ONS Business Data

This is a best practice example of data dissemination by the Office for National Statistics. It highlights the importance of consistency in application of Anonymisation to multiple output formats all originating from the same data source.

When reading this case study it is important to remember that the anonymisation assessment, recommendations and actions are specific to the data examined here although they will have relevance to other similar data and data context.

- **Summary Information**

Organisation disseminating data: Office for National Statistics.
Data: confidential record-level Inter-Departmental Business Register (IDBR).
Data type: Public dataset presented as tables, maps, percentages.
Summary of sharing, dissemination, publication practices: Disseminated online.
Risk type: Identification disclosure of businesses and their confidential returns. Mosaic effect between data outputs/graphics, previous publications, and potential ad hoc customer queries.
Features of risk: Dominating contributions and small cell values in certain sectors. Unusual patterns across time.
Disclosure Control Methods (SDC): Suppression.
Disclosure risk checks: whether any suppressed or rounded value can be 'unpicked'.

- **Background**

Local planners and policy makers need information about businesses. The kinds of questions they ask are: How many businesses are there in my area? What sector of the economy do they belong to? Are they large or small? How many people do they provide employment for? How has the picture changed over time?

The IDBR is a statistical register based on administrative sources and surveys which contains information on businesses in all parts of the economy. It covers nearly all of UK economic activity including that of public sector bodies. The IDBR database is organised on two levels: one dataset contains enterprise-level information while the other has information about 'local units' (the physical locations or workplaces which make up the enterprise).

The two interests in these data were (1) the differences by industrial sector and (2) detail down to low geography, both to provide a time series as far back as possible (to 2001). In the Local Authority tables, the results presented show both numbers of businesses and numbers of employees in each industrial sector in each of the 33 Local Authorities in London split according to whether they work for micro enterprises (employing fewer than 9 employees in the UK), small and medium sized enterprises (employing between 10 and 249 employees in the UK) and large enterprises (employing 250 or more). In the MSOA tables, the data goes down to the small area level presenting analysis of nearly one thousand small areas within the 33 London Local Authorities showing the number of employees working for businesses with fewer than 250 employees in the UK (small- and medium-sized enterprises, or SMEs) and for businesses with 250 or more employees in the UK (large enterprises).

- **The Anonymisation Problem**

The demand for results for small geographical areas or for each industrial sector proved challenging because of the requirement to protect the confidentiality of the IDBR data. This means making sure that it is impossible for users to identify a particular business from the published results. This is more likely to happen when analysing results in small geographical areas than when looking at London or the UK as a whole. In areas where there are few firms or a 'dominant' firm, if results were to be published it might be possible for users to infer how many people were employed by a particular business. Although numbers of people employed is not generally seen as commercially sensitive, it is a legal requirement for ONS to protect this information because it has been provided in confidence by the businesses concerned. This is required under the Statistics of Trade Act 1947.

There are further disclosure complications with this dataset including the disclosure risk from producing outputs illustrating change over time, and the fact that a variety of outputs are published from the same data source.

- **Anonymisation Practices**

Tabular Outputs

In order to publish tables, the method of suppression used was based on standard disclosure rules. Suppression involves blanking out cells if they are considered disclosive. Cells in these tables were suppressed if:
1. There were fewer than 5 local units (businesses) in any cell of a table (primary suppression).
2. One local unit is dominant in any cell making it possible for a reader to work out that local unit's contribution (primary suppression). In most cases, specific sectors in specific locations had low numbers of large businesses, which meant one local unit could easily make up 90% of the total number of employees and thereby dominate.
3. Results in other parts of a table or set of tables allow readers to work out the results that have been removed during primary suppression.

An additional consideration was the fact that the data showed change over time which could aid identification. For example in one area of London, a new business appeared in one year that was the sole business in that category, so if the number of employees in large businesses remained the same in every year it would be possible to deduce the employees in that one business.

Suppression was therefore carried out across geographies, sectors, size of enterprise and time, for example making sure that a partner set of cells in another sector of that authority/year is also suppressed. Suppression in this case was done manually, to ensure usability of the final data based upon knowledge of the London economy and geography, but there are also packages available such as Tau-Argus which could do this automatically.

Percentages

Alongside some tables, percentages of employees by industry type were released which on further investigation allowed unpicking. These percentages use the unrounded figures to present outputs in a different way, but which directly relate to the core tables being released.

Tables 1 and 2 show how the percentage of employees in SME enterprises (table 2) could be used to unpick the suppressed value (marked '..') in table 1 to reveal the employees in medium-sized enterprises.

Table 1: Cutting from Number of Employees by Enterprise size

Industrial sector	Enterprise size	2001
Primary and utilities	0-9 employees	0
Primary and utilities	10-249 employees	..
Primary and utilities	250 or more employees	..
Primary and utilities	Total in all enterprises	394

Table 2: Cutting from Percentage of Employees in Small/Medium Enterprises

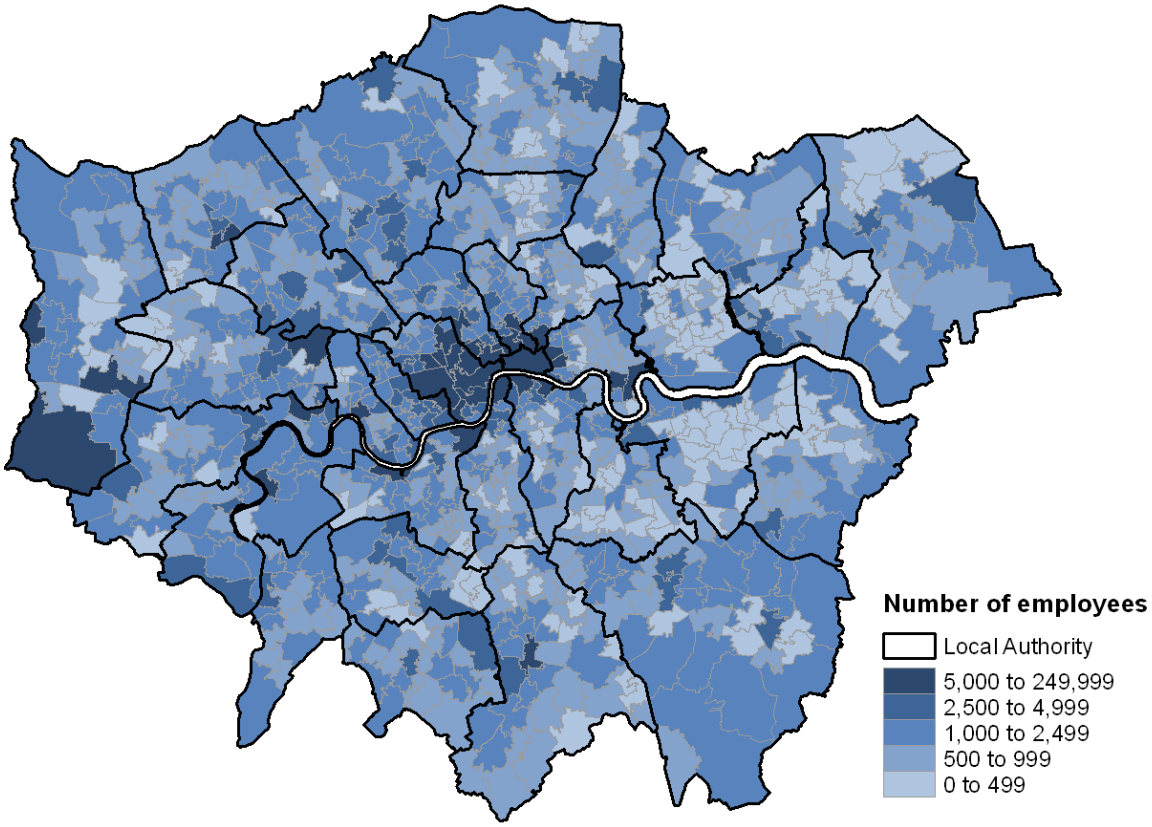
Industrial sector	Enterprise size	2001
Primary and utilities	0-249 employees	19.8

To prevent this kind of disclosure, rounding was also applied to the percentages. A test was done to check values were within bounds to ensure no further suppressions were required.

Maps by London boroughs

Maps were published showing summarised information of the data using bandings of number of employees. It was important to check that these bandings were sufficiently broad so not to unpick the suppressions in the tables to within 10-20% of their original value, but still sufficiently tight to enable utility.

Number of employees working for SMEs in London, 2012



Source: Inter-Departmental Business Register, Office for National Statistics

Maps by workplace zones

Further to the original publication, customers requested extractions of data at very small geographies, specifically the Workplace Zones created for the 2011 Census. These maps indicate the number of employees working in specific sectors in each workplace zone, but there were new disclosure considerations in addition to those mentioned above, including:

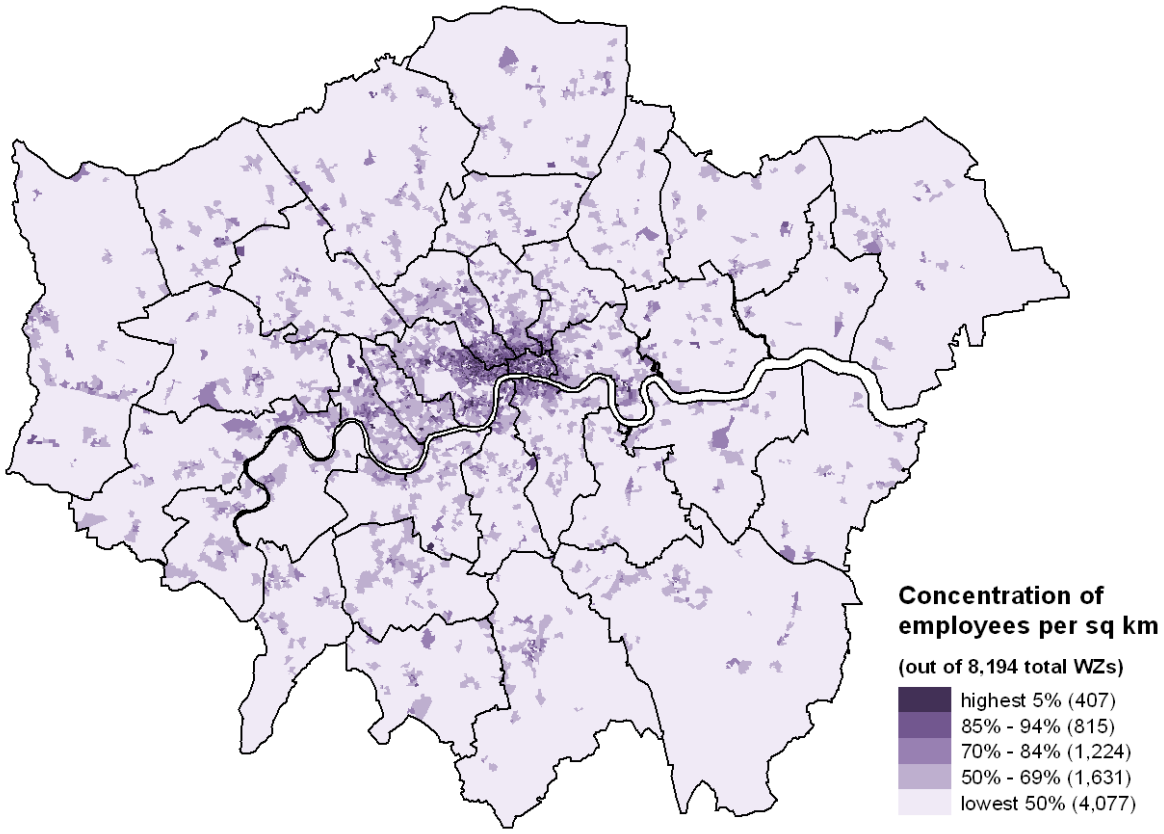
- was it possible to compare these results with previously published results and identify small cell values by "differencing"?
- was it possible to calculate the number of employees in an area given that the size of these areas is available in the public domain?
- is banding the number of employees sufficient, or are those bands too tight?

If it were only the map being published, it may be possible to be less restrictive as it would not be possible to pinpoint exactly what area was of a certain colour and therefore not able to identify exactly where numbers of employees were. However

the map did have to be accompanied by a table also containing the figures and thereby making it possible to identify specific areas.

This information was quite detailed so the disclosure control approach taken was to display the concentration of employment, rather than the exact number of employees, in percentages and again to band the numbers.

London Workplace Zones ranked by number of employees working in 'Science and Technology' per sq km, 2012



Source: Inter-Departmental Business Register, Office for National Statistics

- **Data sharing, Dissemination, Publication**

The data are published online and disseminated in two parts:

The first part about London and its boroughs: www.ons.gov.uk/ons/rel/regional-trends/london-analysis/size-of-firms-in-london--2001-to-2012/index.html and the second part was about the MSOAs and maps: www.ons.gov.uk/ons/rel/regional-trends/london-analysis/small-and-large-firms-in-london--2001-to-2012/index.html

Further ad hoc data requests have also been published based upon this release.

For further information please contact: Elaine Mackey at admin@ukanon.net