

## SECTION 8

# DATA SOURCES: AVAILABILITY AND PROBLEMS

Throughout this handbook we have tried to include detailed information on the sources for specific variables within a substantive discussion of those variables. The aim of this section is to provide an overview of some of the major sources of national data that are available for local areas and note some of the potential problems in using these sources. In choosing to concentrate on national data sets, we know that we fail to cover the many valuable sources of local data that will have been compiled by agencies such as Public Health Observatories and local authorities. Although beyond the scope of this handbook, sources of local data may be the best starting point for anyone wanting to make comparisons within relatively limited areas. The difficulty in trying to use such material for more extensive work is the lack of standardisation, especially with respect to methods of data capture, data definitions and data formats.

### 8.1 Introduction

A surprising range of data relevant to the mapping and analysis of inequalities are becoming increasingly available either free of charge, or for relatively modest payments.

The main suppliers of public data all have searchable websites intended for people who are not already familiar with the range of possible sources.

The range of potentially useful material rapidly reduces if one wants to investigate inequalities at the sub-local authority or sub-health authority level. For example, at electoral ward level, there is no detailed national information on crime or the environment. Many of the more interesting data sources are intended to provide between authority comparisons and are not collected for lower level analyses.

For these reasons, the Decennial Census is still an unparalleled resource.

In addition to the Census, most data used to map inequalities will come from one of three types of sources:

- ◆ Postcoded activity data from public services, other government bodies and some commercial organisations.
- ◆ Activity data from government and public agencies submitted in aggregate form for periodic returns and performance indicators.
- ◆ Data from large-scale surveys.

For the user, the main difference between these is their potential to provide data on small areas. Activity data containing full postcodes are rarely released in order to protect the confidentiality of individuals, but the data may be available aggregated to larger area bases. For example, means-tested benefits data are routinely available at electoral ward and local authority bases, and can be purchased for other areas such as health authorities. A local basket of health indicators has a range of indicators that can be created at a small area level (Section 8.4.11).

## **8.2 The Area Base**

### ***8.2.1 The Desired Base is not Always Available***

Aggregated activity data poses a number of problems for small area analyses because it is often only released at the administrative area base of the agency providing the returns. Typical area bases are local authority areas, school catchment populations and police districts. There are several methods for constructing smaller area estimates from these sources. Where the geographical base is uncertain, as in the case of school catchment areas, it may be possible to use mapping techniques to construct hypothetical boundaries between school areas. The devisers of the Index of Multiple Deprivation used these methods to provide ward level estimates of primary school data. Where the geographical base is known, as for many local authority returns, it may be possible to 'model down' this material to, say, ward level. Various statistical techniques can be employed, but they all rely on establishing a relation between the variable that is only available at the higher level and other variables that are also available at the lower level. Typically, if one can show a relationship at local authority level between Census data and the variable of interest, that relationship could be used to apportion the local authority values to sub-authority areas such as electoral wards. The process is fraught with a number of dangers, notably that the relationship used for the apportioning is distorted by the ecological fallacy or authority level supply factors.

There are a number of national household surveys with samples of 20–30,000, which are too small to provide reliable estimates below local authority level. Again, modelling techniques can be used to make ward level estimates. In this section, the modelling will be carried out with the individual level data from the survey, to produce a predictive model that can be used in conjunction with ward level data to provide ward level predictions for the variable of interest. Whenever possible, these models should also be used to model down local authority data, as they avoid many of the problems of models based on higher level data.

Changes in the boundaries of administrative areas can pose difficulties for data users. In the past decade, there have been major changes to local authority, ward, and health area boundaries. These not only pose difficulties for over-time comparisons but also can complicate the linking of contemporary sources. For example, unemployment and benefits claimants counts are currently not released on the same ward bases – the former is based on 1998/1999 ward boundaries and the latter on 1991 ward boundaries (Table 8.1).

Most developers of indexes of deprivation and other potential indicators of inequality are now interested in providing results for areas smaller than local authorities. The advantage of Census data is that they are generally suitable for this purpose. The 1991 Census data for England and Wales are available to both ward and ED level with two exceptions: certain tables are only included in the LBS set and not the SAS set (the LBS tables are only available down to ward level); and very small cell counts are ‘Barnardised’ – a form of noise is introduced (by adding  $-1$ ,  $0$ ,  $+1$  randomly) to improve anonymity.

Although results for the 2001 Census have been generated for local authority areas and wards, there is a new lower level set of building blocks ‘Output Areas’, constructed from contiguous postcodes to create areas with

**Table 8.1.** Approximate numbers of units and populations in different area bases

	<i>N in England</i>	<i>Approximate average population (persons)</i>
Region	8	6.3M
LAs with respect to social services counties/UAs	149	340K
LAs (districts)	354	150K
Strategic Health Authorities	28	1.8M
Primary Care Organisations	402	120K
Wards	~8640	5.8K
Postcode districts	2260	22K
Postcode sectors	8760	5.7K
Enumeration Districts (Census pre-2001)	110,000 (England and Wales)	490
2001 Census Output Areas	~ 150,000	100–125 households

a target of 100 to 125 households. Boundaries are drawn in order to maximise the homogeneity of populations within OAs. These new OAs offer the possibility of building customised OAs and there is much interest in using them to construct neighbourhood profiles and provide data to other sub-ward boundaries.

### **8.2.2 Postcode to Area Translation**

At present, the full postcode is the most commonly used standard geographical identifier in administrative data sets, though the Ordnance Survey Grid Reference is becoming a contender. In order to aggregate data to larger areas, some version of the National Postcode Directory (NPD) is required. The NPD is released by ONS (under license from the post office) in a number of formats including different sets of electoral and administrative area codes – a subset of this directory is available throughout the NHS as the NHS postcode directory. Other versions of the postcode directory can be purchased for commercial use. Thousands, or even millions of postcoded records can be converted to some other area code such as ward or local authority in a matter of minutes using these directories, a computer and a statistical package, such as SAS or SPSS. The only drawback, apart from the cost, is that the target data has to be fully postcoded. This is not always the case.

In administrative data sets, postcodes are often missing, incomplete, or invalid, though the quality of postcoding is improving as more systems use automated gazetteers for entering and validating addresses.

In some rural areas, a postcode may contain very few households and only one other piece of information, such as age or ethnicity may uniquely identify individuals. For reasons of confidentiality, there may be objections to releasing fully postcoded data, even if the aim is to aggregate the material.

Another problem can arise when the post office updates and expands the postcode base and administrative databases retain old postcodes that eventually drop-out of the postcode directory. A related problem is that there may be insufficient demand to justify the cost of mapping new postcodes to older area bases, as currently happens with 1991 wards.

Much of the new (non-Census) data being incorporated in deprivation indexes is not available with full postcodes, and is only reported for larger areas than those to which most indexes refer. The Index of Multiple Deprivation-2004 is a case in point. Although ward level values are published for this index, not all of its components are based on data that are genuinely available at ward level or lower. Various modelling or apportioning procedures have had to be used to estimate ward level values for these components.

Currently, there is considerable interest in supplementing or replacing postcodes with grid references as the basic locational identifier in administrative data sets. The Gridlink project is developing a database

of grid reference to administrative areas. More details of the project and associated products, such as Ordnance Survey's address-point can be found on the Ordnance Survey website.

### **8.3 Major National Archives and Sources**

#### ***Statbase (Office for National Statistics)***

The largest collection of freely available data sets is held by the ONS managed Statbase. It has to be noted that many of these sets are small single tables, showing data at regional or local authority level. In fact, many seem to be reformatted tables from official publications.

#### ***Neighbourhood Statistics (Office for National Statistics)***

This part of the National Statistics website contains ward and LA level data and is likely to be important for anyone wanting to conduct sub-authority analyses of inequality.

Data are free and can be downloaded as Excel sheets or CSV files.

Some key holdings are:

- ◆ Ward level data on all domains (though not the separate variables) of the IMD 2004.
- ◆ Ward level counts for a large number of means-tested benefits, including income support and family credit and benefits relating to disability.
- ◆ Ward level population estimates for 1998 (from the IMD team).

#### ***NOMIS***

NOMIS is an independently managed database of large-scale data sets, mostly on labour markets, many of which derive from the ONS. It also provides all the LBS tables for the 1991 Census.

Access charges have recently been removed for non-commercial users and most of the NOMIS data sets can now be accessed without charge after a simple registration procedure from ONS.

Some of the key data sets held at NOMIS are:

- ◆ All 1991 LBS Census tables at ward and higher levels.
- ◆ Current unemployment claimant counts.
- ◆ Data from the Labour Force Survey.

#### ***The Data Archive***

This is by far the largest collection of large (mainly survey-based) data sets in the U.K. As almost all data sets hold individual level data, users have to apply to obtain each data set, giving an account of its intended use, and pay a fee that reflects the handling charge for the type of media required. Note that some of these data sets are too large to fit on single CDs.

One of the strengths of the holding is that most projects funded by Economic and Social Research Council (ESRC) and related government monies have been required to lodge their data and an intelligible coding frame at the archive. It holds material such as the General Households Surveys and large health and lifestyle surveys.

Archive staff will also attempt to trace data sets that are not part of their current holding.

### ***Public Health Observatories***

Public Health Observatories were established in each of the nine English regions in order to strengthen the availability and use of information about health at a regional and local level by:

- ◆ Monitoring health and disease trends and highlighting areas for action.
- ◆ Identifying gaps in health information.
- ◆ Advising on methods for health and health inequality impact assessments.
- ◆ Drawing together information from different sources in new ways to improve health.
- ◆ Evaluating progress by local agencies in improving health and cutting inequality.
- ◆ Looking ahead to give early warning of future public health problems.

PHOs increasingly can assist by developing and coordinating analytical expertise across regional and sub-regional networks, sharing methodologies and avoiding duplication of effort as far as possible. PHOs also have a 'critical mass' of analytical skills.

Most of the material produced by the observatories is freely available via their websites. Some of these sites concentrate on reports of the patterning of local health, others also provide data that can be downloaded and re-analysed.

### **Local basket of health inequalities indicators**

The London Health Observatory has led a project to develop a series of indicators for local use of measuring progress in dealing with inequalities. The local basket of health inequalities indicators was released in October 2003. The basket contains an initial set of 70 indicators. It contains measures of health status or health outcomes, measure of the determinants of health, measures of access to services and process measures. The main purpose of the local basket of indicators is to help support local action to achieve the Government's national inequalities targets for life expectancy and infant mortality. The report on the indicator list and the indicators themselves are available on the London Health Observatory website ([www.lho.org.uk](http://www.lho.org.uk)).

## **Other General Sources**

*Local authorities* will employ teams that compile statistics on their area predominately for the purposes of targeting services and other aspects of resource management and allocation. But, as they do not normally regard the public provision of detailed information on sub-authority areas as part of their core function, the range of material that can be found on public websites is both patchy and limited. Nevertheless, they may be prepared to supply detailed information for specific purposes.

Many commercial organisations such as building societies and insurance companies hold important postcode level databases. There are cases of researchers being granted access to these data, but we imagine that uses would be very strictly controlled and likely to be expensive.

## **8.4 Data Sources on Specific Topics**

More details of these sources will be found elsewhere in the handbook, where the topics themselves are discussed. Again, data sets are only mentioned if they have national coverage.

### **8.4.1 Population Estimates**

Population counts are the denominators for many indicators of health and deprivation. During the past decade, ONS has produced rolling population estimates every two years based on the previous Decennial Census (and given some of the doubts of the coverage of the 1991 Census, the estimates were also rooted in the 1981 Census figures). Birth and mortality rates are two of the main factors used to generate the estimates, but they also take account of geographical mobility.

For a number of purposes such as resource allocation, population counts are required in advance of the biennial estimates; to meet this need, ONS provides population projections. In effect, the models used to generate the estimates are run forward, using the last biennial estimate as the base. By this method, projections are prepared for up to three or for years from the last estimate.

The main sources of population estimates and projections are as follows:  
On the NOMIS site:

- ◆ 1991 Census counts (LBS tables only).
- ◆ Resident population estimates for local authorities.
- ◆ Ward population estimates are not routinely released by ONS, but the NOMIS site contains the denominators for the Index of Multiple Deprivation-2004.

### **Other sites/sources**

Full versions of the Census data are held at a number of sites and can be also purchased for local use. The English academic community has traditionally accessed Census data via MIDAS – a data set system hosted by the University of Manchester computing service.

The Oxford Group is currently running a national project to improve small area estimates for the numbers of elderly people, a group whose numbers are not always accurately estimated in simple inter-censile projections.

The Compendium of Health Indicators produced by NCHOD for Department of Health includes ONS estimates for health authorities in five year age–sex bands (also on Statbase). It also contains figures for the populations of primary care trusts and groups, based on primary care registrations data, reconciled to ONS estimates and projections at the health authority level.<sup>7</sup>

Results from the 2001 Census were released in late 2002. Access to Census material has radically changed with the 2001 project. Although much of the 1991 material could be got from public websites towards the end of the 1990s, the 2001 results are freely available from the start, though there are handling charges for very large data requests on CD, and additional charges for material specially aggregated to client-defined areas.

The best introduction to the availability of the 2001 Census is the Census Output Prospectus. It is downloadable via the Census links from the National Statistics website. It describes the OAs, and methods for obtaining data.

In addition to the Prospectus, critical documents include the Output Classification Manual (which describes the data definitions and classifications used in the reporting) and various downloadable manuals that list output tables and area options. Although this documentation will be invaluable to regular and heavy-duty users, great efforts have been made to provide a user-friendly graphical interface, so that many users will find that they can get all they want by navigating the website without downloading supporting documents.

### **8.4.2 Health**

#### **NHS activity data**

Two factors have had a major impact on the ability of the NHS to provide data on patient care: first, NPfIT which provides a framework for the development of NHS information systems; second, structural changes, such as the re-structuring of health authorities and regional offices and the increasing role for primary care trusts and the resulting impact on community trusts.

---

<sup>7</sup> The Reason for this is that there is substantial 'list inflation' so that overall, primary care registrations are ~ 6% higher than population estimates.

### General practice

There is one officially supported data warehousing scheme, now known as the General Practice Research Database, which is maintained by the Medicines and Health Care Regulator for the Department of Health.

The one comprehensive national resource on general practice activity covers *prescribing*. Details are available from the PPA or the Prescribing Support Unit (PSU).

For several decades, the main national source of data on English general practice have been the surveys that have supported the series of Morbidity Survey in General Practice publications. These data sets are lodged at the data archive.

The triage database systems used by *NHS Direct* are a potentially rich source of information on population morbidity.

There are many examples of local projects successfully approaching general practices and primary care groups and trusts for data on the incidence of specific problems or conditions. There may be no alternative to such local approaches if details on the severity of symptoms or clinical outcomes are required. The most promising conditions are those for which standard care management protocols are established, such as diabetes and asthma.

Some information on *dental care* can be found in the Korner statistics on the Department of Health website. Low level data from the Adult Dental Health Survey and the *Children's Dental Survey* may be available on request from the Department of Health.

The same source should be contacted for access to low-level versions of the 5% sample of dental treatment claims.

### The secondary sector

There are two main types of activity data routinely collected and made available for secondary care in the NHS: Hospital Episode Statistics (HES) and Korner data.

Sectors for which information is particularly patchy are the former community health trusts and mental health trusts. Korner returns are the only consistent source of data from these trusts. Several national disease registers are either established or in development, but the cancer registry is probably the only one with national coverage at present.

Access to HES data has much improved with the establishment of an HES site within the Department of Health website and an HES enquiry desk. The HES data dictionary and guide to accessing the data (both can be downloaded from the same source) are good starting points. Many tables of results can be freely downloaded from the website; alternatively, customised requests will be accepted, subject to the usual restrictions to protect patient confidentiality. Most requests can be processed through a

low-cost fast-track scheme and free estimates are provided for all requests. Public Health Observations are now providing a HES service to their local public health communities.

The HES and PAS data systems provide the foundation for another major collection of activity statistics, with attached costs. The Healthcare Resource Groups reference data set (available on CD – though access outside the NHS may be restricted) presents patterns of activity for both secondary and some primary care institutions. The PAS/HES material has been processed by a package called ‘Grouper’ – maintained by the NHS Information Authority (NHSIA). Its aim is to group procedures and episodes of care, within specialities, by the demands they make on healthcare resources. Healthcare providers are then required to estimate the costs of each of these groupings and the reference costs CD reports the age costs by HRG both nationally and by individual trusts. Coverage of all procedures and episodes of care is not yet complete, but the current database is very extensive. The NHSIA website reports progress on extending HRGs beyond the present procedures. The NHS accounting manual (from the Department of Health main site) lists the HRGs that are presently in use for the costings.

### **8.4.3 Health Surveys**

Low level data from most of the main health and health and lifestyle surveys, as well as more general surveys with supplements on health, are lodged with the Data Archive. These include: HSE, the Surveys of Psychiatric Morbidity in Great Britain, GHS and the Omnibus Survey (more details of these can be found in [Section 4.2](#)). The SEPHO Lifestyle Toolkit contains online information about lifestyle surveys undertaken.

### **8.4.4 Social Care**

Social services departments in England make a number of annual returns on the services they provide and the numbers of clients. All these data are only available at the level of Local Authorities Responsible for Social Services (LAWRSS) of which there are approximately 150 in England. Some central returns for children’s social services now require postcoded data on individual clients and it may be possible to negotiate access to an anonymised ward-based version of this. However, these can be very sensitive data – especially details from the child protection register – and permission may not be granted for ward-level access.

Key Statistics (KS1) is the first and central source for data on personal social services in England.

These are some of the main pieces of information that are available on PSS activity. All are presented by LAWRSS area.

For children:

- ◆ Numbers 'being looked after' sub-divided into numbers in different types of care: e.g. residential homes, secure accommodation, fostering and being placed back with families under supervision orders.
- ◆ Children with disabilities and special needs.
- ◆ Numbers of places in children's homes.

For adults and older adults:

- ◆ Numbers of people supported by LAWRSS in residential and nursing homes (also the number of homes and potential places).
- ◆ Numbers of people receiving domiciliary care.
- ◆ Numbers of people with learning disabilities.

Benefits data (for example on disability) may be a better source of information on the geographical distribution of some of these groups.

#### **8.4.5 Housing**

Apart from the Census, very little information on housing is available at sub-local authority level. These are several national surveys, but their samples are too small for sub-LA breakdowns. It is possible to *model* down the survey results, as, for example, in the IMD and Welsh Deprivation Index.

##### Physical condition of housing

The Census includes questions on amenities such as baths, showers and central heating and the data on numbers of people in the household and the number of rooms occupied are used to compute a measure of overcrowding.

The main sources on the structural condition of housing are two infrequent surveys and local authority returns.

The English House Condition Survey (there are parallels for Wales and Scotland) includes a professional assessment of physical condition and a valuation, as well as an interview with the residents. It is based on a sample of 25,000 dwellings and is repeated every five years. Half of the properties in the 1991 survey were reassessed in 1995 in order to record any changes. The data sets are available from the ODPM.

The ODPM conducts a second national survey, the Survey of English Housing. Here, the emphasis is on the type of accommodation, tenure, the experiences of the household in finding accommodation, moving and their views of the accommodation and the area. The survey is repeated annually and is based on a sample of 20,000 households.

Local authority returns for the HIP give some data on housing stock, vacancies, lettings and homelessness. More specific information is

provided by local authority returns on the numbers of unfit dwellings and the reason for their being classified as unfit.

The Housing Needs Index is another LA level data source. It is based on (amongst others) data from the Survey of English Housing, the General Households Survey and the English House Condition Survey.

### Housing – tenure

The Survey of English Housing collects information on tenure, but the Census is the main source here. The 1991 Census asked if rented accommodation was furnished or unfurnished. In 2001, this question is only asked in Scotland.

There are several sources relating to particular types of tenure. Housing association statistics (collated by the ODPM) give details of the numbers of lettings and new lettings, also a considerable range of information on the tenants. This information is released for local authority areas.

Local authority housing performance indicators (ODPM) will have some information on tenure.

### **8.4.6 Employment and Unemployment**

Both of the main English sources of unemployment data – the claimant counts and the Labour Force Survey – can be found on the NOMIS website.

*Unemployment benefit claimant counts* are presented in a number of forms down to ward level. The main options for counts are: counts (age and duration of unemployment), seasonally adjusted counts, unadjusted counts – with rates. The same set of options are available for claimant flows and there is an additional data set of claimant ‘off-flows’, showing the reason for ceasing to be a claimant in addition to the claimant’s age and duration of the most recent claim.

The Labour Force Survey provides an alternative estimate of unemployment independent of the periodic changes in the criteria for claiming unemployment benefit. It is a quarterly sample survey and collects information on personal circumstances and employment status of respondents. The full data sets are lodged with the Data Archive; subsets of recent data are available on-line from NOMIS.

The Annual Employment Survey replaced the annual Census of Employment in late 1995. This survey of approximately 130,000 businesses collects information on the nature of the business, the gender of employees, the types of jobs and whether they are full- or part-time. Data are produced by ward and there is considerable detail on the type of business activity. It is available on-line from NOMIS, but there is a special registration procedure for this data set which requires a statement of intended use.

Other relevant data sets held by NOMIS include details of job centre vacancies, broken down by occupation and industry, but most of these are only available at unitary authority level or higher.

### **8.4.7 Environment**

Local authority returns to the ODPM provide some information here, but there are no national small area data sets.

### **8.4.8 Crime**

Neither of the two main sources of data on crime in England are available at ward level and for this, and related reasons, the compilers of the ODPM reluctantly omitted a crime domain.

The two sources are the annual British Crime Survey and the quarterly returns from police forces to the Home Office. The self-report data from the British Crime Survey can be disaggregated to ACORN type of community within each government region. The annual recorded crime data is available at the basic command unit level (there are about 300 such units per local authority).

Home Office data on notifiable offences recorded by the police are available on the Neighbourhood Statistics Area of the ONS website.

Whenever crime data are presented by electoral ward, the material will have been obtained directly from local police forces. The West Midlands MAIGIS project contains such data on a variety of offences including burglary and car crime for the local area.

The Scottish Area Deprivation Index uses house contents insurance rating at postcode sector as a proxy for crime. Similar data may be available from insurance companies in England.

### **8.4.9 Education**

A limited amount of data on education is provided by the Decennial Census, available from NOMIS. It asks people aged 16 and over to list their educational attainments, and there are boxes to record whether or not each person in the household is a full-time student. The Census also includes a question on professional and vocational qualifications. From 2001, the Census material is available via the main ONS site.

An indirect estimate of those staying in full-time education after the compulsory leaving age can be obtained from child benefit data, which is broken down by the age of child for those aged 16 and over. The Labour Force Survey (from the Data Archive) collects data on adult qualifications.

The DES collects data on educational attainment and absenteeism, but these are nearly always recorded by local educational authority, or educational institution. In the latter case, the catchment area has to be inferred, as there are no published details of the distribution of pupil postcodes. The compilers of the IMD index use GIS techniques to approximate ward maps of primary school catchment areas, but admit that this method is unlikely to be valid for the much larger and complicated

secondary school areas. IMD ward-level estimates of *primary school pupils at Key Stage 2* are available on the ONS Neighbourhood Statistics site.

The Universities and Colleges Admissions Service keeps postcoded records of university applicants with details of the outcome of the application. Ward-level university admissions data can be found on the ONS Neighbourhood Statistics site.

#### **8.4.10 Income and Benefits**

The New Earnings Survey and Index of Average Earnings are two of the major sources of information on pay and income from work. The former is based on a 1% national sample of employees whose tax is handled via PAYE. Area analyses are available on the NOMIS site. The Index of Average Earnings results are available from the ONS.

Ward-level data on various means-tested and disability-related benefits can be freely downloaded from the ONS Neighbourhood Statistics site.

Income support data is provided by age and household structure in receipt of pensioner, disability or lone parent premiums. Data are also available on numbers of claimants and dependants of claimants. Other counts available at ward level include:

- ◆ Family credit claimants.
- ◆ Attendance allowance claimants.
- ◆ Disability living allowance.
- ◆ Jobseeker's allowance.
- ◆ Incapacity benefit.
- ◆ Severe disablement allowance.

#### **8.4.11 Composite Indicators and Geo-classification Systems**

The ONS Classification of local and health authorities, revised for 1999, can be downloaded from the ONS website, or purchased in hard-copy form.

The classification of wards, based on the same system must be purchased from ONS.

Details of the ACORN and SuperProfiles classification systems can be found on-line at [www.CACI.co.uk](http://www.CACI.co.uk) and [CLARITAS.co.uk](http://CLARITAS.co.uk) respectively. Descriptions of both are available for free; the databases are for purchase. A licensed version of the higher-level SuperProfile classification may be locally available from the Department of Health.