations for each local authority in England and Wales using local data, or the application of the methodology on national data. There are pros and cons to each option. Creating the outputs at local level to feed into a national hub would have access to a wider range of datasets and have a faster turnaround. The top down approach using national data and produced centrally would provide greater consistency and economies of scale.

Overall the methodology provided fit-for-purpose results for some of the most difficult to count areas in England with acceptable trade-offs in getting those results quickly and at granular level.

6.3 Overall Summary

The four papers contained in this thesis illustrate that routinely collected administrative data in the UK can be accessed and linked, for the purpose of population estimation. They also demonstrate the policy relevant applications the outputs can be used for.

This research preceded the current paradigm shift in the UK for research and national statistics to move towards the use of linked administrative data, and as such has been part of this inevitable evolution and has helped pave the way for this.

APPENDIX 6.A

Timeline of use of administrative data by the author/co-authors/*nkm* and the Office of National Statistics (ONS)

Year	Author/co-authors/*nkm*	ONS
2000	• First use of linked local authority and health trust administrative data for LB Brent Health Action Zone • Information Commissioner Data Protection Act compliance confirmation (March)	
2001		• Census 2001 carried out (March)
2003	• *nkm* (neighbourhood knowledge management) brand name formed • *nkm* pilot project with Tower Hamlets Partnership	• Census 2001 results deemed to under-count in some cities in England and parts of London • Administrative data used to improve 2001 Census address lists and revise MYE for 2001, 2002 and 2003 in Manchester and Westminster. Fed into improvements of 2011 Census • Proposals for Integrated Population Statistics System combining census, survey and administrative data at individual level
2004		• Improving Migration and Population Statistics Programme (IMPS) set up. One working area was proposed collaboration with local authorities to investigate potential for making greater use of administrative data to improve local estimates
2005	• First *nkm* administrative data population estimation (ADPE) for LB Brent Accurate Census project	
2006	• *nkm* ADPE for Doncaster • *nkm* ADPE for LB Enfield • *nkm* ADPE for LB Southwark • The 'Impact of Asthma, Ethnicity, and Social Adversity on educational Achievement' project with Centre for Health Sciences, Queen Mary University • LB Brent use *nkm's* LB Brent Accurate Census results to estimate the number and composition of households and vulnerable people within the exclusion zone of the December 2006 tornado to calculate need for rest centres and support (December)	• Use of administrative data to improve migration statistics

2007	• First use of *nkm* household classification for LBs Brent and Newham • *nkm* ADPE for North East Lincolnshire • *nkm* ADPE for LB Hackney • *nkm* ADPE for LB Newham • Sarah Teather MP refers to *nkm's* LB Brent Accurate Census results as evidence of under-count in official statistics and under-funding in a House of Commons debate (November)	
2008	• Professor Mayhew and LB Brent provide written evidence to the House of Commons' Treasury Committee Report 'Counting the Population' on how *nkm* ADPE is an alternative to the census and proves census under-count and under-funding in LB Brent (January) • *nkm* ADPE for Birmingham • *nkm* ADPE for LB Barking & Dagenham • *nkm* Cheredi project based on religious group carried out for LB Hackney • Author funded as ESRC UPTAP Research fellow at Cass Business School to write up methodology (June 2008 to December 2009)	• Census results described as "unfit for purpose" by the House of Commons Treasury Select Committee (May) • Rt. Hon Francis Maude MP, Minister for the Cabinet Office, announces that the 2011 Census would be the last
2009	• LINK (Local Information and Knowledge Management) public health intelligence toolkit created for Department of Health based on *nkm* methodology • *nkm* ADPE for Mid-Essex PCT • *nkm* ADPE for LB Greenwich • *nkm* ADPE for LB Tower Hamlets • *nkm* ADPE for LB Waltham Forest	
2010	• First use of *nkm* ethnicity predictor algorithm for LB Waltham Forest • *nkm* ADPE for Luton with focus on immigration and ethnicity • UPTAP report 'Using administrative Data to Estimate the Population and Applications' published. Graded 'Good'	• ONS begin the BY2011 (Beyond 2011) programme. This gained access to record level administrative data thus enabling previous concepts to become a reality (May)

2011	• *nkm* obtains Information Governance Toolkit certification • *nkm* ADPE for six Olympic boroughs (Greenwich, Tower Hamlets, Newham, Barking & Dagenham, Hackney, Waltham Forest) to coincide with Census, incorporating *nkm* household classification and ethnicity predictor (March) • *nkm* 2011 ADPE for Newham used in Census 2011 quality assurance • Household paper published as Cass Business School Actuarial Research Paper number 128	• Census 2011 carried out (March) • BY2011 create administrative data based population estimates for local authorities and compared with Census 2011 results to assess the quality of administrative data
2012	• LB Waltham Forest publish a guide to population sources referring to *nkm* LB Waltham Forest ADPE. They state that "this dataset is by far the most up to date and accurate source of population data that we have. The ethnic breakdown is also unique and provides a far broader breakdown of ethnicities than those in the census" (February) • Department of Health strategy paper uses *nkm* Tower Hamlets integrated health, social care and population data work as example of best practice to shape commissioning and delivery of services (May) • Paper 1 'Using administrative Data to Count Local Populations' published in Applied Spatial Analysis and Policy (June) • Letter from London Borough of Brent councillor published in The Guardian quoting *nkm* LB Brent ADPE and discussing ONS under-count of population in Brent (August) • Paper 2 'Applications of Population Counts based on Administrative Data at Local Level' published in Applied Spatial Analysis and Policy (September) • House of Commons Science and Technology Committee report 'The Census and Social Science' published. *nkm*'s evidence quoted as "confirmation that there is a credible alternative to the census for the purposes of local government" (September) • Article in The Guardian quoting *nkm* LB Hackney ADPE and discussing the problem of ONS under-counting and its effect on the local government Formula Grant system allocations for East London Boroughs (September) • Paper 4 'Impact of Asthma on Educational Attainment in a Socioeconomically Deprived Population: A Study Linking Health, Education and Social Care Datasets' published in PLOS ONE (November) • Health and Social Care Act 2012 and NHS reform ends PCTs and local health data now held centrally by HSCIC/NHS Digital	• Administrative Data Taskforce (ADT) model of linking and de-identifying data using a 'trusted third party' before use for research purposes

2013	• Last full *nkm* ADPE carried out for LB Haringey • Consultations with General Register Office for Scotland/National Records of Scotland, NHS Greater Glasgow and Clyde and Glasgow City Council on alternatives to the Census and under-counting of population in Glasgow • *nkm* equality assessment for LB Havering found that without access to the GP Register, the *nkm* ADPE methodology counts less than half of the population using only local authority datasets	• ONS BY2011 carry out public consultation on 2 front-running options for 2021 – an online census or ADC plus annual surveys • ONS BY2011 indicate that an initial assessment of 6 options under consideration found that option 5 (administrative data linkages, plus an annual circa 1% coverage survey with a one-off circa 10% coverage survey in 2021 to validate the method) was the most cost-effective and produced the best quality data. This reflects the author and co-author's research (May) • BY2011 consider 'associative matching' of administrative data (already used by *nkm*) • BY2011 methodological reports. First basic SPD by linking PR, CIS and HESA and applying rules to create LA basic age and sex counts (July) • BY2011 Independent Review of Methodology (October)
2014	• Consultation with ONS BY2011 team (July)	• National Statistician recommends that the census in 2021 should be predominantly online, making increased use of administrative data and surveys to enhance the statistics from the 2021 Census (March) • Consultation with *nkm* (July)
2015	• Francis Maude, Minster for the Cabinet Office announces "Our ambition is that censuses after 2021 will be conducted using other sources of data and providing more timely statistical information" (July) • Paper 3 'Using Administrative Data to Count and Classify Households with Local Applications' published in Applied Spatial Analysis and Policy • *nkm* diversify into projects identifying PRS and HMOs from administrative data	• ONS establish Census Transformation Programme (CTP) January 2015 to deliver predominantly online census in 2021, and develop alternative administrative data census estimates for comparison, and improve population statistics through increased use of administrative data and surveys • First set of CTP admin data research outputs (October)

2016	• *nkm* diversify into providing evidence on how to improve grant allocation for children's charity Buttle UK	• ONS CTP Programme Conference to engage with users • Independent Review of UK Economic Statistics recommend that ONS "make the most of existing and new data sources" to improve economic statistics • Better Use of Data in Government consultation paper supports changes in legislation to improve ONS access to identifiable administrative data for official statistics purposes (February) • CTP find that linking anonymised data does not deliver level of quality required to produce PEs of required accuracy • CTP consider use of UPRN (already used by *nkm*) • CTP plan to access and assess electoral register and council tax data (already used by *nkm*) • CTP SPD considering special population data e.g. prisons and military (already used by *nkm*) • Revised ADC size of population outputs referencing *nkm* methodology (November) • ADC outputs on income (December) • SPD V2 methodology changes more in line with *nkm* and includes school census data (already used by *nkm*)
2017	• *nkm* re-investigating access to GP Register to enable ADPE	• ONS take over from DCLG for household projections • CTP release Research Outputs on households referencing *nkm* methodology (February) • RSS households in administrative data meeting (September)
2018		• ADC Population Coverage Survey Test
2021		• ONS to have replicated as many census outputs as possible using administrative data and surveys
2023		• ONS to make recommendation about future provision of population statistics

References

Administrative Data Taskforce. 2012. The UK administrative data research network: improving access for research and policy. Report from the Administrative Data Taskforce. Available: http://www.esrc.ac.uk/_images/ADT-Improving-Access-for-Research-and-Policy_tcm8-24462.pdf

Alder, J., Mayhew, L., Moody, S., Morris, R. & Shah, R. 2005. *The chronic disease burden-An analysis of health risks and health care usage.* Cass Business School, City University London. London.

Altman, D. G. 1999. *Statistics for Medical Research,* London, Chapman & Hall.

Anderson, H. R., Bailey, P. A., Cooper, J. S., Palmer, J. C. & West, S. 1983. Morbidity and school absence caused by asthma and wheezing illness. *Archives of Disease in Childhood,* 58**,** 777-784.

Asher, M. I., Montefort, S., Björkstén, B., Lai, C. K. W., Strachan, D. P., Weiland, S. K. & Williams, H. 2006. Worldwide time trends in the prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and eczema in childhood: ISAAC Phases One and Three repeat multicountry cross-sectional surveys. *The Lancet,* 368**,** 733-743.

Austin, J. K., Huberty, T. J., Huster, G. A. & Dunn, D. W. 1998. Academic achievement in children with epilepsy or asthma. *Developmental Medicine & Child Neurology,* 40**,** 248-255.

Baffour, B., King, T. & Valente, P. 2013. The Modern Census: Evolution, Examples and Evaluation. *International Statistical Review,* 81**,** 407-425.

Baker, D., Taylor, H. & Henderson, J. 1998. Inequality in infant morbidity: causes and consequences in England in the 1990s. ALSPAC Study Team. Avon Longitudinal Study of Pregnancy and Childhood. *Journal of Epidemiology & Community Health,* 52**,** 451-458.

Barr, N. 2004. *The Economics of the Welfare State*, Oxford University Press.

Bevan, G. 2009. The search for a proportionate care law by formula funding in the English NHS. *Financial Accountability & Management,* 25**,** 391-410.

Bowley, G. 2003. The last census? *Prospect Magazine*. November 20. Available: https://www.prospectmagazine.co.uk/magazine/thelastcensus

Bowling, A. N. N. 1991. Social Support and Social Networks: Their Relationship to the Successful and Unsuccessful Survival of Elderly People in the Community. An Analysis of Concepts and a Review of the Evidence. *Family Practice,* 8**,** 68-83.

Brackstone, G. 1987. Statistical Purposes. *Survey Methodology***,** 29.

British Thoracic Society, S. I. G. N. 2007. *BTS/SIGN British Guideline on the Management of Asthma* [Online]. Available: http://www.brit-thoracic.org.uk/c2/uploads/asthma_fullguideline2007.pdf [Accessed Sep 20 2009].

Bronte-Tinkew, J. & Dejong, G. F. 2005. Do household structure and household economic resources predict childhood immunization? Evidence from Jamaica and Trinidad and Tobago. *Population Research and Policy Review,* 24**,** 27-57.

BSPS meeting. May 18. 2015. *The 2012-based household projections for England: methodological issues.* LSE. London. Available: http://www.lse.ac.uk/socialPolicy/BSPS/dayMeetings/Home.aspx

Burghardt, J. A. & Geraci, V. J. 1980. State and local annual population estimation methods employed by the Bureau of the Census. *Review of Public Data Use,* 8**,** 339-354.

Bush, A. & Saglani, S. 2010. Management of severe asthma in children. *The Lancet,* 376**,** 814-825.

Central London Forward. 2010. *Census coverage survey and imputation. Deliberate event. Meeting minutes.* Available: http://www.centrallondonforward.gov.uk/news/clf-census-event-7-october-2010/

Clark, N. M., Feldman, C. H., Evans, D., Wasilewski, Y. & Levison, M. J. 1984. Changes in Children's School Performance as a Result of Education for Family Management of Asthma. *Journal of School Health,* 54**,** 143-145.

Coleman, D. A. & Schofield, R. 1986. *The state of population theory: forward from Malthus*, Blackwell Oxford.

Communities and Local Government. *Indices of Depivation 2007* [Online]. Available: http://webarchive.nationalarchives.gov.uk/20100806161347/http://www.communities.gov.uk/communities/neighbourhoodrenewal/deprivation/deprivation07/ [Accessed Jun 7 2010].

Cook, L. July. 2003. *A demographics statistics service for the 21st Century. Introductory covering letter.* Office for National Statistics. Available: http://www.statistics.gov.uk/about/methodology_by_theme/Dem_Stat_Ser_21ST_Cen.asp

de Bruin, A., Kardaun, J., Gast, F., de Bruin, E., van Sijl, M. & Verweij, G. 2004. Record linkage of hospital discharge register with population register: experiences at Statistics Netherlands. *Statistical Journal of the United Nations Economic Commission for Europe,* 21**,** 23-32.

Department for Communities and Local Government. 2008. *Options for the future of the household projection model.* Communities and Local Government. London. Available: http://www.communities.gov.uk/documents/housing/pdf/housingprojectionmodel.pdf

Department for Communities and Local Government. March. 2010a. *Consultation on proposed changes to the national statistics on household projections.* Communities and Local Government. London. Available: http://www.communities.gov.uk/documents/housing/pdf/1487114.pdf

Department for Communities and Local Government. November. 2010b. *Updating the Department for Communities and Local Government's household projections to a 2008 base - methodology.* Communities and Local Government. London. Available: http://www.communities.gov.uk/documents/statistics/pdf/1780350

Department for Communities and Local Government/Royal Statistical Society meeting. 2013. *DCLG Household projections.* Royal Statistical Society. London.

Department of Health. 2012. *The power of information: Putting all of us in control of the health and care information we need.* Available: https://data.gov.uk/sites/default/files/DH%20Open%20Data%20Strategy_10.pdf

Diette, G. B., Markson, L., Skinner, E. A., Nguyen, T. T. H., Algatt-Bergstrom, P. & Wu, A. W. 2000. Nocturnal Asthma in Children Affects School Attendance, School Performance, and Parents' Work Attendance. *Archives of Pediatrics & Adolescent Medicine,* 154**,** 923.

Dorling, D. 2007. How many of us are there and where are we? A Simple Independent Validation of the 2001 Census and its Revisions. *Environment and Planning A,* 39**,** 1024-1044.

Duran-Tauleria, E. & Rona, R. J. 1999. Geographical and socioeconomic variation in the prevalence of asthma symptoms in English and Scottish children. *Thorax,* 54**,** 476-481.

Eagan, T. M. L., Gulsvik, A., Eide, G. E. & Bakke, P. S. 2004. The effect of educational level on the incidence of asthma and respiratory symptoms. *Respiratory Medicine,* 98**,** 730-736.

Egton Medical Information Systems Limited. 2007. *EMIS* [Online]. Available: http://www.emis-online.com [Accessed May 16 2008].

Ellison-Loschmann, L., Sunyer, J., Plana, E., Pearce, N., Zock, J. P., Jarvis, D., Janson, C., Anto, J. M. & Kogevinas, M. 2007. Socioeconomic status, asthma and chronic bronchitis in a large community-based study. *European Respiratory Journal,* 29**,** 897-905.

Ericksen, E. P. & Kadane, J. B. 1986. Using administrative lists to estimate census omissions. *Journal of official statistics,* 2**,** 397.

Eurostat Statistical Books 2009. *Reconciliation between work, private and family life in the European Union, section 4.1.4,* Luxembourg, Office for Official Publications of the European Communities.

Fender, V. 2013. Household satellite accounts: valuing informal childcare in the UK.

Fowler, M. G., Davenport, M. G. & Garg, R. 1992. School functioning of US children with asthma. *Pediatrics,* 90**,** 939-944.

Free, S., Howden-Chapman, P., Pierse, N. & Viggers, H. 2010. Does more effective home heating reduces school absences for children with asthma? *Journal of Epidemiology & Community Health,* 64**,** 379-386.

Freedman, M., Lane, J. & Roemer, M. I. 2008. New Approaches to Creating Data for Economic Geographers. *Journal of Official Statistics,* 24**,** 133-56.

Frost, M. & Harper, G. 2007. The challenge of profiling communities. *The National Evaluation of Sure Start: Does area-based early intervention work?* : Policy Press.

Gehlke, C. E. & Biehl, K. 1934. Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material. *Journal of the American Statistical Association,* 29**,** 169.

Gill, L. 2001. Methods for automatic record matching and linkage and their use in national statistics. National Statistics Methodological Series No. 25. *London: National Statistics*.

Greenland, S. & Robins, J. 1994. Invited Commentary: Ecologic Studies—Biases, Misconceptions, and Counterexamples. *American Journal of Epidemiology,* 139**,** 747-760.

Gutstadt, L. B. 1989. Determinants of School Performance in Children With Chronic Asthma. *Archives of Pediatrics & Adolescent Medicine,* 143**,** 471.

Harper, G. 2002. Using surfaces to inform local policy–modelling deprivation in Brent. *Unpublished MSc dissertation.*

Harper, G. & Mayhew, L. 2012a. Using Administrative Data to Count Local Populations. *Applied Spatial Analysis and Policy,* 5(2)**,** 97-122.

Harper, G. & Mayhew, L. 2012b. Applications of Population Counts Based on Administrative Data at Local Level. *Applied Spatial Analysis and Policy,* 5(3)**,** 183-209.

Harper, G. & Mayhew, L. 2016. Using Administrative Data to Count and Classify Households with Local Applications. *Applied Spatial Analysis and Policy,* 9(4)**,** 433-462.

Hayward, J., Murray, D., Iny, I., Jarrett, J., Lonergan, K., Pillas, D. & Seager, S. 2010. *London TB Service Review and Health Needs Assessment.* . Available: http://www.csl.nhs.uk/Publications/Documents/Tuberculosis/PHAST_London_TB_report_I.pdf

Hill, D. 2012. How to count East Enders. *The Guardian*, September 2012.

HM Government. 2010. *The coalition: Our programme for Government.* Available: http://www.cabinetoffice.gov.uk/media/409088/pfg_coalition.pdf

HM Government. 2011. *Growing the social investment market: A vision and strategy.* Cabinet Office. London. Available: http://www.parliament.uk/deposits/depositedpapers/2011/DEP2011-0271.pdf

Hollis, J. & Chamberlain, J. 2009. *GLA 2008 round demographic projections. DMAG Briefing 2009–02.* Data Management and Analysis Group, Greater London Authority.

Holloway, S., Short, S. & Tamplin, S. 2002. Household satellite account (experimental) methodology. *London: ONS*.

Holmans, A. E. 2012. *Interim revised estimates of future demand and need in England in 2006 – 2026 with 2008-based demography.* Cambridge Centre for Housing and Planning Research. Available: http://www.cchpr.landecon.cam.ac.uk/Downloads/Future%20demand&need%20WEB%20COPY.pdf

Hope, C. 2010. National census to be axed after 200 years. *The Daily Telegraph*.Available: http://www.telegraph.co.uk/news/newstopics/politics/7882774/National-census-to-be-axed-after-200-years.html

House of Commons Public Administration Select Committee. 2014. *Too soon to scrap the census. Fifteenth report of session 2013–14.* House of Commons London.

House of Commons Science and Technology Committee. 2012. *The Census and social science: Third Report of Session 2012-13.* The Stationery Office Limited. London. Available: https://publications.parliament.uk/pa/cm201213/cmselect/cmsctech/322/322.pdf

House of Commons Treasury Committee. 2008. *Counting the population: Eleventh Report of Session 2007–08.* London. Available: https://publications.parliament.uk/pa/cm200708/cmselect/cmtreasy/183/183.pdf

House, T. & Keeling, M. J. 2008. Household structure and infectious disease transmission. *Epidemiology and Infection,* 137**,** 654.

Jenkins, S. P., Lynn, P., Jäckle, A. & Sala, E. 2008. The Feasibility of linking household survey and administrative record data: New evidence for Britain. *International Journal of Social Research Methodology,* 11**,** 29-43.

Jones, L., Bellis, M. A., Dedman, D., Sumnall, H. & Tocque, K. 2008. *Alcohol-attributable fractions for England: alcohol-attributable mortality and hospital admissions*. Available: https://www.alcohollearningcentre.org.uk/_assets/AlcoholAttributableFractions.pdf

Keohane, N. 2008. *Local counts – the future of the census.* National Local Government Network. London. Available: http://www.nlgn.org.uk/public/wp-content/uploads/local-counts.pdf

King, K., Martynenko, M., Bergman, M. H., Liu, Y. H., Winickoff, J. P. & Weitzman, M. 2009. Family Composition and Children's Exposure to Adult Smokers in Their Homes. *PEDIATRICS,* 123**,** e559-e564.

Kitch, B. T., Chew, G., Burge, H. A., Muilenberg, M. L., Weiss, S. T., Platts-Mills, T. A. E., O'Connor, G. & Gold, D. R. 2000. Socioeconomic Predictors of High Allergen Levels in Homes in the Greater Boston Area. *Environmental Health Perspectives,* 108**,** 301.

Kleinert, S. 2007. Adolescent health: an opportunity not to be missed. *The Lancet,* 369**,** 1057-1058.

Larsson, K., Thorslund, M. & Kåreholt, I. 2006. Are public care and services for older people targeted according to need? Applying the Behavioural Model on longitudinal data of a Swedish urban older population. *European Journal of Ageing,* 3**,** 22-33.

Lawrence, D., Payel, I. & Sievwright, M. 2007. Brent primary care trust shortfall in revenue allocations for 2006/7 and 2007/8. Report. Brent PCT.

Leeds City Council. 2011. *Leeds strategic housing market assessment update.* Leeds City Council. Available: http://www.leeds.gov.uk/files/Internet2007/2011/27/final%20leeds%20shma%2027-05-11(1).pdf

Lipschutz, S. 1998. Schaum's outline of set theory and related topics.

Local Government Association. 2007. *Estimating the scale and impacts of migration at the local level. Report.* LGA. Available: https://www.local.gov.uk/research-population-and-migration-estimating-scale-and-impacts-migration

London Borough of Hackney. 2012. Comparative analysis of the resident population of the six Olympic host boroughs: sources and uses of locally owned administrative data, a report by nkm. Available: https://www.hackney.gov.uk/media/2671/Comparative-analysis-of-the-resident-population-of-the-six-Olympic-host-boroughs/pdf/Six_borough_nkm_summary_population_analysis

London Borough of Waltham Forest. August. 2012. *A Brief Guide to Population Sources.* Research & Consultation Strategy & Communications.

Malthus, T. R. 1888. An essay on the principle of population: or, A view of its past and present effects on human happiness, Reeves & Turner.

Marmot, M. 2010. Fair Society Healthy Lives. *Inequalities in Health.* Oxford University Press.

Martin, D. 2006. Last of the censuses? The future of small area population data. *Transactions of the Institute of British Geographers,* 31**,** 6-18.

Mayhew, L. 2002. The neighbourhood health economy: A systematic approach to the examination of health and social risks at neighbourhood level. Actuarial Research Paper 144. Cass Business School, City University, London.

Mayhew, L. & Harper, G. 2010a. *Counting the population of Tower Hamlets – A London borough in transition. Report.* Available: http://www.towerhamlets.gov.uk/lgsl/901-950/916_borough_statistics/population_growth.aspx

Mayhew, L. & Harper, G. 2010b. *Counting with confidence: The population of Waltham Forest. Report.* Available: http://www.walthamforest.gov.uk/index/community/wf-statistics/mayhew-report.htm

Mayhew, L., Harper, G. & Waples, S. 2011. *Counting Hackney's population using administrative data – an analysis of change between 2007 and 2011.* Available: http://www.hackney.gov.uk/Assets/Documents/estimating-and-profiling-the-population-of-hackney.pdf

McDonald, N. & Williams, P. 2014. Planning for housing in England: Understanding recent changes in household formation rates and their implications for planning for housing in England. London: RTPI.

Mielck, A., Reitmeir, P. & Wjst, M. 1996. Severity of childhood asthma by socioeconomic status. *International journal of epidemiology,* 25**,** 388-393.

Milton, B., Whitehead, M., Holland, P. & Hamilton, V. 2004. The social and economic consequences of childhood asthma across the lifecourse: a systematic review. *Child: Care, Health and Development,* 30**,** 711-728.

Mitchell, R., Dorling, D., Martin, D. & Simpson, L. 2002. Bringing the Missing Million Home: Correcting the 1991 Small Area Statistics for Undercount. *Environment and Planning A,* 34**,** 1021-1035.

Moher, J. 2012. Local services now a postcode lottery. *The Guardian*, August 1.

Myrskyla, P. 1991. Census by questionnaire-census by registers and administrative records: the experience of Finland. *Journal of Official Statistics,* 7**,** 457.

Netuveli, G., Hurwitz, B., Levy, M., Fletcher, M., Barnes, G., Durham, S. & Sheikh, A. 2005. Ethnic variations in UK asthma frequency, morbidity, and health-service use: a systematic review and meta-analysis. *The Lancet,* 365**,** 312-317.

NHS. 2007. Tuberculosis prevention and treatment: A toolkit for planning, commissioning and delivering high-quality services in England. Report. Department of Health. Available: http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@en/documents/digitalasset/dh_075638.pdf

Nordholt, E. S. 2005. The Dutch virtual Census 2001: A new approach by combining different sources. *Statistical Journal of the United Nations Economic Commission for Europe,* 22**,** 25-37.

O'Neil, S. L., Barysh, N. & Setear, S. J. 1985. Determining School Programming Needs of Special Population Groups: A Study of Asthmatic Children. *Journal of School Health,* 55**,** 237-239.

Office for National Statistics. 2000. *Household Satellite Account (Experimental).* Office for National Statistics. London.

Office for National Statistics. October. 2003b. *Census strategic development review alternatives to a census: Linkage of existing data sources. Information Paper.* Available: https://www.ons.gov.uk/file?uri=/census/2011census/whywehaveacensus/onlyoptionfor2011/availablesources_tcm77-384209.pdf

Office for National Statistics. 2009. *UK National Statistics: Education* [Online]. Available: http://www.statistics.gov.uk/cci/nugget.asp?id=268 [Accessed Aug 2 2010].

Office for National Statistics. 2010a. 2009 mid-year population estimates frequently asked questions. Report.

Office for National Statistics. 2010b. Feasibility linkage of births records to school census records. Report.

Office for National Statistics. March. 2014. The census and future provision of population statistics in England and Wales: Recommendation from the National Statistician and Chief Executive of the UK Statistics Authority. Available: https://www.ons.gov.uk/census/censustransformationprogramme/beyond2011censustransformationprogramme/thecensusandfutureprovisionofpopulationstatisticsinenglandandwalesrecommendationfromthenationalstatisticianandchiefexecutiveoftheukstatisticsauthorityandthegovernmentsresponse

Office for National Statistics Geography 2007. Coverage of address registers for 2007 Census test, phase 1. Report.

Ohwaki, K., Hashimoto, H., Sato, M., Tamiya, N. & Yano, E. 2009. Predictors of continuity in home care for the elderly under public long-term care insurance in Japan. *Aging Clinical and Experimental Research,* 21**,** 323-328.

Openshaw, S. 1984a. Ecological Fallacies and the Analysis of Areal Census Data. *Environment and Planning A,* 16**,** 17-31.

Openshaw, S. The modifiable areal unit problem. 1984b. Geo Abstracts University of East Anglia.

Openshaw, S. & Taylor, P. 1981. The modifiable areal unit problem. In 'Quantitative Geography: a British View'.(Eds N Wrigley, R Bennett) pp. 60–69. Routledge and Kegan Paul: London.

Parcel, G. S., Gilman, S. C., Nader, P. R. & Bunce, H. 1979. A comparison of absentee rates of elementary schoolchildren with asthma and nonasthmatic schoolmates. *Pediatrics,* 64**,** 878-881.

Penneck, S. The future of using administrative data sources for statistical purposes. Proceedings of the Third International Conference on Establishment Surveys, Survey Methods for Businesses, Farms, and Institutions, 2007. 18-21.

Pharoah, R. & Hale, T. August. 2007. Behind the numbers: Migrant living patterns in Westminster. ESRO report. esro.

Poulsen, M. E. 1999. Maintaining the quality of the registers used in the Danish Census. *Statistical Journal of the United Nations Economic Commission for Europe,* 16**,** 155-163.

Redfern, P. 1986. Which countries will follow the Scandinavian lead in taking a register-based census of population? *Journal of Official Statistics,* 2**,** 415.

Redfern, P. 1990. A Population Register or Identity Cards for 1992? *Public Administration,* 68**,** 505-515.

Redfern, P., 2004. An Alternative View of the 2001 Census and Future Census Taking. *Journal of The Royal Statistical Society A,* 167(2), pp. 209-228

Resource Allocation Working Party. 1976. *Sharing resources for health in England.* Resource Allocation Working Party. London. Available: http://webarchive.nationalarchives.gov.uk/20130123205152/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/PublicationsPolicyAndGuidance/DH_4121873

Richards, M. & Wadsworth, M. E. J. 2004. Long term effects of early adversity on cognitive function. *Archives of Disease in Childhood,* 89**,** 922-927.

Roderick, P. J. & Connelly, J. B. 1992. The problems of monitoring tuberculosis in an inner-city health district: Integrated information is required. *Public Health,* 106**,** 193-201.

Savage, M. & Burrows, R. 2009. Some Further Reflections on the Coming Crisis of Empirical Sociology. *Sociology,* 43**,** 762-772.

Silverstein, M. D., Mair, J. E., Katusic, S. K., Wollan, P. C., O'Connell, E. J. & Yunginger, J. W. 2001. School attendance and school performance: A population-based study of children with asthma. *The Journal of Pediatrics,* 139**,** 278-283.

Simpson, L. 2007. Fixing the Population: From Census to Population Estimate. *Environment and Planning A,* 39**,** 1045-1057.

Simpson, L. & Brown, M. 2008. Census Fieldwork in the UK: The Bedrock for a Decade of Social Analysis. *Environment and Planning A,* 40**,** 2132-2148.

Spencer, N. 2000. *Poverty and child health,* Oxford, Radcliffe Medical Press.

StataCorp. 2007. *Stata Statistical Software: Release 10* [Online]. StataCorp LP. [Accessed].

Statistics Commission. 2004. Census and population estimates and the 2001 Census in Westminster: Final Report. London.

Steffey, D. & Bradburn, N. 1994. *Counting people in the information age,* Washington, National Academy Press.

Sturdy, P., Bremner, S., Harper, G., Mayhew, L., Eldridge, S., Eversley, J., Sheikh, A., Hunter, S., Boomla, K., Feder, G., Prescott, K. & Griffiths, C. 2012. Impact of Asthma on Educational Attainment in a Socioeconomically Deprived Population: A Study Linking Health, Education and Social Care Datasets. *PLoS ONE,* 7(11)**,** e43977.

Taras, H. & Potts-Datema, W. 2005. Childhood Asthma and Student Performance at School. *Journal of School Health,* 75**,** 296-312.

Taylor, W. R. & Newacheck, P. W. 1992. Impact of childhood asthma on health. *Pediatrics,* 90**,** 657-662.

Teather, S. 2007. Hansard House of Commons Debates 8 November 2007 column 352/353.

Thompson, G. 2010. P*rimary care trusts: Funding and expenditure*. Standard Note: SN/SG/5719. . London: House of Commons Library.

Tranmer, M. & Steel, D. G. 1998. Using Census Data to Investigate the Causes of the Ecological Fallacy. *Environment and Planning A,* 30**,** 817-831.

Tudor Hart, J. 1971. The Inverse Care Law. *The Lancet,* 297**,** 405-412.

Ulker, A. 2008. Household structure and consumption insurance of the elderly. *Journal of Population Economics,* 21**,** 373-394.

UNECE. 2011. Measurement of different emerging forms of households and families. Prepared by the Conference of European Statisticians Task Force on Families and Households. Available: http://www.unece.org/fileadmin/DAM/stats/publications/Families_and_Households_FINAL.pdf

Vale, S. 2006. *The use of administrative sources for economic statistics – an overview. .* Regional Workshop on Use of Administrative Data in Economic Statistics, Moscow. Available: http://unstats.un.org/unsd/economic_stat/Web/PDF/Overview%20-%20Vale.pdf

Van der Heyden, J. H. A., Demarest, S., Tafforeau, J. & Van Oyen, H. 2003. Socio-economic differences in the utilisation of health services in Belgium. *Health Policy,* 65**,** 153-165.

Varjonen, J. & Aalto, K. 2006. Household production and consumption in Finland 2001-Household satellite account.

Vaupel, J. W. 2010. Biodemography of human ageing. *Nature,* 464**,** 536-542.

Viel, J.-F. & Tran, A. 2009. Estimating denominators: satellite-based population estimates at a fine spatial resolution in a European urban area. *Epidemiology,* 20**,** 214-222.

Voas, D. & Williamson, P. 2001. The diversity of diversity: a critique of geodemographic classification. *Area,* 33**,** 63-76.

Wagner, K. D., Ritt-Olson, A., Soto, D. W. & Unger, J. B. 2008. Variation in Family Structure Among Urban Adolescents and Its Effects on Drug Use. *Substance Use & Misuse,* 43**,** 936-951.

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.

West, A. M., Mackintosh, N. J. & Mascie-Taylor, C. G. N. 1992. Cognitive and educational attainment in different ethnic groups. *Journal of Biosocial Science,* 24.

Westminster City Council. 2002. Evaluation of accuracy and reliability of 2001 Census.

White, H. 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica,* 48**,** 817.

Winkler, W. E. 2011. Matching and Record Linkage. *In:* COX, B. G., BINDER, B. N., CHINNAPPA, A., CHRISTIANSON, A., COLLEDGE, M. A. & KOTT, P. S. (eds.) *Business Survey Methods.* New York: John Wiley & Sons, Inc.

Zhang, L.-C. 2011. Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica,* 66**,** 41-63.