Technical Briefing 6

August 2009

Using small area data in public health intelligence

Purpose

This is the sixth in a series of technical briefings produced by the Association of Public Health Observatories (APHO), designed to support public health practitioners and analysts and to promote the use of public health intelligence in decision making.

In this briefing we assess the types of small area data available to analysts and local policy makers across the UK and Ireland, reflecting, as much as possible, the differences between the constituent nations. We describe a range of analytical and presentational methods that may be used and some of the analytical issues that may be encountered.

Further materials, including tools to support our technical briefing series, are available through our website at http://www.apho.org.uk

Contents

Introduction	2
Types of small area	2
Analytical issues	6
Analytical solutions	8
Further resources	11
Glossary and abbreviations	11
References	12

Authors

Jo Watson
Stacey Croft
Heather Heard
Philip Mills

Contributors

Jennifer Bishop
Alison Burlison
Anne Cunningham
(NHS Blackburn with Darwen)
Mark Dancox
Lorraine Fahy
Paul Fryers
Cheryl Heeley
Clare Humphreys
John Langley
James Nelson-Smith
Stuart Simms
(NHS Stockton-on-Tees)
Diane Stockton

Introduction

Each nation within the UK and Ireland has primary geographical units for local administration of public services and, naturally, data tend to be widely available to describe and compare these populations. However, these areas can have widely varying, and sometimes quite large, populations. For example, Hampshire and Surrey Primary Care Trusts (PCTs) both have populations of over one million, as does one local authority (LA), Birmingham City. Such large populations are very diverse and contain communities with different characteristics. Detailed knowledge of the demography, socio-economic structure and usage of services in small areas is required in order to assess their needs and to provide appropriate services.

For the purposes of this briefing, small areas are taken to include any geography below LA district in England and Wales; LA or NHS board or community health partnership (CHP) area in Scotland; local government district (LGD) in Northern Ireland; or county in the Republic of Ireland.

Types of small area

There are four principal categories of small area which can be used to produce data: i) Census geography, ii) electoral areas, iii) postal geography and iv) other ad hoc local areas, such as service delivery or catchment areas, neighbourhoods, and target areas for special initiatives. This briefing concentrates on area-based populations, but clearly there are other non-geographical ways of defining populations, e.g. users of a particular service (patients registered with a GP practice, children attending a particular school, etc.) or people with particular characteristics (aged over 65, Asian ethnic origin, etc.). Many of the analytical issues discussed apply also to non-geographically defined populations.

Census geography

Census geography is the hierarchical set of areas defined for the release of 2001 Census data in the UK. They include Census output areas (OAs) and super output areas (SOAs) in England, Wales and Northern Ireland, and OAs, data zones and intermediate geography (or zones) in Scotland. A summary of small area Census geography is provided in Table 1.

Output areas were introduced in Scotland in 1991 and in England, Wales and Northern Ireland in 2001 and are the smallest areas in common use across the UK. Prior to this, UK Census data had been published at enumeration district level, where enumeration districts were the groups of properties covered by single Census enumerators (the people who delivered and collected the Census forms). OAs were defined by computational analysis to be as homogeneous as possible, within the constraints set on population size, and the intention was that they would reflect natural communities as much as possible. Most of the Census results are published at this level, with small area counts adjusted to prevent the disclosure of individuals' responses from the figures.

Following the 2001 Census, the Office for National Statistics (ONS) introduced the SOA, which was intended to provide a stable, permanent geography that could be used to publish a wide range of statistics on a consistent basis. In order to allow for their use in a wide range of applications, three levels of SOA were proposed for England and Wales: lower super output areas (LSOAs) were generated automatically by ONS, constrained to the electoral ward boundaries at the time (see later) and hence could (and still can) be grouped to be coterminous with LAs; middle super output areas (MSOAs) nest within LAs and are not constrained by ward boundaries; upper super output areas have never been defined but were planned to be grouped to be coterminous with LAs.[1]

The first data to be published at SOA level were the Indices of Multiple Deprivation 2004. These indices, revised in 2007, provide data on various aspects of socio-economic status at small area level and provide very useful information to support the assessment of health needs and interpretation of health outcomes.[2]

The Scottish Government has developed data zones as a relatively consistent and stable geography, and statistics across a number of policy areas are readily (and regularly) available at this level for use on the Scottish Neighbourhood Statistics website.[3] The data zone geography covers the whole of Scotland and nests within LA boundaries. Data zones are groups of OAs and some effort has been made to respect physical boundaries. In addition, they have compact shape and contain households with similar social characteristics. Aggregates of data zones called intermediate zones were introduced when it became clear that not all statistics are suitable for release at the data zone level because of the sensitive nature of the statistics or for reasons of statistical reliability.

The Northern Ireland Statistics and Research Agency (NISRA) has also developed SOAs (of which there is only one level) as the core geography to improve the reporting of small area statistics.

The Republic of Ireland does not currently have a small area geography corresponding to those in the UK, but plans to introduce a new small area geography for their 2011 Census (Censuses are undertaken every five years).

Electoral areas

These include wards, civil parishes and parliamentary constituencies in England, Wales and Scotland; wards and parliamentary constituencies in Northern Ireland; and electoral divisions in the Republic of Ireland.

The electoral ward's primary purpose is to provide the constituencies for local elections; each ward elects one or more councillors to the LA. However, until the publication of the 2001 Census, wards were also used by ONS and others as the standard geography for data publication below LA level. Each LA is made up of a number of wards and, while wards vary hugely in size between districts (from a few hundred residents to over 30,000), they tend to

be reasonably consistent within a district, to ensure residents are fairly represented by their local councillors. For the same reason, they are subject to regular review and in areas with substantial levels of development or population movement they may undergo frequent change.

Table 1: Summary table of Census/electoral small areas

Area name	Total number of areas in 2001	Minimum size in 2001	Average size in 2001
Census output area (OA)	165,665 (England) 9,769 (Wales) 5,022 (NI) 42,604 (Scotland)	40 households, 100 residents (E&W, NI) 20 households, 50 residents (Scotland)	150 households, 400 residents (E&W) 125 households, 340 residents (NI) 120 residents (Scotland)
Lower super output area (LSOA), SOAs in NI	32,482 (England) 1,896 (Wales) 890 (NI)	1,000 residents (E&W) 1,300 residents (NI)	1,500 residents (E&W) 1,900 residents (NI)
Data zone *(Scotland only)*	6,505	500 residents	800 residents
Middle super output area (MSOA)	6,780 (England) 413 (Wales)	5,000 residents (E&W)	7,200 residents (E&W)
Intermediate zone *(Scotland only)*	1,235	2,500 residents	4,200 residents
Upper super output area (planned but not defined)	–	–	*25,000 residents (E&W)*
Electoral ward (2009)[1] (multi-member ward in Scotland)	7,943 (England) 881 (Wales) 582 (NI) 353 (Scotland)	<100 residents (England) 763 residents (Wales) 761 residents (NI) 1991 residents (Scotland)	6,400 residents (England) 3,400 residents (Wales) 3,000 residents (NI) 14,600 residents (Scotland)
2003 statistical ward[2]	8,868 (E&W)	<100 residents (England) 783 residents (Wales)	6,000 residents (E&W)
Census Area Statistics (CAS) ward[3]	8,850 (E&W) 1,222 (Scotland)	40 households, 100 residents (E&W) 650 residents (Scotland)	6,000 residents (E&W) 4,100 residents (Scotland)
Standard Table (ST) ward[4]	8,800 (E&W) 1,176 (Scotland)	400 households, 1,000 residents (E&W & Scotland)	6,000 residents (E&W) 4,300 residents (Scotland)
Parish[5]	10,397 (England) 868 (Wales) 871 (Scotland)	– 51 residents (Scotland)	3,000 residents (E&W) 6,000 residents (Scotland)
Electoral division (2006)[6] *(Republic of Ireland only)*	3,440	76 residents	1,230 residents

Notes:
1 2009 wards, 2007 mid-year population estimates (Source: ONS Experimental Population Estimates & Scottish Neighbourhood Statistics) except NI (2001 populations – wards unchanged between 2001 and 2009).
2 Equivalent to electoral wards in place or in statute by 31 December 2002 and used for publication of 2001 Census data.
3 Variant of statistical wards with particularly small wards (fewer than 40 households or 100 residents) merged to protect data confidentiality, created for 2001 Census Area Statistics outputs. No requirement in NI as all electoral wards exceed the threshold.
4 Further variant of statistical wards such that those with fewer than 400 households or 1,000 residents have been merged. Used to publish 2001 Census Standard Table outputs. No requirement in NI.
5 Defined for rural areas of England and Wales only, as at 1 April 2003. Known as communities in Wales.
6 Wide range of population from 76 to 32,051.

The electoral wards that were in place or on the statute books at the time the 2001 Census was published have continued to be used for data publication by ONS, to provide comparability over time, and are referred to as (2003) statistical wards. Census Area Statistics and Standard Tables were released using slight variants (CAS wards and ST wards) which merge the smallest statistical wards.[4] Civil parishes are now defined only in rural areas and are not commonly used outside a very local context.

UK parliamentary constituencies have an average total population of about 90,000. They are not coterminous with either LAs or PCTs, but are used for a range of statistics of particular interest to members of parliament and constituency organisations.

In the Republic of Ireland, electoral divisions (EDs) are currently the most disaggregated geographical areas for which Census data are available. However, they vary greatly in population from 76 to 32,051, which makes it difficult to use them to assess the needs of local communities. Work is being carried out to develop a new small area geography for the Republic of Ireland that will nest into the existing EDs. Data from Census 2011 will be released at this new small area level.

Postal geography

While postal geography is primarily designed for delivering mail, addresses can provide a point location for individuals or events, which can be used to place them within other geographies. Full addresses are most valuable, but are often left off records for reasons of confidentiality, and can be difficult to work with because of the scope for erratic spelling and formats. All UK addresses have a postcode. Individual postcodes are shared, on average, by 15 addresses and may cover between one and around 100 addresses within a single street, giving the smallest level of aggregation of households routinely available. Postcode geography divides the whole country into segments, each containing all the addresses with the same postcode. These can be aggregated to sectors, districts and areas (see Box 1).

Box 1

UK postcodes take the format AA99 8BB, e.g. YO10 5DG.
AA is a one- or two-letter abbreviation of the main post town (or part of London) defining the postcode area, e.g. YO – York, M – Manchester, EC – East Central London.
99 is a one- or two-digit number defining the postcode district, e.g. YO10, M60, or, in London, can be one digit and one letter, e.g. W1A, EC1A.
8 is a one-digit number defining the postcode sector, e.g. YO10 5.
BB are two-letters defining the full unit postcode, e.g. YO10 5DG, M1 1AA, EC1A 1BB.

For more information about the parts of the postcode (including 'outward code' and 'inward code', full rules and exceptions) see the Cabinet Office website.[5]

It is currently difficult to analyse data accurately at postcode level, because of the lack of reliable population estimates required to calculate rates and allow comparison between areas. Hence postal geography itself is little used in public health intelligence. However, in the absence of full addresses, postcodes themselves provide a vital link to other geographies for many individual-level data sources, such as death registrations, hospital admissions and public service registration systems (GPs, schools, etc.) which include postcodes. It is therefore very important to ensure postcodes are recorded accurately on as many datasets as possible. Lookup tables, such as the NHS Postcode Directory,[6] are used to allocate individual records to any of the administrative or electoral geographies referred to in this briefing. Postcodes are not permanent and change in accordance with the needs of the postal system, sometimes being re-used for different addresses. Hence, care should be taken when analysing historical postcode data.

Commercial social marketing companies use data from financial and retailing sources to characterise the residents of a full postcode and group these into categories accordingly. This type of analysis is known as geodemographic segmentation and its uses in public health intelligence are discussed in *Technical Briefing 5: Geodemographic Segmentation.*[7]

The Republic of Ireland does not have a postcode system which covers the whole of the country.

Table 2: Average population size for postcode geographical areas

	Average population		
Area name	**England & Wales**	**Scotland**	**Northern Ireland**
Postcode area (e.g. YO)	500,000	400,000	1,750,000
Postcode district (e.g. YO10)	22,500	12,000	21,500
Postcode sector (e.g. YO10 5)	600	600	600
Full postcode (e.g. YO10 5DG)	40	40	40

Service delivery or catchment areas

Many local services define catchment areas to plan and manage demand for services. These include GP practices, schools, social services and Sure Start. Other services may divide their area of responsibility into zones for teams to cover, e.g. police beats, health visitors' patches, LA neighbourhood services, etc. Catchment areas will often overlap with each other and the boundaries may have been defined without considering data availability. This means they may not be coterminous with Census geography, have a limited range of datasets available, and there may be no demographic data to enable rates to be calculated and compared robustly.

GP practices have to define catchment areas, within which they are willing to register residents as new patients. However, the practices will also have registered patients who live outside those areas. In general, the catchment areas have large overlaps so most households live within the catchment areas of several practices, giving choice to patients. This makes GP practice catchment areas unsuitable for comparative analysis. When analysing data at practice level it is preferable to use the registered practice population, rather than a geographically defined one, but this type of analysis is limited to datasets which include the individual's GP practice. The introduction of the Quality and Outcomes Framework (QOF)[8] for GPs has led to the publication of some aggregated data at practice level, but in general there is very limited availability of data.

In England and Wales, practices are grouped into practice-based commissioning groups, but these are defined in very different ways, even within a single PCT, and range in size from single practices to whole PCTs. GP practices range in size from small single-handed practices with only a few hundred patients to those with over 25,000 patients, and constantly change: practices open, close, merge and split, and GPs move between practices, sometimes taking their registered patients with them. This makes the monitoring of changes over time in practice-based needs or outcomes particularly difficult.

In some circumstances we may wish to calculate an approximate geographical area for the practices, representing the localities where their patients actually live, rather than their nominal catchment areas. This can be done in a number of ways, as discussed in the section on aggregation on page 8.

Many special initiatives, such as Sure Start in the UK and New Deal for Communities in England, target defined geographical areas. The areas are defined in various ways, not necessarily tied to any geography for which population data are available, and do not comprehensively cover the whole country, or even the whole of a LA.

In many LAs, neighbourhoods have been defined for the organisation of local services, analysis of needs and inequalities, or community engagement. Again, these areas are not necessarily defined in ways that allow denominator populations to be estimated accurately, so comparative analysis can be difficult.

Table 3: Summary attributes for different types of small area

Attributes	Census areas	Electoral areas	Postcode areas	Catchment areas
Sub groups	OA, LSOA, MSOA, data zone, intermediate zone	ward, parish, constituency	area, district, sector	GP practice, school
Are decision makers familiar with the concept?	No	Yes	Yes[1]	Yes
Can areas be easily defined geographically?	Yes	Yes	Yes	No
Is there hierarchical consistency allowing data to be published at the most appropriate level?	Yes	Yes	Yes	No
Are these areas of similar size in terms of resident/ registered populations?	Yes	No	No	No
Are reliable population data available for these areas?	Yes	Yes	No	Varies
Do these areas remain constant over time?	Yes	No	No	No
Do issues of disclosure prevent publication of data?	Some (OAs, LSOAs, data zones)	Some (smaller wards)	Yes	Some (e.g. small GP practices)

Note:
1 Postcodes are familiar to most people, but the areas and hierarchy are not necessarily well-known.

Analytical issues

Analysing data at the smallest area level increases the granularity of the results, giving several advantages:

i) more precise description of individuals in the area

ii) minimising ecological effects

iii) more stark description of inequalities

iv) more accurate targeting of interventions.

As mentioned in the previous section, data may not be available to be attributed to small areas (e.g. where data are published aggregated to a higher level of geography). Furthermore, variations in data quality or completeness become more important factors as the numbers get smaller. A key example is population data, required as denominators for comparison of areas through calculation of rates and discussed below. However, even where the availability and format of data technically allow very small area analysis, there can be problems that prevent such analysis or render the results unreliable or even meaningless.

The most obvious and common problem with small areas is a simple function of their small populations: as the size of the population reduces, so does the reliability of statistics calculated for those areas. Hence it becomes increasingly difficult to observe any statistically significant differences between areas, as the size of random variation in the area statistics can mask the underlying differences. Similarly, real changes over time can be hidden by massive year-to-year random variation. It is particularly important to present small area data with confidence intervals or p-values to avoid over-interpretation of apparent differences (see *Technical Briefing 3: Commonly used public health statistics and their confidence intervals*[9] for further information).

Confidentiality and disclosure

When publishing health data, it is a primary concern that individuals should not be identifiable from the data, either directly or indirectly. If data are published for very small populations, small counts can, in certain circumstances, have the effect of disclosing information about individuals. For example, if there is known to be only one elderly man in a Census OA, and a published breakdown of the ages of residents gives a '1' for one particular age, with zeros for all other categories around it, local people can deduce the age of that neighbour. This would undermine the confidentiality of the Census. For many datasets analysed within the health and LA sectors, data disclosed could be a great deal more sensitive than just a person's age. Hence a great deal of effort goes into disclosure control when data are published for small areas. The usual approach is simply to suppress counts of, say, three or less and rates derived from them, but there are many alternative methods. The UK Census 2001 used a combination of record-swapping (randomly muddling the individual records prior to any analysis to a degree that would not affect aggregate analyses), minimum thresholds for population and households and rounding of cell counts.[10] Specific guidelines are often set by data owners and vary in detail. Some data owners, such as UK cancer registries, also specify a minimum denominator population for publication of results.

Guidelines set by data owners should be carefully observed, but in the absence of specific guidelines it will generally be safe to suppress cells in tables that are based on fewer than five individuals, and any other cells which might enable data users to derive suppressed results. In most situations, results based on fewer than five cases will have very wide confidence intervals and hence be uninformative in any case. This would not be true when analysing potential clusters of very rare conditions, where even two or three cases together may be highly significant, but while such data may be appropriate for public health surveillance, they should not be made available outside the appropriate authorities (note that the Freedom of Information Act 2000[11] does not require data to be made available if there is a risk of disclosing confidential information).

It is important to ensure that suppressed values cannot be derived from other data in a table, e.g. if a column of data has one suppressed value, but the correct total, then the suppressed value can easily be obtained by subtraction. This is a simple example of an effect called disclosure by differencing. In fact, reasonable efforts must be made to ensure that no suppressed values can be derived through combination of the data being published and any other published data. A similar situation arises when comparable data are published for a range of different geographies. For example, where results are published at both ward and LSOA level, by differencing the results for areas with slightly different boundaries it may be possible to produce information about an identifiable individual. It is wise to be very wary of producing the same datasets on the basis of more than one type of small area and be aware of other publications of the same data.

More detailed confidentiality guidance has been published by ONS,[12] and by NHS National Services Scotland in the ISD Statistical Disclosure Control Protocol.[13]

Population denominators

UK resident population estimates are published annually (usually at least a year behind) for administrative and electoral area residents. Since the underlying source is the decennial Census, the population estimates have the limitations of all Census datasets, including the reliability of the ages and ethnicity recorded and the under-recording of people in categories such as young adults, travellers, homeless people and illegal immigrants. Figures are adjusted to take account of these problems, but the effectiveness of the adjustments is difficult to assess.

The Census figures have to be adjusted further to take account of the ageing of the population, births, deaths, migration and other changes (such as movements of armed forces). Some additional information is provided by GP practice registers and electoral rolls which are updated

annually. However, in areas with high levels of population movement, such as inner-cities and university cities, practice populations and electoral returns are unlikely to be more reliable than Census returns for the categories mentioned.

Scottish data zone population estimates are published by sex and five-year age band; in England and Wales, MSOA estimates are published with this level of detail, but LSOA estimates are published only by broad age groups. In Northern Ireland, mid-year population estimates are routinely published to local government district and parliamentary constituency level. Population estimates for OAs and SOAs were generated for mid-2003 in order to construct the Northern Ireland Multiple Deprivation Measure 2005. However, these figures are rounded totals, not broken down by age or gender. Later in 2009 NISRA intends to publish updated mid-year estimates by OA and SOA, in order to construct the Northern Ireland Multiple Deprivation Measure 2008.

GP patient registers tend to produce higher numbers than resident population estimates: for England and Wales the aggregate of practice populations is about 5% higher than the aggregate of ONS mid-year population estimates. This difference is thought to result from a combination of delays in updating practice records, particularly in areas with large ethnic minority populations, and the inclusion of some people who may have been omitted or excluded themselves from the Census. Patient registers are also subject to the same problems as Census results in areas with high levels of population movement.

Numbers, rates and ratios

In public health terms, rates are generally of greater interest than numbers because different populations can be compared with each other and with national or regional averages and the statistical significance of the results can be assessed. Methods for calculating and comparing rates are described fully in *Technical Briefing 3*[9] and all of the principles apply equally to small area data. However, there are some issues that arise when using these methods for small area analysis which are highlighted here. These comments should be read in conjunction with *Technical Briefing 3*.[9]

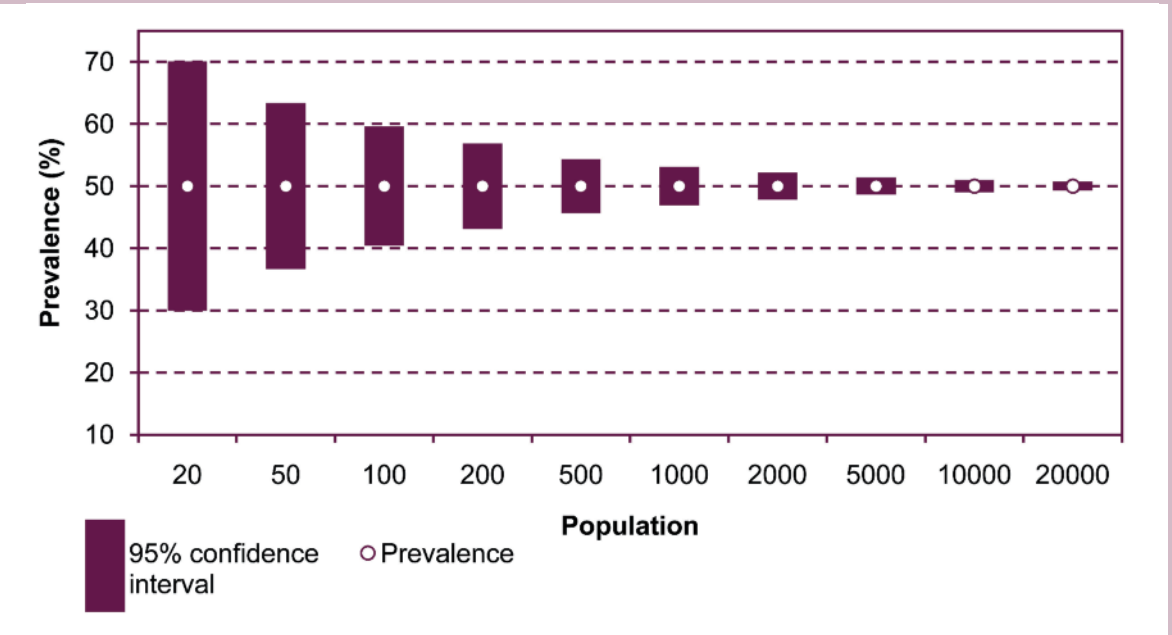
When standardising rates to account for the different age and sex structures of different areas, there are some advantages to using indirect standardisation rather than direct standardisation for areas with smaller populations. First, the numbers of events in each age group, required for direct standardisation, can be too small to permit its use. Second, if the observed and expected numbers of events for the indirect standardisation calculation are provided for small areas, these can each be aggregated to provide the correct numerator and denominator for the aggregated area. By contrast, to calculate directly standardised rates for aggregated areas it is necessary to recalculate the rates from scratch. It must be noted, however, that indirectly standardised ratios only allow comparison between the small area rate and the reference rate: comparisons between different small areas can be misleading.

Confidence intervals and significance testing

The calculation of confidence intervals and significance testing of results are also dealt with in *Technical Briefing 3*[9] and with small numbers of cases in small areas, it is particularly important that underlying variability is highlighted clearly by use of confidence intervals. Significant variations may still be observed in small areas, especially where there is a specific factor, such as the presence of a nursing home or prison.

Figure 1 below indicates the way that confidence intervals increase as the population size is reduced. The larger the confidence interval, the lower the likelihood of achieving statistically significant differences.

Figure 1: Effect of population size on confidence intervals for a prevalence estimate



Analytical solutions

Aggregation

The problems of small numbers and disclosure can both be avoided by aggregating data from small areas into groups, defined in various ways. Data can be aggregated on the basis of geography, time, age group or other characteristics such as socio-economic group or deprivation score. In each case some detail or specificity of the information is lost in order to improve its statistical robustness.

Aggregation over time involves grouping, for example, three or five years together. For visual purposes, a moving average can be graphed to smooth annual variations and give a clearer impression of underlying trends. A disadvantage of grouping data over several years is that the data become less timely. Time series or forecasting methods can be used to estimate more up-to-date underlying rates where observed values fluctuate from year to year. These methods will be covered in a future technical briefing on analysing trends and forecasting.
Geographical aggregation involves grouping smaller areas together and need not necessarily be constrained to adjacent geographical areas (see Box 2).

Box 2: Deprivation categories

Small areas are often aggregated into quintiles or deciles using deprivation scores, e.g. the OAs, SOAs, wards, neighbourhoods or data zones within a LA can be grouped such that the most deprived 20% ('quintile') can be compared with the remainder of the LA population. These groups are geographically defined and can usually be compared using a wide range of routine data, but do not form contiguous areas, i.e. the deprived quintile can be scattered across the LA. While this is useful for analysis of inequalities, it can be difficult to target interventions at these particular areas.

Another approach to geographical aggregation is the use of cluster generation or cluster analysis (not to be confused with the investigation of disease clusters). This aims to define areas that are as homogeneous as possible, and as distinct from each other as possible, based on a wide range of different variables. It is used in geodemographic segmentation[7,14,15] and involves a number of fairly complex statistical processes, which are beyond the scope of this briefing.

Small areas may also be aggregated to approximate service catchment areas, such as schools or GP practices, as mentioned on page 4. There are several approaches to this; the simplest is to allocate small areas to the practice or school with the nearest premises. This may be reasonable for schools, but is rarely useful for GP practices, especially where town centre practices are all based very close to each other, and draw their patients from different areas of the town, for various historical, socio-economic or organisational reasons. Figure 2 illustrates the dispersion in location of residence at LSOA level for patients registered with just one GP practice (the purple areas are those with the highest number of patients; pale green shows the areas with the lowest number of patients).

If data such as patient registers are available to provide a breakdown of the registered practice population by area of residence (e.g. OA) then these areas can either be allocated to the practice with which most residents are registered, or the OAs can be proportionally allocated to the practices with which residents are registered. Shewan[16] provides a description of the different methods for estimating catchment populations and mapping catchment areas, with recommendations for the best methods to use in different circumstances.

Geographical information systems (GIS)

Mapping data using GIS can be a helpful way of presenting small area data. In particular, for abstract areas without names, such as OAs, showing the data on a map can make the data comprehensible – people can see where the data relate to. However, displaying the data on a map does nothing to overcome the small number issues discussed on page 6 – if the data are dominated by random variation, a kaleidoscopic effect is created, obscuring any geographical patterns.

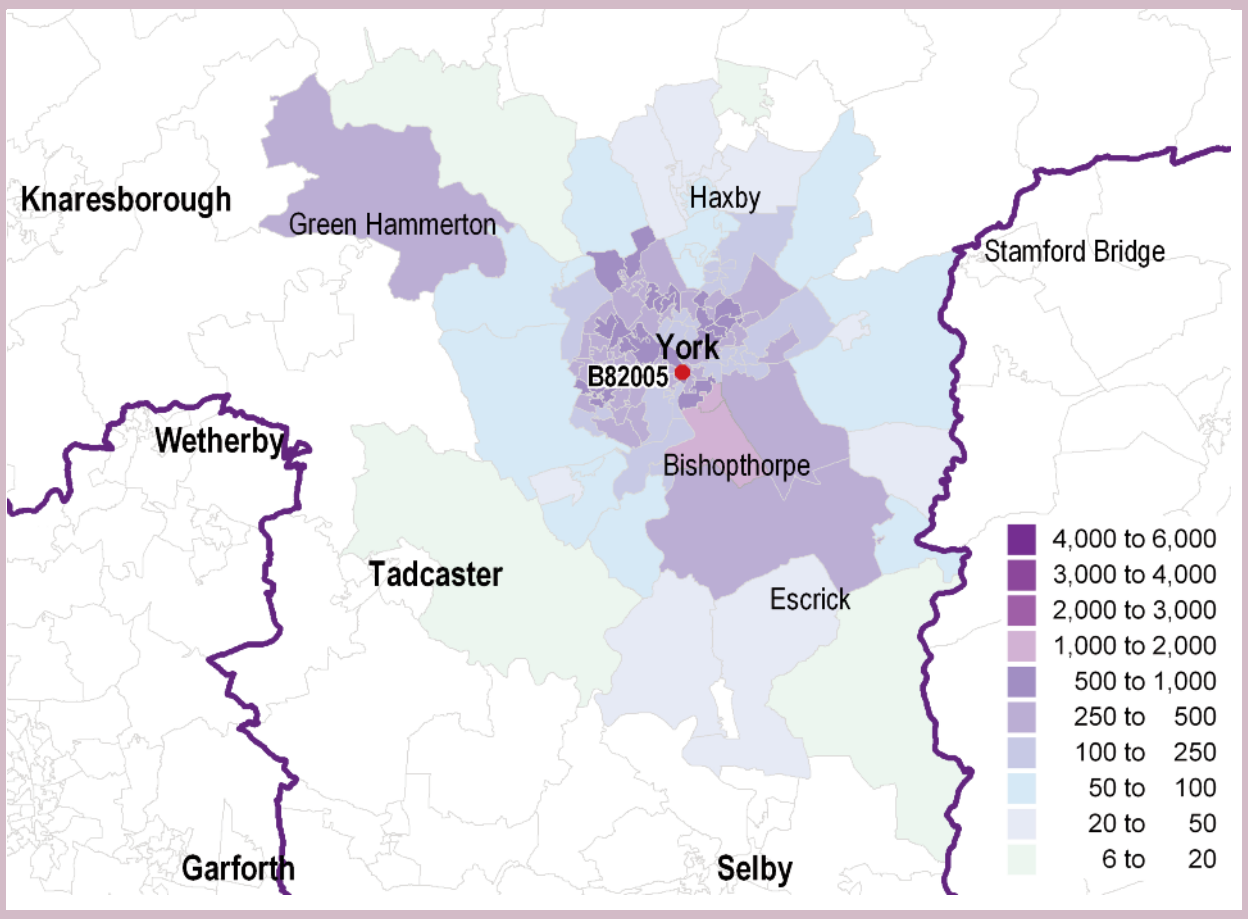
There are many spatial analytical methods which can help to interpret small area data. Detailed description of these methods is beyond the scope of this briefing, but will be covered in a future technical briefing on GIS methods. However, methods that should be considered include spatial smoothing[17,18] (whereby each area's data value is replaced by an average for the area and its immediate neighbours) and Bayesian methods for identifying significant clusters (e.g. WinBUGS).[19]

When displaying data on maps, it should be noted that areas with a low population density tend to be most prominent on such maps because of their size, whereas areas with a high population density are so much smaller that they may be almost invisible. This can be tackled partially by presenting separate maps on different scales for urban and rural areas, or more radically by using cartograms.[20]

Prevalence modelling

To assess needs in local populations, it is necessary to know about the prevalence of disease and demographic, economic, social and behavioural risk factors. Measuring the prevalence of diseases in small areas is usually impractical and, because of the small number issues discussed earlier, unreliable. An alternative approach is to create a model based on national or regional data from robust surveys or analysis of routine data, and apply the model to small area populations (see Box 3). These estimates should be treated as expected values and not true observed figures.

Figure 2: Distribution of patients registered with a GP practice in York by LSOA of residence



Box 3: Example of a disease prevalence model

A model developed to assess the prevalence of chronic obstructive pulmonary disease (COPD) uses the age, sex, smoking status and ethnicity characteristics of an area to estimate the expected prevalence of the condition in that area's population.

The expected prevalence from the model can be compared with the observed prevalence from GP practice registers of patients with COPD, and a standardised morbidity ratio can be calculated.

The results generated by such models are dependent on the robustness of the research evidence used to define the model. Local factors not accounted for by the model could result in inaccurate estimates. However, estimates derived in this way can often be a lot more accurate than incomplete data or observed measurements based on small numbers. Estimates modelled in this way cannot be used to monitor changes over time.

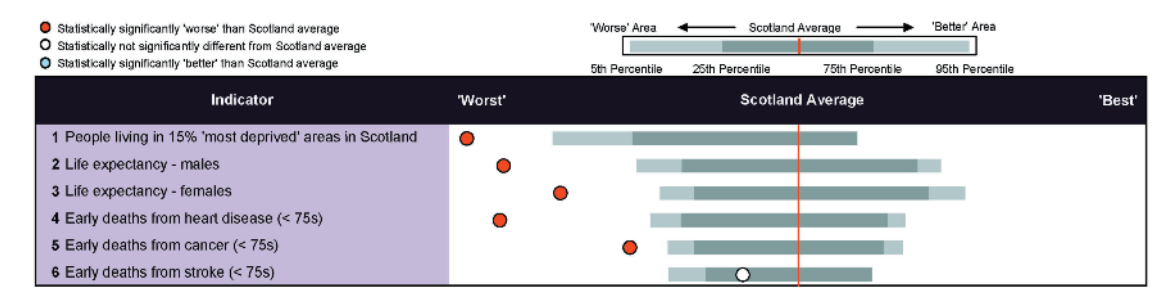
Prevalence models for a range of diseases have been produced and are available on the APHO website.[21] They are populated for England at LA and PCT level but small area populations can be entered by the user. Future technical briefings will cover prevalence modelling methodology and smoking prevalence in particular.

Presenting small area data

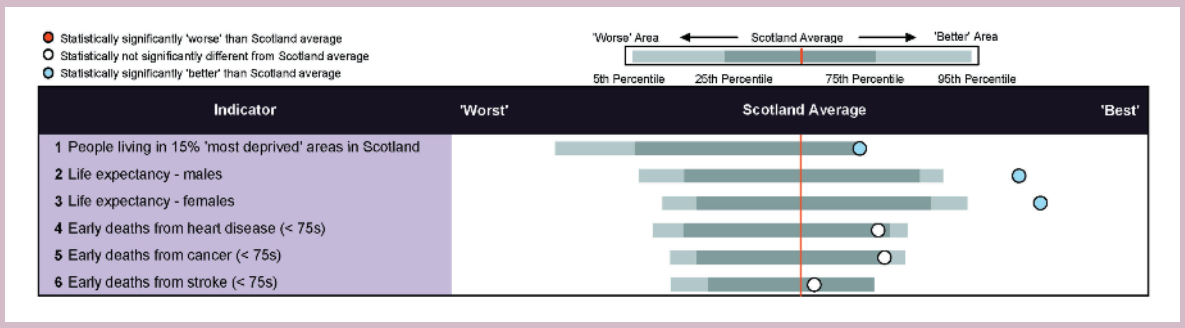
There are many ways of presenting area-based data, either as profiles of an area or comparatively across areas. In general, best practice methodology applies to small area data exactly as for data for larger areas, so display methods such as spine charts (Figure 3) as used in the *APHO Health Profiles*[22] or funnel plots (Figure 4), which clearly show statistical significance of variations should be considered. Funnel plots, which are explained fully in *Technical Briefing 2: Statistical process control methods in public health intelligence*,[23] have the advantage of being able to present data for hundreds of areas on a single graph, highlighting statistical outliers very clearly. As mentioned above, GIS software can be used to present data on maps, which helps people relate the data to their own local knowledge of areas.

Figure 3: Spine chart presentation describing selected health indicators for two contrasting intermediate zones in Scotland

Greendykes and Niddrie Mains

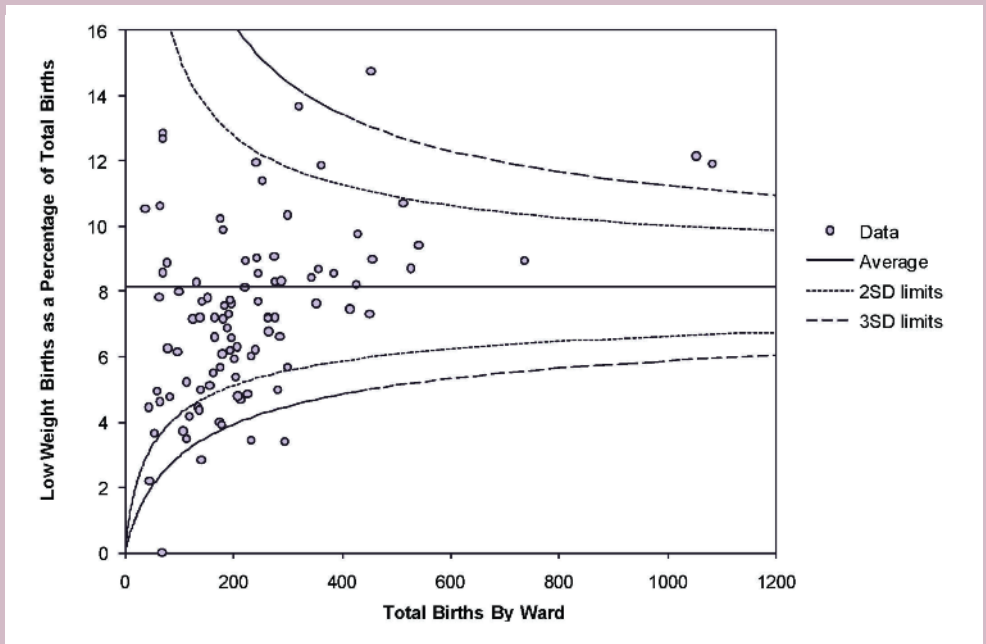


St Andrews South West



Source: Scottish Health and Wellbeing Profiles 2008: http://www.scotpho.org.uk/Profiles

Figure 4: Example funnel plot showing percentage low birth weight births by wards in Luton and Bedfordshire



Source: Births Register 2005

Further resources and links to sources of small area data

ONS Neighbourhood Statistics: http://www.neighbourhood.statistics.gov.uk/

Data for Neighbourhood Renewal (Data4NR): http://www.data4nr.net

Scottish Neighbourhood Statistics: http://www.sns.gov.uk/

Scottish Health and Wellbeing Profiles: http://www.scotpho.org.uk/Profiles

Northern Ireland Neighbourhood Information Service (NINIS): http://www.ninis.nisra.gov.uk/

Northern Ireland Statistics and Research Agency: http://www.nisra.gov.uk

Republic of Ireland Central Statistics Office: http://www.cso.ie/census/SAPs.htm

Republic of Ireland Small Area Health Research Unit (SAHRU): http://www.sahru.tcd.ie/ (work stream includes formulation of the SAHRU deprivation index at small area level based on Census data)

Republic of Ireland Statistics Act, 1993 (covers confidentiality & disclosure): http://www.cso.ie/census/documents/statsact93.pdf

ONS offers useful guidance in using neighbourhood statistics which can be accessed at http://www.neighbourhood.statistics.gov.uk/dissemination/Info.do?page=analysisandguidance/analysis.htm

ONS has also recently published a prototype visual tool which aids users in the analysis of change over time for small areas. The Change over Time Analysis (CoTA) Viewer can be requested from better.info@ons.gov.uk and further information can be found at http://www.neighbourhood.statistics.gov.uk/dissemination/Info.do?page=analysisandguidance/analysisarticles/change-over-time—-guidance.htm

The Department for Communities and Local Government has published Practical Guides for Using Neighbourhood-Level Data, available at http://www.communities.gov.uk/documents/localgovernment/pdf/932145.pdf

All links accessed 27 July 2009.

Glossary and abbreviations

Bayesian inference: Statistical methods in which evidence from observations is used in conjunction with prior knowledge to infer and update the probability that a hypothesis is true. In traditional statistical inference, evidence is used simply to accept or reject a hypothesis depending on whether the probability that it is true is above or below an arbitrary p-value (see below). In geographical analysis, Bayesian methods are used to balance the weights given to national, regional and local observations respectively.

Cartogram: A type of thematic map in which the areas of spatial features are distorted to present them in proportion to the value of an attribute, e.g. population size.

CHP: Community Health Partnership. Local bodies in Scotland responsible for the delivery of health services provided in health centres, clinics, schools and homes. Some are also responsible for delivering local social work services.

Cluster analysis: Methodology for grouping similar units (e.g. neighbourhoods) together into groups or 'clusters', which are both as homogeneous as possible and as different as possible from each other.

Confidence interval: A range of values used to describe the degree of uncertainty around a point estimate of a value, e.g. a prevalence rate. This uncertainty results from random variation in the observed value or estimate. The width of the confidence interval for a defined level of confidence depends on the sample size from which the estimate is derived and the underlying variability in the phenomenon being measured. A 95% confidence interval implies that 95 times out of 100 the interval will include the true underlying rate.

COPD: Chronic obstructive pulmonary disease.

Disclosure by differencing: When data are published for geographical areas with different boundaries (e.g. wards and LSOAs) it is possible that the difference between numbers may allow an individual to be identified and personal information might be disclosed. Disclosure by differencing can also be possible when the values for cells suppressed in a data table may be calculated by using other information presented in the table.

Ecological effects/ecological fallacy: This is the assumption that average statistics for a population can be attributed to every individual within the area in question, ignoring possible variations within the area. It is wrong to assume that factors which explain variation between groups or populations necessarily also operate at the level of individuals.

EDs: Electoral divisions. The smallest level of political geography in the Republic of Ireland. Not to be confused with enumeration districts, previously but no longer used for Census data publication in the UK.

Geodemographic segmentation: The classification of geographically defined populations based on characteristics derived or estimated by combining a variety of data sources.

GIS: Geographical information systems. Software for spatial analysis and mapping.

Homogeneous: Similar to each other, exhibiting little variation within the group.

ISD Scotland: Information Services Division Scotland, the business operating unit of NHS National Services Scotland – formerly known as the Common Services Agency.

LA: Local authority. These are the basic units of administration of local government in England, Wales and Scotland, but local government structures vary between nations and within England. The terms used also vary, including district, borough or city councils and unitary authorities.

LGD: Local government district. The basic unit of administration of local government in Northern Ireland.

List inflation: The difference between the number of registered patients listed on a GP register and the number on the register who are alive and appropriately registered with that practice. List inflation tends to be more of a problem in areas of high migration, where delays occur in deleting patients from practice lists.

NISRA: Northern Ireland Statistics and Research Agency.

OA: Census output area. The smallest area for which 2001 Census results are released. OAs are based on postcodes and fit within 2003 ward boundaries.

ONS: Office for National Statistics.

PCT: Primary Care Trust. Local bodies responsible for public health and provision and commissioning of health services in England.

Prevalence: Prevalence is a statistical concept defined as the number of cases of a disease or characteristic that are present in a particular population at a given time.

p-value: The statistical probability of obtaining by chance a result at least as extreme as the one observed. If the p-value is low (e.g. 0.05 or 0.01), this indicates that a result as extreme as the one observed would occur by chance in only a small proportion (1 in 20 or 1 in 100) of all possible samples. On this basis, the conclusion is usually drawn that it probably didn't occur by chance, but resulted from some specific cause or causes.

QOF: Quality and Outcomes Framework, the mechanism for rewarding general practitioners in the UK for meeting a defined set of quality criteria.

Random variation: Variability of a process (which is operating within its natural limits) caused by many irregular and erratic (and individually unimportant) fluctuations or chance factors that (in practical terms) cannot be anticipated, detected, identified or eliminated. [Source: BusinessDictionary.com] Even when areas have the same underlying risk of mortality, say, the actual number of deaths will vary from year to year and from area to area, without implying any difference in the underlying risk of death for individuals in the areas.

SOA: Super output area. Statistical areas built up from groups of OAs. There are two types: lower super output areas (LSOAs) with an average population of 1,500 and middle super output areas (MSOAs) with an average population of 7,200.

Spatial smoothing: Spatial smoothing is a statistical technique applied to try and show the underlying trends across areas, by combining data for each area with data from its neighbours. This can reveal patterns in the data which are otherwise hidden behind large random variations between areas.

Sure Start: A UK Government initiative launched in 1998 with the aim of 'giving children the best possible start in life' through improvement of childcare, early education, health and family support.

Time series: Statistical methods for interpreting trends in data over time in order to forecast future values.

References

1. ONS. Beginner's Guide to UK Geography. Available at http://www.statistics.gov.uk/geography/beginners_guide.asp
2. Department of Communities and Local Government. Indices of Deprivation 2007. Available at http://www.communities.gov.uk/communities/neighbourhoodrenewal/deprivation/deprivation07/
3. The Scottish Government. Scottish Neighbourhood Statistics. Available at http://www.sns.gov.uk/
4. ONS. Statistical Wards, CAS Wards and ST Wards. Available at http://www.statistics.gov.uk/geography/statistical_cas_st_wards.asp
5. Cabinet Office. UK Government Data Standards Catalogue. Available at http://www.govtalk.gov.uk/gdsc/html/frames/Postcode.htm
6. NHS. NHS Postcode Directory. Available at http://www.datadictionary.nhs.uk/web_site_content/supporting_information/nhs_postcode_directory.asp?shownav=1
7. Abbas J et al. Technical Briefing 5: Geodemographic Segmentation. York: APHO; 2009. Available at http://www.apho.org.uk/resource/item.aspx?RID=67914
8. Department of Health. Quality and Outcomes Framework. Available at http://www.dh.gov.uk/en/Healthcare/Primarycare/Primarycarecontracting/QOF/DH_099079
9. Eayres D. Technical Briefing 3: Commonly used public health statistics and their confidence intervals. York: APHO; 2008. Available at http://www.apho.org.uk/resource/item.aspx?RID=48457
10. ONS. 2001 Census Disclosure Control. London: ONS; 2001. Available at http://www.statistics.gov.uk/about_ns/downloads/info_to_commission/AG(01)06_Disclosure_Control.doc
11. OPSI. Freedom of Information Act 2000. Available at http://www.opsi.gov.uk/acts/acts2000/ukpga_20000036_en_1
12. ONS. Review of the Dissemination of Health Statistics: Confidentiality Guidance (Working Paper 4: Glossary). Available at http://www.statistics.gov.uk/about/Consultations/downloads/Health_Stats/Health_Stats_4_Glossary.pdf
13. ISD Scotland. ISD Statistical Disclosure Control Protocol. 2009. Available at http://www.isdscotland.org/isd/4489.html#smallNumbers
14. Arrundale J et al. Handbook and Guide to the Investigation of Clusters of Diseases. London: Leukaemia Research Fund; 1997.
15. Vickers D. Introducing Clustering, Area Classification and Geodemographics. In: Multi-Level Integrated Classifications: Based on the 2001 Census. PhD thesis. Leeds: Department of Geography, University of Leeds; 2006.
16. Shewan J. Catchment areas and populations. Cambridge: ERPHO; 2003. Available at http://www.erpho.org.uk/viewResource.aspx?id=14754
17. Holmes N. Spatial Smoothing. BURISA 2006;No.167:9-13. Available at http://www.burisa.org/Temp/167.pdf
18. Baker A, Ralphs M, Griffiths C. Standardised Mortality Ratios – the effect of smoothing ward-level results. Health Statistics Quarterly 2008;48.
19. Lunn DJ et al. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 2000;10:325-337.
20. Krygier JK, Wood D. Making Maps: A Visual Guide to Map Design for GIS. New York: Guildford Press; 2005.
21. APHO. Prevalence Modelling. Available at http://www.apho.org.uk/resource/view.aspx?RID=48308
22. APHO. Health Profiles. Department of Health; 2009. Available at http://www.healthprofiles.info/
23. Flowers J. Technical Briefing 2: Statistical process control methods in public health intelligence. York: APHO; 2008. Available at http://www.apho.org.uk/resource/item.aspx?RID=39445

All links accessed 27 July 2009.

About the Association of Public Health Observatories (APHO)

The Association of Public Health Observatories (APHO) represents and co-ordinates a network of 12 public health observatories (PHOs) working across the five nations of England, Scotland, Wales, Northern Ireland and the Republic of Ireland.

APHO facilitates joint working across the PHOs to produce information, data and intelligence on people's health and health care for practitioners, policy makers and the public.

APHO is the largest concentration of public health intelligence expertise in the UK and Republic of Ireland, with over 150 public health intelligence professionals.

APHO helps commissioners to ensure that they get the information they need and our websites provide a regular stream of products and tools, training and technical support.

We work with partners to improve the quality and accessibility of the data and intelligence available to decision makers.

We are constantly developing and learning new and better ways of analysing health intelligence data. We use these new methods to improve the quality of our own work, and share them with others.

Updates and more material, including methods and tools to support our technical briefing series are available through our website at http://www.apho.org.uk

For further information contact:
Association of Public Health Observatories
Innovation Centre, York Science Park,
Heslington, York YO10 5DG

Telephone: 01904 567658
http://www.apho.org.uk