**Local authority procurement data**

There are 24 files in the “Chunked\_LA\_Data\_04112017.zip” archive. Each is in the following format:

|  |  |
| --- | --- |
| Variable | Description |
| Amount | Value of transaction |
| Date | Date of transaction – usually DD/MM/YYYY |
| Normalized\_beneficiary | Recipient of transaction; name standardised to assist in matching (e.g. consistent recording of terms, such as “limited” rather than “Ltd” for a company |
| Provider | Provider of funds – named local authority |
| Raw beneficiary  | Name of recipient of funds as it originally appeared in source |
| Suggested\_charity | Best match with list of registered charities |
| Suggested\_company\_name | Best match with list of companies from Companies House |
| Suggested\_education\_name | Name of education provider (from Department of Education lists) |
| Suggested-health\_name | Name of health provider (NHS Trust) |
| Suggested\_namedindividual | Best match is an individual name |
| Suggested\_public\_name | Suggested name of public authority, e.g. another local authority |
| Suggested\_sports\_name | Suggested match against list of community sports clubs (CASCs) |

These were captured by a combination of web scraping and manual searches and downloaded as found on local authority websites; as far as possible, the information was then rendered into a consistent format but we were at the mercy of individual local authorities whose practices varied greatly.

We do not recommend use of this data as we have provided a summary by funder (local authority) recipient and financial year – and we believe that this is the format in which users will find the data most useful. The format of that data is now described.

***Summary data on local authority procurement (file: LAprocur.dta)***

Each row is a summary of the total amount paid to a given provider of service (or recipient) in a given financial year. This reduces the scale of the data to c. 1.8 Mn observations, representing 560 000 distinct organisations.

To aid analysis we have also attached ONS codes for local authorities, rather than their names; and we have also included charity commission or company numbers were

|  |  |
| --- | --- |
| Variable | Description |
| Funder | Local authority – identified by ONS code, e.g. “E08000025” (Birmingham) |
| Recipient  | Name of organisation receiving funds |
| Recipient\_charity\_number | Charity Commission registration number, if matched (blank otherwise) |
| Recipient\_company\_number  | Companies House registration number, if matched (blank otherwise) |
| Recipient\_organisation\_type | Type of organisation, based on most plausible match: types include Charity, company, educational establishment, health provider (various: GP surgeries, pharmacies, but a range of other providers also included), individual (this is when our matching procedures produce no match and the text suggests a named individual), public sector (principally named public authorities, e.g. parish councils, district or county councils) and sports organisations (from lists of Community Amateur Sports Clubs) |
| Year | Calendar year of payment |
| Amount | Value of transaction |

Classification of organisation type is inherently challenging and approximately one quarter of organisations were unclassifiable given the time and resources we had available. However they only account for some 5% of the total value of transactions.

These data were compiled from information that is **publicly available** – all we have done is put it together in a relatively accessible form. We assume that anyone using it has as a primary purpose the identification of broad patterns in the data rather than the identification of named individuals, of whom there are many in this dataset.

Are there significant risks to individuals from this data? Firstly, all that is available is the name of a person. There is no address information given and we do not know, therefore, whether someone who receives money from a given local authority actually resides in that area. The probability of being able to identify someone uniquely is therefore low. To consider the risk involved, we undertook some further investigations of the data.

Our data contain information on the following entities:

Charity 23640 4.2 4.22

Company 255199 45.5 49.7

Education 6023 1.1 50.8

Health 7633 1.4 52.2

Individual 64275 11.4 63.4

Public-Sector 5861 1.1 64.8

Sports 708 0.1 64.9

Unallocated 197221 35.1 100.00

The least problematic categories are charities and companies, where we are very confident that we have successfully matched names recorded in the procurement data. Over 98% of organisations classed as charities have a charity registration number; almost all those organisations identified as companies have a Companies House number, or their names contain the string “limited” or “Ltd”, denoting company status.

Investigation of a sample of records shows that the matches against lists of education providers, and health providers, almost wholly identify organisations, such as NHS Trusts, hospitals, GP Practices, or schools; the same is true also for the “Public sector” category, where we were matching lists from the procurement data against lists of known public authorities, and the “sports” category (lists of sports clubs).

This leaves a large “unallocated” category, accounting for 35% of observations (197000 entities). From an investigation of a sample of these, it is clear that well over 95% of records contain the names of companies, charities, or other organisations that have not been successfully matched to lists of such entities. This is almost invariably because the way in which their name is recorded differs in non-trivial ways from the way it is captured on the list of a regulator.

It also leaves a list of some 64 000 individuals, approximately 11% of the total, but accounting for less than 5% of total expenditure recorded. At least 15% of these appear to be organisations, companies or charities, but the majority of cases contain what are recognisable as the names of people. We suggest these are dropped. Many are very small payments indeed. One way for analysts to proceed with this would be to recode the data into size bands for payments of £0 – 100, £100-999, etc., and dropping anything categorised as a payment to an individual as well as any payments below a certain threshold. For example, dropping individuals, and setting the payment threshold at £1000, would remove half a million transactions (leaving a total of 1346251) but only 4.3% of the value of those transactions. Most of the small amounts paid to individuals are likely to be reimbursements of expenses, although there is no way of judging that from the source material.

An alternative is to drop entries on purely financial grounds; for example, a threshold for payments of at least £100k in any one year would reduce the dataset to 153000 observations, of which only 2.3% would be in the “individual” category.