

# The Keys to Unlocking Public Payments Data

Charles Rahal

Department of Sociology and Nuffield College  
University of Oxford

Accepted for publication at Kyklos, 25th of August 2017

## Abstract

We mechanize some of the richest yet significantly under-utilized data resources within developed, ‘Open Data’ economies. We show how it is possible to scrape, parse, clean and merge tens of thousands of disaggregated public payments datasets in an attempt to bridge the methodological gap between newly available data from the administrative sphere and applications in empirical social science research. We outline techniques to unambiguously link records to various freely available institutional registers. In particular, we offer guidance on overcoming the substantial challenges of heterogeneous provision and administrative recording errors in the absence of Uniform Resource Identifiers, namely in the form of an approximate, domain-specific ‘record-linkage’ type matching algorithm. As an illuminating example, we construct a cleaned database of 24,581,192 local government payments subject to the Local Transparency Codes which total £169.87bn in value. We overcome various challenges in a detailed examination the procurement of services by local government from the voluntary sector: an important contemporary issue due to the rise of the ‘Big Society’ political ideology of the early 21st century. Finally, we motivate future work in this area and discuss potential international applications and practical advancements.

**Keywords:** Public Administration, Public Economics, Local Governance, Open Data

**JEL Classification:** C81; H50; L30

---

For correspondence: Dr. Charles Rahal, Department of Sociology, University of Oxford, Manor Road Building, Manor Road, Oxford, OX1 3UQ. E-mail: [charles.rahall@sociology.ox.ac.uk](mailto:charles.rahall@sociology.ox.ac.uk). The code libraries for this project – written in Python 3.5.2 – are available upon request and at [github.com/crahal](https://github.com/crahal), and the raw database used in the empirical example is available via the UK Data Service. The author would like to thank John Mohan for his stewardship of the work. Thanks are also due to David Kane for sharing insightful Python code, and to Marc Lawson and Cin Man Winnie Yeung for research assistance. Peter Backus, Neil Owen, an anonymous referee and journal editors provided a substantial number of constructive comments. Financial support gratefully received for this project by the Economic and Social Research Council (grant number ES/M010392/1). Useful feedback was received from participants at the Charity Data Dive (Manchester) and the ESRC Data Partnerships Event (London).

“We [economists] have been using the same datasets over and over again, and since we wanted new answers, we have been developing new econometric techniques to try to transform the data, and get more meaningful information out of them. Just having a fresh, new data set brings a whole new perspective, and I think people are starting to realize that, and gradually people are becoming more interested in data collection itself”

---

– A. Cavallo, 15 November 2012 in [Taylor et al. \(2014\)](#)

## 1 Introduction

The emergence of ‘Big Data’ has strong parallels to the Industrial Revolution of the 19th century, bringing with it the potential to transform the way in which we approach all sub-disciplines of social science. The way in which governments make administrative data available to citizens and researchers is of particular significance, representing a potential goldmine for policy analysts. This is especially true with regard to public procurement data which forms a fundamental part of modern economies, and typically accounts for tens of percents of gross domestic product. Historically, economists have dealt only with data that fits in a spreadsheet, but that is changing rapidly. [Einav and Levin \(2013\)](#) provide several extremely relevant examples as merging large-scale administrative records becomes an increasingly critical component of cutting edge empirical research in many social science sub-disciplines. Governments collect huge amounts of data which can be used for guiding policy decisions, but this data is, as of now, significantly under-utilized. The ‘Open Data’ revolution creates in its wake a tremendous resource that is as yet largely untapped, and this is truer nowhere more so than for open government data. The value of being able to operationalize ‘Open Data’ is huge, and several studies estimate the economic benefits in the regions of tens to hundreds of billions of US dollars (and the Open Data Institute compiles a comprehensive list of estimates of value-added from various sources). Within the academic record-linkage literature, [Koudas et al. \(2006\)](#) cite figures which claim that the economic loss of poor administrative data is as large as \$611bn per year. Because this data is both immensely detailed and incredibly meaningful, profound conclusions can often be drawn from simply describing basic patterns. However, distinct epistemological, methodological and computational challenges remain before the full potential can be realized. Despite the recent push for greater transparency – data ‘dumping’ by public bodies alone is not in itself generating any significant value.

Government transparency has become a major objective within the public administration sphere, but in the realm of public finance, the value of granular spending data cannot be realized without

tractable methods for tracking and mapping the payments. There are numerous daemons which have prevented this type of data from being analyzed on a national, or even international scale, despite global motions for change. These originate from concerns of ‘raw’ input fidelity – the focus of our dataset construction diverges from traditional ‘user-generated content’ which is prevalent in much existing research. While the data is ‘open’, it is often extremely difficult to aggregate into a suitable format for further analysis due to a lack of standardization in supplier names. The only way of mapping this data is through this supplier name (a string literal) typed by a capricious civil-servant.

Existing ‘Big Data’ research across social science utilizes recent technological advancements such as distributed computing, and is typically split into one of three themes: administrative, social, or private sector, and exciting developments are emanating from all three. The general approach considered within this paper falls within the ‘administrative’ sphere, with an overlap into the ‘private sector’, given the general direction of local government expenditure. One of the prohibiting features of some ‘Big Data’ research is that of its proprietary nature. Our contribution is to outline a methodology (in the form of algorithms and general analytical tools), which makes such deep – almost intractable ‘nano’ level data readily accessible (and to show what it is capable of). This is despite the extremely heterogeneous, noisy raw format in which it is made available at source.

In this paper we outline methods for unlocking access to a unique source of data which hitherto remains unexplored: perfectly granular, high velocity public finance transactions data. We build tools to create a national database of localized government spending data which is similar in structure to national spending aggregates (such as OSCAR in the U.K., or NIPA in the U.S.) and we develop algorithms to map every payment to a terminal recipient across various registers, despite the lack of a Uniform Resource Identifier (URI). Our approaches can be applied to a range of international contexts. As an example application which showcases the value of our methodology, we provide a thorough examination of public procurement from the third sector by analyzing 24,581,192 local government payments subject to the Local Transparency Codes in England totaling £169.87bn. New data-sources such as this have never been more important given the need to evaluate political ideologies such as the ‘Big Society’ (announced by David Cameron in July 2010) which was intended to empower local people through a civic sense of community. Founded upon a post-Thatcherite brand of ‘one-nation’ conservatism, the concept relied heavily on the ideas of localism, devolution and volunteering during a time of austere fiscal policy and subsequent funding cuts to local public services by central government. Our approach herein represents an ideal way to analyze the financing of this broad policy change.

The remainder of the paper is structured as follows. We first review the technical work in this area, followed by the local government spending and voluntary sector financing literature. We

then give an overview of the appropriate policy sphere relevant to our methods and application and provide a detailed description of the two primary sources of data which we utilize: local authority (hereafter LA) procurement data and open source institutional registers. Within the methodology section we describe (and then evaluate) our approach to reconciliation in detail, including discussions and examples of name ‘normalization’ and our approach to ‘targeted’ approximate string matching. We showcase what our algorithmic treatment of local government procurement makes possible with a specific focus on procurement from the third sector. We then describe a number of further potential applications which utilize other types of procurement (such as from the private or health-care sectors) as well as international opportunities for further research before finally concluding.

## 2 Literature Review

### 2.1 Technical Literature

The most similar method conceptually to ours is the Company, ORganization and Firm name Unifier (CORFU) approach of [Alvarez-Rodríguez \*et al.\* \(2015\)](#) which is validated against the procurement dataset of supplier names in Australia between 2004-2012, containing 77,526 unique names in 430,188 payments. Further comparisons to our approach are drawn later on, as, despite their ultimate objective being different, they are still interested in the issue of unifying ‘n string literals  $\rightarrow$  1 company  $\rightarrow$  1 URI’. In another similar piece of work which formed part of the ‘LOD2 – Creating Knowledge out of Interlinked Data’ project, [Svátek \*et al.\* \(2014\)](#) outlines the different, yet interrelated tasks of data extraction, publishing (presented as a ‘Public Contracts Ontology’), buyer/supplier matchmaking and aggregated analytics at both a national and EU-level. The MOLDEAS project – which developed a linked pan-European e-procurement platform (outlined in [Alvarez \*et al.\*, 2012](#)) – addressed the matching task using techniques such as spreading activation and resource description framework (RDF) classification. However, in comparison to most of the existing work, our methodological approach is constrained by the fact that we are dealing with a simple string literal to identify and link payments. [OpenCorporates \(2017\)](#) also provide a brief explanation of the scoring system utilized by the OpenCorporates API – specifically related to how they weight their heavily normalized company names with supplementary information on inactivity and jurisdiction.

Moving away from the specific applied task of company name reconciliation, there exists a substantial body of work which outlines the general theory of ‘record linkage’. The primary applications in this line of research involve the creation, updating, and un-duplicating of survey data (using non-unique identifiers, such as names, dates of birth, addresses), and serves as a

method to link individuals via names and addresses from multiple administrative files.<sup>1</sup> The earliest contributions to modern record linkage date back to [Newcombe \*et al.\* \(1959\)](#) and then [Fellegi and Sunter \(1969\)](#) who provide a more formal definition of the problem. Since then, however, a number of approaches have been developed which rely heavily on data mining and machine learning. One especially relevant piece of work is [Enamorado \*et al.\* \(2017\)](#),<sup>2</sup> which develops ‘a fast and scalable algorithm to implement the canonical probabilistic model of record linkage’ able to ‘efficiently handle millions of observations while accounting for missing data and measurement error’. While the scope of this literature is too broad to discuss in detail here, we refer where appropriate to the applicable work in this field at the relevant points in our discussion throughout the remainder of this paper. Our algorithmic approach discussed below (‘targeted approximate matches’) can be seen as one ‘domain specific’ way of achieving this (in a similar way to [Jin \*et al.\* \(2003\)](#) which ‘preserves domain-specific similarity’).

## 2.2 Relevant Local Government Spending Literature

Our work is motivated not only from a desire to utilize this new source of data, but also because the lack of local expenditure studies with an empirical focus is striking.<sup>3</sup> From a theoretical perspective, the study of local expenditures can be traced back to what remains the most influential work in the field – [Tiebout \(1956\)](#), which develops a model which ‘yields a solution for the level of expenditures for local public goods which reflects the preferences of the population more adequately than they can be reflected at the national level’. This was based on the idea that local governments attract perfectly mobile, fully informed citizens through provision of public good packages until they reach an optimum community size. [Yinger \(1982\)](#) expands the Tiebout concept by fully accounting for the capitalization of local fiscal policies into house values. However, barriers to citizen mobility, a lack of co-ordination, and a limited choice of municipalities typically leads to failure of the mechanism as envisaged by Tiebout. [Besley and Coate \(2003\)](#) outline a political economy approach which considers the trade-off between centralized and decentralized provision of local public goods.

[Oates \(1969\)](#) began a discussion on the efficiency of local public good supply, observing that property values were positively related to the amount of public spending, and negatively to local taxation – providing some support for the Tiebout hypothesis that citizens are mobile and migrate. The underlying idea here is that house prices reflect discounted rents plus the imputed net value of services less taxes. Another line of inquiry follows the test of allocative efficiency of [Brueckner](#)

---

<sup>1</sup>[Jaro \(1995\)](#), for example, is a well cited example which is primarily concerned with the probabilistic linkage of large public health data files, and other applications involve population census’s or death indexes.

<sup>2</sup>This is accompanied by the ‘fastLink: Fast Probabilistic Record Linkage with Missing Data’ package in R.

<sup>3</sup>For a thorough review of the local public finance literature, the reader is directed to [Blankart and Borck \(2005\)](#).

(1982), which showed that property values are an inverted U-shaped function of public service provision.

One relevant empirical study is that of Solé-Ollé (2006), which takes a cross section of 2,610 Spanish municipalities to examine two types of LA expenditure spillovers. At a national spending level, a vast body of work investigates the relationship between government spending and economic growth (such as Barro, 1991), with various disputed conclusions. Fewer papers consider the relationship at a ‘sub-federal’ level, such as Schaltegger and Torgler (2004), which examines Cantons in Switzerland between 1981-2001. Another influential study is that of Gyourko and Tracy (1991), which uses an inter-city random effects model to show how local fiscal climates affect the quality of life across metropolitan areas.

## 2.3 Financing the Third Sector

Problems in analyzing the funding base of the third sector abound – both in the U.K. and abroad. Administrative data (gathered by regulators such as the Charity Commission in England and Wales) provides only a partial picture of key financial characteristics of the aggregate sector. This is partly due to financial thresholds (such as reporting requirements only being binding for incomes above £500,000), and the fact that reporting categories are ambiguous (as discussed in Morgan, 2012). The initial analyses of Posnett (1990), Osborne and Hems (1995) and others has been subsequently advanced by the National Council for Voluntary Organisations (NCVO hereafter) in their annual publication, the ‘Civil Society Almanac’. This regular work generates estimates of the total amount of voluntary sector income that comes from different sources (see NCVO (2016) for further details). The legwork involved in the analysis of third sector financing is emphasized by Clifford and Mohan (2016). In this paper, some 500,000 lines of data were captured manually as accounts data is not readily available in machine readable format. Making use of a custom web-based form which allowed the data entry staff to enter data in a way that replicated the hierarchy of data in the accounts, they managed to capture information on a sample of 7,000 charities in England and Wales in order to describe the distribution of charities across income – highlighting the diversity of organizations with charitable status.

Utilizing another type of dataset altogether, Clifford *et al.* (2010) use data from the National Survey of Third Sector Organisations (NSTSO) – a representative sample of 48,000 third sector organizations in England. They estimate that around 36% of third sector organizations receive some public money (and that 14% of these regarded statutory funding as their most important source of income). The authors show that charities which were bigger, newer, and serving socially excluded or vulnerable people were more likely to receive public funding than other organizations, and we are able to directly analyze these findings with observational data. Backus and Clifford (2013) analyze

the ‘dominance’ of big charities from both a cross-sectional and longitudinal perspective using a panel dataset with information on charities’ income in England and Wales between 1997 and 2008. A number of commentaries and analyses cite the important role of the third sector within the ‘Big Society’ envisaged by David Cameron. While we cite further relevant studies as appropriate in Section 5, no other study has, to our knowledge – within the third sector or otherwise – either developed the necessary tools to construct, or had access to a dataset with the same magnitude or granularity as that which we discuss here.

## 2.4 The Open Data Agenda and General Policy Background

The UK government’s Open Data agenda first became apparent in 1998 when the Cabinet Office published a document called Crown Copyright in the Information Age. In 2011, the U.K. government issued the Code of Recommended Practice for Local Authorities on Data Transparency in order to increase democratic accountability, which was reviewed in 2012, revised in 2014, and further updated in February 2015. Stemming from this, the Local Authority Transparency Code (hereafter LATC) has been made legally binding for most LAs as of 31 October 2014; it requires all councils to publish (on a quarterly basis) items of spending above £500 and to publish contracts and tenders in full. It applies to all of the categories of authority covered by the Code (such as county and district councils), with the exception of parish and town councils with either a gross income or expenditure of under £6.5 million (for who it is recommended, but not legally binding). This Code is issued by the Secretary of State for Communities and Local Government under Section 2 of the Local Government, Planning and Land Act (1980) as a Code of Recommended Practice concerning the release of information by LAs about their functions and other matters. Supplementary guidance in addition to the Code encourages data to be made available for payments of a minimum of £250, reported on a monthly frequency, both of which are seldom undertaken.

While the legislation is bold in its ambition, there are numerous problems with the initial implementation and interpretation of the requirements. There is no requirement for the data to be entered accurately: we frequently observe obtusely entered recipient string names proliferated with spelling and grammatical mistakes. The Department for Communities and Local Government (DCLG) advocates the use of the ‘five-star’ approach to open data as outlined by [Berners-Lee \(2016\)](#). However, there is significant difficulty in aggregating the individual datasets due to the fact that while the legislation encourages the distribution of ‘machine-readable’ files (which would garner a minimum of two stars: “make it available as structured data (e.g., Excel instead of image scan of a table)”), a proportion of them are merely one-star ‘human-readable’ (“make your stuff available on the Web (whatever format)”). This is far removed from the more desirable ‘linked open data’ (“link



your data to other people’s data to provide context”) discussed in the articles mentioned above.<sup>4</sup>

## 3 Data

### 3.1 Local Government Procurement Data

The LATC contains supplemental guidance specifically relating to what exemptions or redactions may or may not apply. However, two main challenges are the extremely heterogeneous format of the raw data at source (and how we deal with this is discussed below), and the considerable problem of reconciling the raw data with registers despite the frequent occurrence of spelling and grammatical errors, arbitrary use of punctuation and the use of abbreviation and acronyms in recipient names. The second issue originates from the fact that the only way to generate a mapping from the raw data to a register is via a string literal typed by a civil-servant without the use of any automated assistance or auto-complete method. Personal information is generally excluded in accordance with the Data Protection and Freedom of Information Acts. The data should also be redacted if it jeopardizes security or an investigation into criminal matters. Schools are themselves excluded, other than when expenditure is incurred directly. However, named individuals can legitimately occur in the context of sole traders or individuals trading under his or her name for business purposes.

We traverse a list of 326 sub-domains of LAs which are subject to the LATC and utilize 11,751 individual datasets, totaling 32,057,175 transactions. We retain 24,581,192 transactions after they are cleaned as part of the process described below, with the majority dropped due to issues which make it unclear as to who the provider is, the date at which the invoice is paid or the invoice amount in question. The data origination begins at different times across different LAs from 2010 onwards, and as of the 31st December, 2015, 302 LAs achieve a minimum of the ‘2 star’ provisions discussed by Sir Berners-Lee and make their data available in a machine readable format for at least one of their monthly datasets.<sup>5</sup> Despite the supplemental guidance on what is required of the files, they are so heterogeneously presented that it remains tractable to only consider retaining (semi-structured) information on the (numerical value) payment amount, date and recipient name (a string literal).

However, even within each LA, there remains a huge array of naming conventions used for each variable. We construct our database through recognition of 88 different naming conventions for the payment recipient (‘supplier name’, ‘beneficiary’, ‘vendor’,...) and 162 different names for the payment amount (‘amount’, ‘value’, ‘paid’,...) and 60 variants of the date of payment

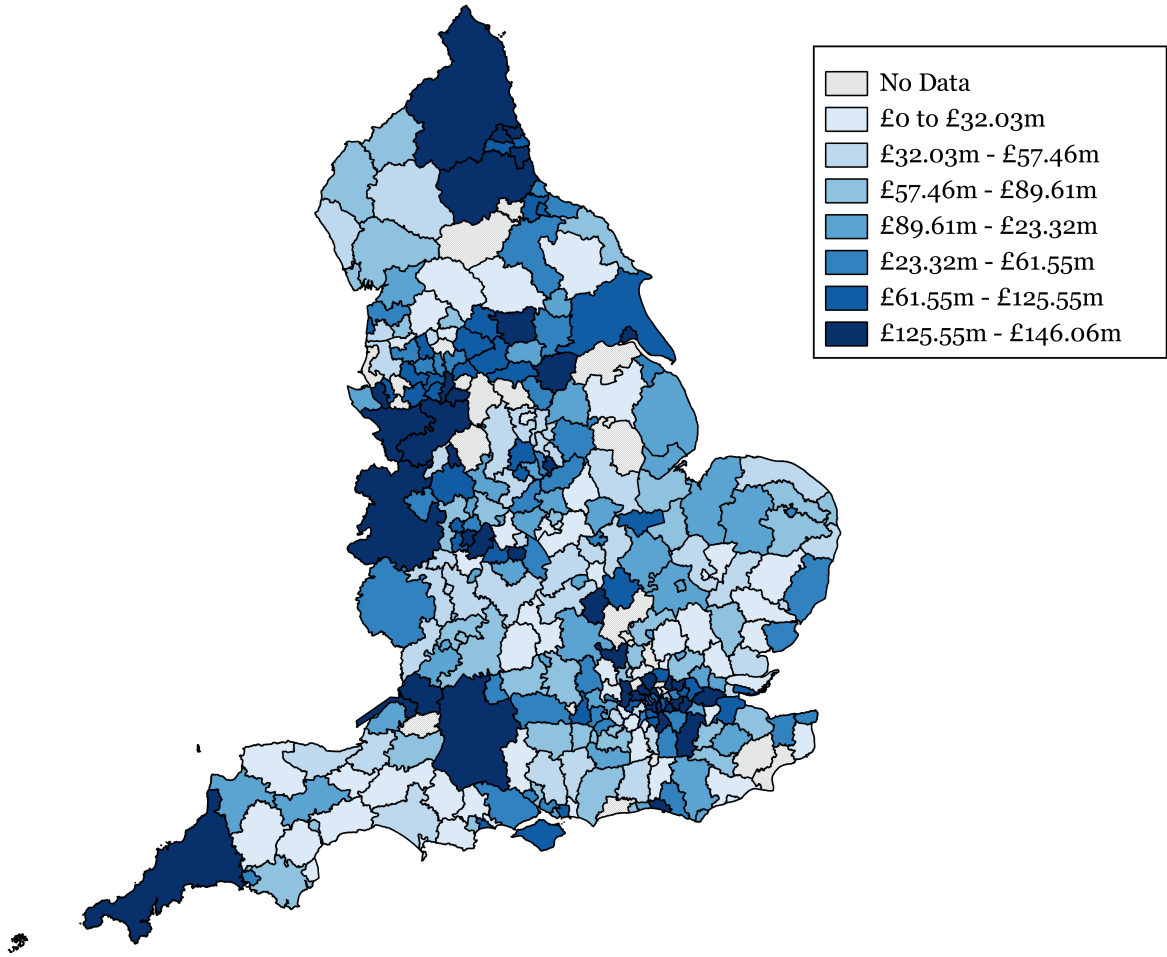
---

<sup>4</sup>However, in the midst of the low quality provision of data, there remain some LAs such as Lichfield District Council and a small minority of exemplary others who even provide unique URI numbers.

<sup>5</sup>The majority of the ‘2 star’ proprietary formats are .xls and .xlsx, and ‘three star’ open, non-proprietary formats are .csv or .json.



Figure 1: Spending Coverage Achieved within the Database Across 302 LAs



variable ('entry date', 'invoice date', 'date paid,...')<sup>6</sup>. We also redact any 'total value' style amount entries which are summations within each dataset, and we manually verify that these have been redacted by sorting payments by magnitude at the back-end to check. The total value of payments is £169.87bn. The mean value is £7,075 and the median is £855. Table 1 displays the ten highest value receivers in the raw dataset (payments aggregated over the unique raw\_beneficiary field), and shows a range of private and public recipients of LA funds. Figure 1 shows a spatial choropleth map of the coverage across England.

---

<sup>6</sup>Although these different string names for dates represent different specific definitions, and while this may be a cause for concern in some specific scenarios, none of our analysis in Section 5 takes a longitudinal perspective which requires further decomposition or classification of this field.

**Table 1: Highest Value String Literal Recipients in Raw LA Data**

Recorded Name	# Payments	Total (£m)	Av. Value (£)
hanson aggregates	5,199	1,999	384,412
greater london authority	1,086	1,380	1,270,876
tyne & wear integrated transport	38	909	23,932,786
eden brown ltd	21,685	825	38,041
sunderland care and support	17	690	40,581,303
hampshire county council	10,370	673	64,922
comensura ltd	9,7623	669	6,852
dept for communities & local goverment	259	647	2,498,788
venn group limited	5,420	613	113,076
hbhc synergy ltd	3,088	566	183,152

### 3.2 Open Source Registers

The ‘open’ nature of the U.K. more generally makes it amenable to the type of work being undertaken because institutional registers are freely available. We attempt to map the data to six key types of register (although multiple sub-registers can form a type) which we believe might contain potential recipients of LA expenditure. Of critical importance is the availability of a free, publicly available corporates register. For this reason, we make use of the Basic Company Data products made available by Companies House. We also make use of the register of charities provided by the Charity Commission (CC), the Office of the Scottish Charity Regulator (OSCR) database of charities and finally, the Charity Commission for Northern Ireland (CCNI) registers. We filter for active charities only.

For our register of health-care institutions, we utilize the official NHS database of 25,195 organizations in England and Wales from the Health and Social Care Information Center. We use the register of public bodies made available by [public-body.register.gov.uk](http://public-body.register.gov.uk), which is then augmented with a list of councils. For our ‘education’ registry, we utilize a list of Key Stage 4 schools made available by the Department of Education and a list of all Higher Education Institutions. For sports clubs, we utilize a government list of all Community Amateur Sports Clubs (CASCs), augmented with a list of sporting institutions provided by Sport England. Finally, we utilize a list of common western forenames from [nrscotland.gov.uk](http://nrscotland.gov.uk) augmented by a list of all known titles (‘Mr.’/‘Mrs.’/‘Dr.’, etc) in order to check for payments going to named individuals. Figure 2 shows the most frequent words from each of the six registers discussed above. The figure helps to conceptualize the types of institution which form the constituent parts in each of the registers, and a longer list (up to 100) of the most frequent words from each register are utilized in our ‘targeted’ approximate

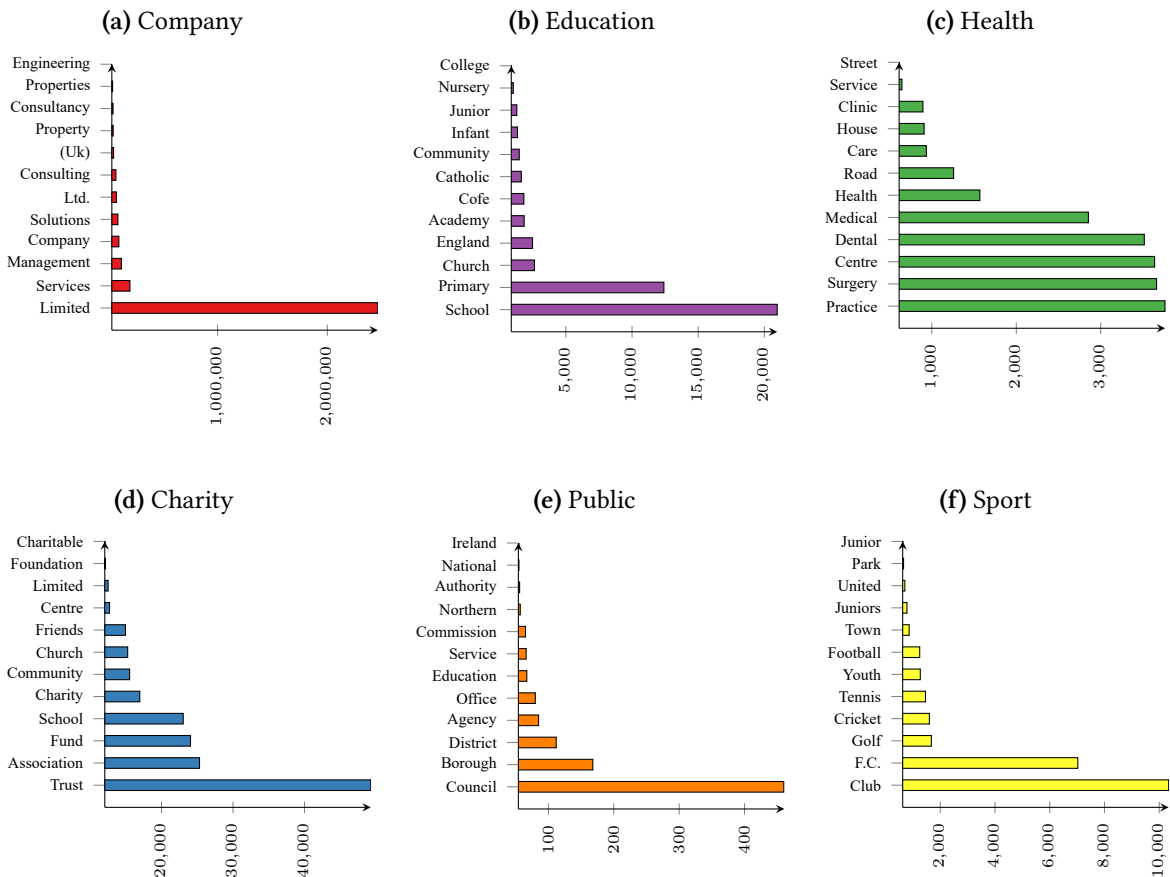
string matching algorithms discussed below.<sup>7</sup>

## 4 Methodology

### 4.1 Parsing the Raw Data

An outline of our algorithmic approach is shown in Figure 3. The first stage of the process is to download all of the data and use a simple script in order to consistently export the raw, extremely noisy data into a structured database. We parse each data record (payment) into a tabular format (comma separated values), retaining three key, consistent fields which we label ‘amount’, ‘supplier’ and ‘date’ on our local database. We drop all redacted payment records instantly in order to retain anonymity where required by the source based on a list of 70 separate ways of citing redaction which occur within the raw files (such as named vendor, private individual, withheld, etc). Each

Figure 2: Most Common Words by Register



<sup>7</sup>The code which calculates these word counts is termed `wordcountfunction` in the accompanying replication files.

raw payments dataset is ran through a subroutine which parses the information while attributing a filename and LA to each record, parsing files with a .tsv, .json, .csv, .xls or .xlsx extension. Our scripts perform a number of critical tasks to make this possible. They look for problems which hinder machine-readability, such as bypassing image files (typically council logos), rows at the top of the dataset which seldom contain information about extraction from LA accounting systems, and a variable number of blank or noisy (non-data record) rows within the header of the file. They automatically remove payment/transaction numbers, which are sometimes entered within the ‘recipient’ field due to administrative error (and for one LA – a trait of suffixing an ‘expense’ area into the recipient field is problematic). If a field of structured string literals is expected, and the data contains an array of predominantly numeric data, the file is discarded (and vice versa when expecting string literals in the numeric/date fields). We retain string literals which are character encoded within ASCII only. The critical feature which makes the reconciliation algorithms computationally tractable is the extraction of a unique list of supplier names. This is a computational necessity as every supplier (or rather: every string literal) appears to receive multiple (potentially contracted) payments. The 24,075,158 payments reduce to 562,019 unique suppliers. From this contraction, we can assert that every unique string literal receives a mean number of 42.837 payments (although several string literals can reference the same supplier due to spelling or administrative errors).

## 4.2 Normalization

Once the data is parsed into a tabular database and an array of unique supplier names is created, we generate a normalized variant of each individual record (in addition to creating a set of normalized registers to match the unique normalized entries to). This stage is especially important,

**Figure 3: Simplified Process Chart**

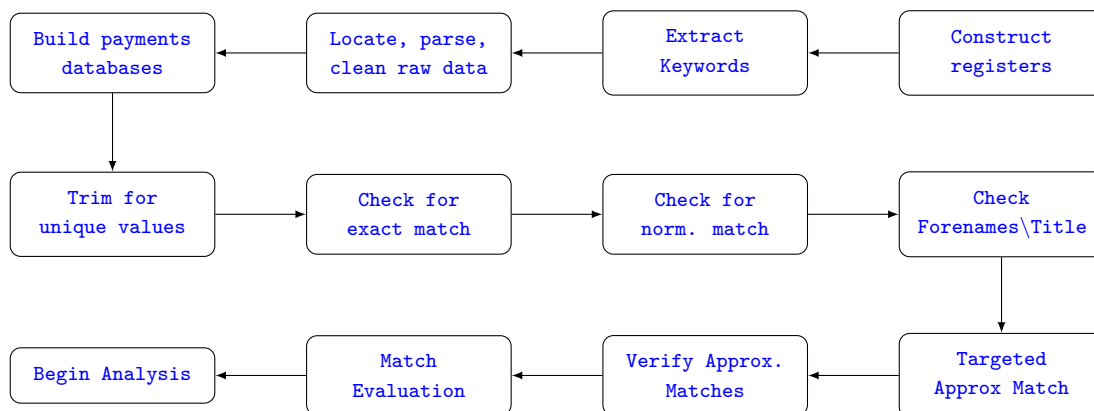
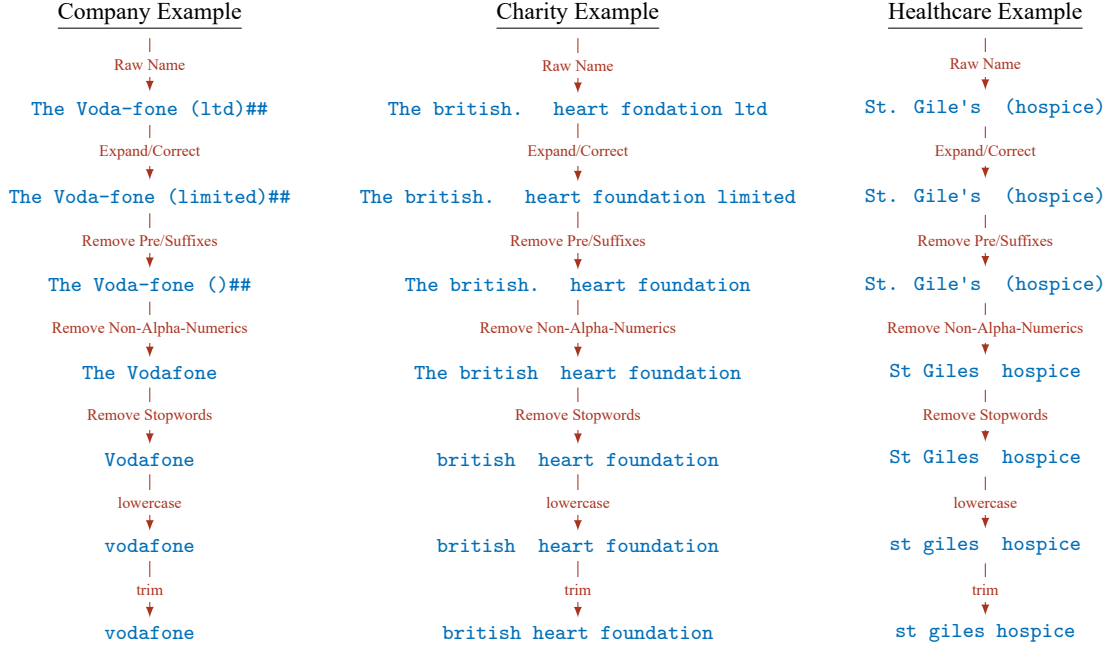


Figure 4: Normalizing the String Literals



as acknowledged in the record linkage literature: ‘With good standardization, effective comparison of corresponding components of information and the advanced methods described in this paper become possible.’ (Porter *et al.*, 1997, p.190). Our normalization function<sup>8</sup> takes a stepwise approach, and is not entirely dissimilar to Alvarez-Rodríguez *et al.* (2015). Our approach is calibrated to be more appropriate for the data source in our application (U.K. LA expenditure), and only requires a set of standard regular expression tools. The first step is a dictionary-based expansion/replacement of common acronyms or mis-spelled words. We expand/replace a list of 522 commonly used acronyms (such as c.i.c. expanding to ‘community interest company’) and spelling mistakes particular to this type of data. The second step removes 197 common prefixes and suffixes from the strings (such as ‘associates’ or ‘services’). We then remove any non-alpha-numeric characters. Following this, we remove any of a list of 179 stopwords to filter non-relevant words (such as conjunctions and prepositions) – an approach prevalent in the natural language processing literature. We then make all case-based characters lowercased, trim leading and trailing whitespace, delete double spaces and then nullify strings less than 3 characters in length. Figure 4 provides three examples of how normalizing noisy, manually recorded string literals can result in a suitably contracted set of characters which can be reconciled with a normalized registry entrant. We do not deploy an automated spell check algorithm with training data. This decision should be driven by the

<sup>8</sup>This is called `normalization_function` in the replication files.

empirical success of such an approach, and [Alvarez-Rodríguez \*et al.\* \(2015\)](#) cites preference for a stopwords set/dictionary due to the fact that ‘some spelling corrections are not completely adequate for corporate names’.

### 4.3 Reconciliation Strategy

The matching script (called `matching_script` in the replication files) directly maps payment recipient names from the procurement data to institutional registers in a variety of ways, and forms the ‘master’ file of the code which calls other functions. Such an approach is necessary as: “Although improvements in available computing power have to some extent mitigated against the effects of this accelerating growth in the size of the data sets to be linked, large-scale probabilistic record linkage is still a slow and resource-intensive process” ([Christen \*et al.\*, 2002](#)). A detailed schematic of our reconciliation approach is shown in Figure 5. We also take an iterative, stepwise approach to matching in order to minimize computational burden on a huge set of string based searches. We first search for exact matches on raw and normalized unique suppliers and with raw and normalized registers respectively. Following this, we bring in a series of pre-computed matches generated from the OpenCorporates registers. Specifically, OpenCorporates provides a highly popular reconciliation API for Open Refine, which allows matching company names to legal entities.<sup>9</sup> After this, we check the string literal for either a common title (such as ‘mr’, ‘mrs’, ‘dr’, etc), or a common forename from the register detailed above in Section 3.2.

#### 4.3.1 ‘Targeted’ Approximate String Matches

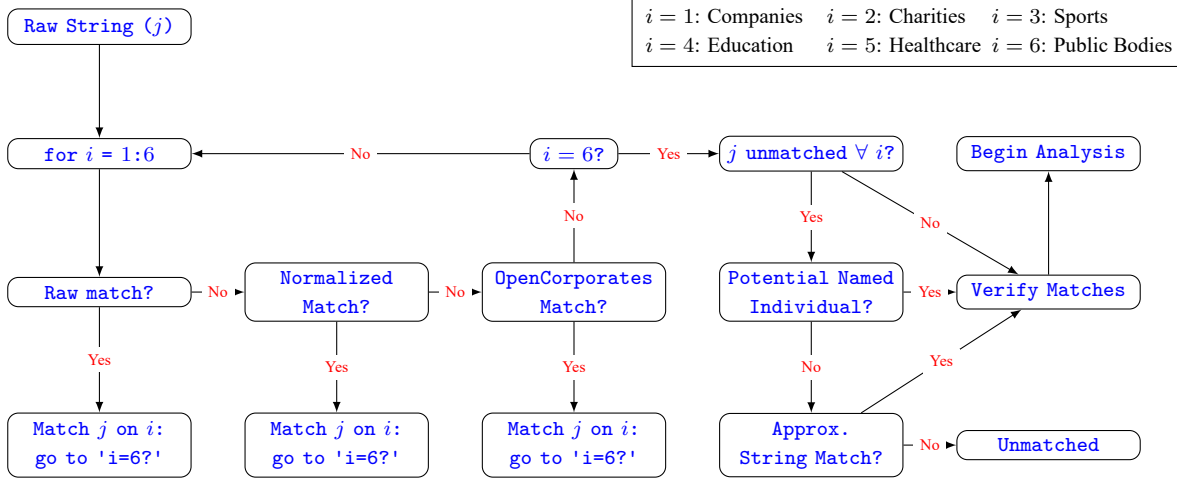
Due to the computational intensity of traversing a list of 562,019 string literals of with an average length of 23.7 characters against multiple registers (some as long as 3.4 million entrants) with no other means of reconciliation, we undertake approximate string matching last as part of the iterative procedure described above and shown in Figure 5. One of the main methodological contributions of this paper to the reconciliation literature is to advance the approximate string matching strategy for procurement data by considering ‘targeted approximate matches’, where the goal is to ‘reduce number of pairs on which similarity is computed’ ([Koudas \*et al.\*, 2006](#)).<sup>10</sup> The accompanying function is called `approximate_matches` in the replication files, and works as follows. In order to condense

---

<sup>9</sup>The basis for reconciliation at OpenCorporates is not all that dissimilar from approach we detail: “The search is case-insensitive and returns companies with previous names matching the term as well as current name, and some normalization of the company names is done, removing non-text characters (e.g. dashes, parentheses, commas), common ‘stop words’ (e.g. ‘the’, ‘of’), and normalizing common company types (e.g. Corp, Inc, Ltd, PLC) so that both the short and long versions can be used.” ([OpenCorporates, 2017](#)).

<sup>10</sup>This is akin to [Enamorado \*et al.\* \(2017\)](#), who aim to develop a fast and scalable algorithm to implement the canonical probabilistic record linkage model originally proposed by [Fellegi and Sunter \(1969\)](#) by aiming to ‘reduce the number of assignments by focusing on those pairs whose posterior probabilities are greater than a threshold’.

Figure 5: A Flow Diagram For Full Matches



the computational burden and make our algorithm completable on local machines with a single core (as opposed to distributed clusters of machines ran in parallel), we utilize a list of the 100 most common keywords from the register which we are searching across (the 20 most common from each register are shown in Figure 2). If the string contains one of these 100 keywords (and is above 5 characters in length), it is approximately string matched against a contracted series of the appropriate register (which is also contracted to contain only elements which contain the keyword), with the keyword removed from both the string to matched and the contracted register. We then use a standard Dice Coefficient based calculation (utilizing integer based bigrams in conjunction with numba's @autojit for to enhance speed of calculation), returning the highest scoring match from the candidate register for strings with a score above 85.

All normalized and approximate string matches are then manually verified. This is because there is no way of automatically creating such a mapping with perfect confidence in an approximate match without human validation. An example of the need for human validation can be seen by the fact that despite being two separate entries on the Companies House register, 'stagecoach east midlands' and 'stagecoach west midlands' requires two 'edits' only, and this is the closest approximate match: but clearly not the exact same subsidiary. We also examine the unmatched strings for patterns which can improve our normalization technique in future iterations.

#### 4.4 Match Evaluation

Figure 6 shows the distribution across the various types of reconciliation. In particular, it highlights the importance of the normalization and approximate string matching parts of the matching algorithm – especially for matches to the Companies House register. The large proportion of



Figure 6: Distribution Over Types of Match

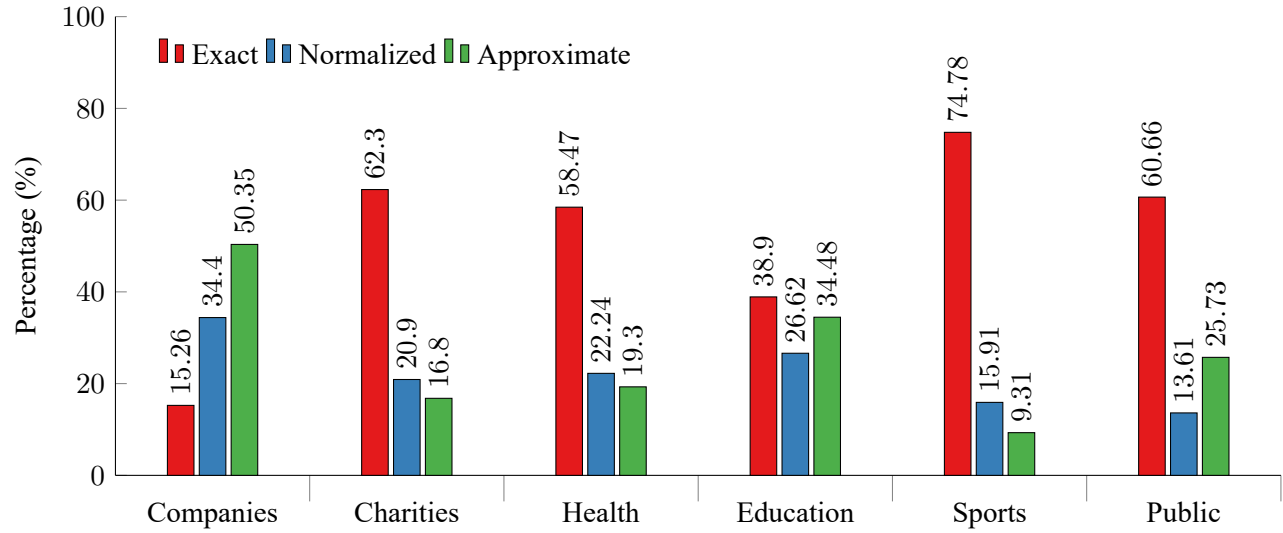
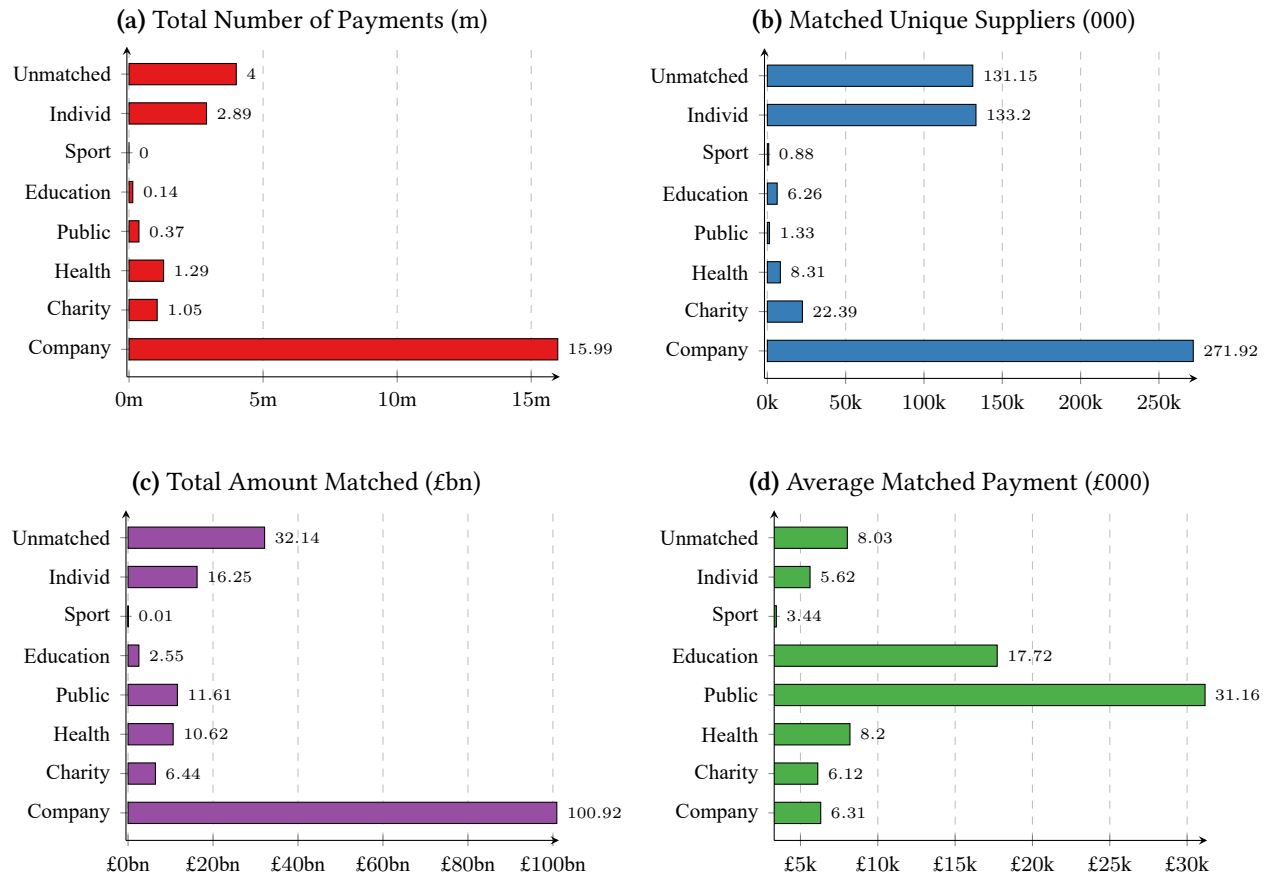


Figure 7: Reconciliation Evaluation Metrics

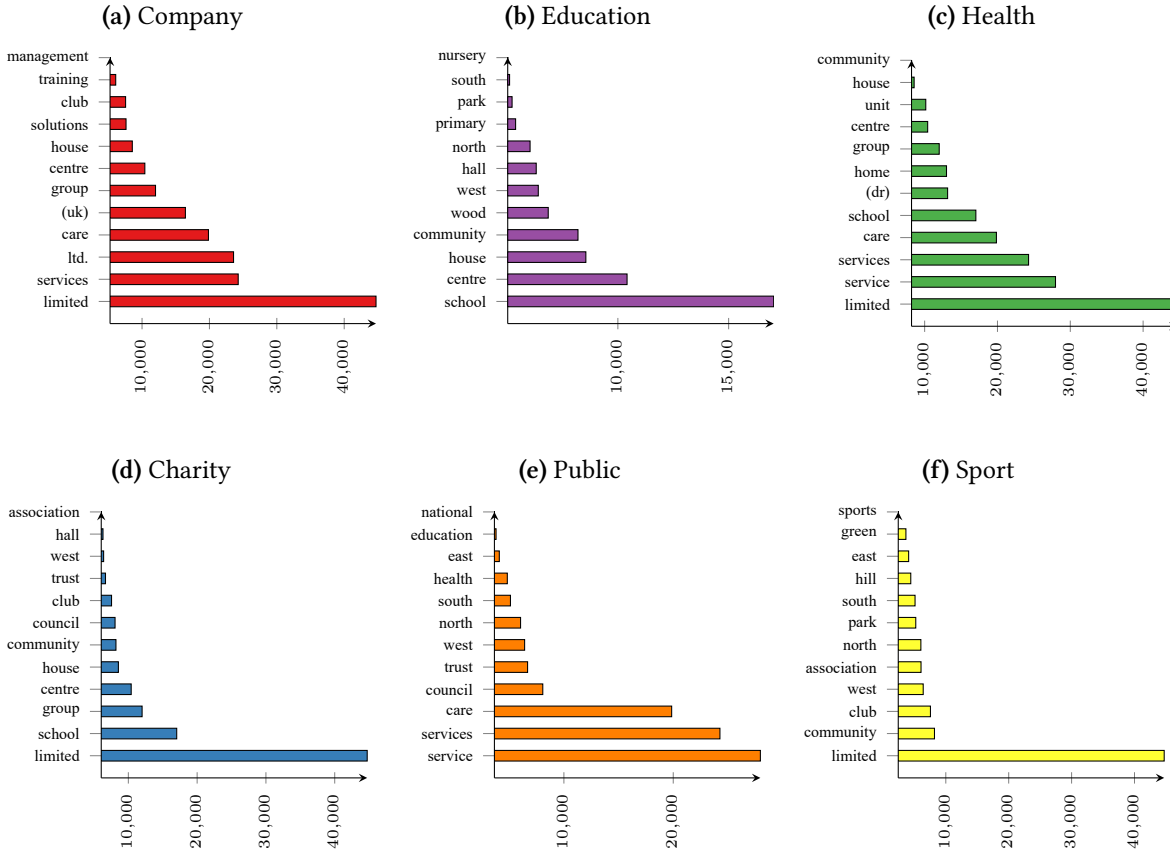


approximate matches to Companies House is driven by the inclusion of the (company entity-specific) OpenCorporates API. This indicates the returns to investing in the development of an entity-specific normalization routine. The gains from the approximate routine are considerably higher than the 30% improvement achieved by [Enamorado \*et al.\* \(2017\)](#) when compared against exact matches only. Figure 7 shows what matches occur across which register, highlighting the dominance of the private sector in the local government procurement process. In particular, it shows that entities matched to the Companies House register supply services worth over £100bn in our incomplete dataset alone. We can successfully attribute around 77% of `raw_beneficiary` string literals to one of the matching registers (only 131,150 out of 562,019 remain unmatched). These supplier matches represent all but about four million out of twenty four and a half million individual payments (about 84%). This, in turn represents 88% of payments by value matching successfully to at least one, but potentially more registers. While the objective and the evaluation matches differ slightly, our method compares favorably to the ‘unification’ of 77,526 matches in of Australian data of [Alvarez-Rodríguez \*et al.\* \(2015\)](#) who achieve 48% when evaluated across the entire dataset (rising to 100% when only considering Forbes 100 companies). A more conventional metric where we undertake searches for true and false positives and negatives by hand indicates that while the matching algorithm performs extremely robustly, there exists a slight problem in the reconciliation strategy for named individual suppliers. For example: the raw supplier ‘national door and domelight company’ contains the Irish forename of ‘iona’ (and ‘iona’ is matched the most, at 221 times), generating a false positive. While this is left for further research, the position of this function in the iterative approach renders it largely insubstantial.

Table 2 represents a frequency tabulation of the most frequently unmatched `raw_beneficiary` strings. From this table alone we are able to make four important observations. The first is that while we map against six different institutional registers, there is clearly the potential to expand this to additional registers (such as, for example – a list of libraries and police stations). Secondly, some of the highest frequency recipients are receiving a large number of individual payments, as indicated by the low average payment value column. Thirdly, the table also provides an insight into some of the difficulties in matching such messy data. For example, the seventh most frequently observed, yet unmatched string is ‘o’rourke construction and \* 1174630’ – which is naturally extremely challenging to match appropriately. The final observation to make regards ‘cornwall external supplier’ – payments made by Cornwall Council to a nondescript ‘external’ supplier – further highlighting the complexity of some of the challenges which we are facing.

Figure 8 details the frequency distribution of the most common individual strings in the unmatched recipients using a list of the 100 most frequently occurring strings in each of the six institutional registers. The figure shows no real systematic bias against any particular unmatched

**Figure 8: Most Commonly Unmatched Words by Register**



type of institution: while ‘limited’, ‘Ltd’ and ‘services’ are the most frequently unmatched, we would naturally expect them to be so given the aforementioned concentration of procurement from the private sector. It does, however, provide evidence of some slight biases in our registers, with a potential need to incorporate more detailed registers which contain ‘school’ and ‘council’ type institutions. <sup>11</sup>

## 5 Analysis

In this section we merge the dataset of raw payments with the unique reconciled strings as described in Section 4. We then compliment this with a range of relational tables from the Charity Commission database in order to examine procurement by LAs within Great Britain from the voluntary sector. Such an evaluation is important and timely, as the third sector in the UK has been growing in reliance on public funding (Clifford *et al.*, 2010). NCVO (2016), for example, estimates that £15bn

<sup>11</sup>The script which conducts the match evaluation is named `evaluation_script` in the code repository.

**Table 2: Ten Most Frequent Unmatched Recipients**

raw_beneficiary	# Payments	Total (£m)	Av. Value (£)
bertram library services	27,689	12.21	441
mears ltd (diskette)	26,227	19.90	759
london care plc	18,452	36.04	1,953
west mercia supplies	17,776	33.80	1,902
brake grocery	16,929	2.18	129
askews library service	15,766	0.87	55
o'rourke construction and * 1174630	15,346	3.62	236
phs group plc	15,259	11.10	727
the imprest holder	14,129	2.41	171
cornwall external supplier	14,120	43.08	3,051

of the total £43.8bn voluntary sector income was from government in 2013/2014. Concerning the reconciliation to specific charities, our matching algorithm ‘suggests’ the best charity through a confidence based preference ordering: exact matches take precedence over normalized matches which take precedence over approximate matches. In each of the following subsections we attempt to disaggregate the overall picture and contribute to existing analysis using our new reconciled dataset.

## 5.1 Highest Value Third Sector Recipients

The first application which we showcase is similar in design to Tables 1 and 2. Table 3 details the ten (reconciled) charities which receive the highest total amount of procurement by value across all LAs. In contrast to the other tables, the final column (CC Income (£m)) details the corresponding income

**Table 3: Ten Highest Value Charity Recipients**

Charity	CC Number	ICNPO SubGroup	Total Value (£m)	CC Income (£m)
United Response	265249	Social Services	123.42	79.08
North East Autism Society	1028260	Social Services	93.03	17.37
Action for Children	1068215	Social Services	80.16	160.88
Orders of St. John	1048355	Nursing Homes	79.92	110.61
St Anne’s Comm. Services	502224	Housing	26.03	43.23
Anchor Trust	1052183	Housing	42.29	367.33
Addaction	1001957	Social Services	53.14	66.94
The Brandon Trust	801571	Social Services	55.23	46.63
Real Life Options	1156258	N/A	53.43	34.63
Comm. Int. Care	519996	Nursing Homes	47.94	107

for the appropriate charity number for the last financial year.<sup>12</sup> The charity which receives the largest value of payments is United Response (registered charity number 265249), which provides a range of support and services for more than 2,000 people with learning disabilities, mental health needs or physical disabilities across England and Wales. The charity which received the largest number of payments by frequency is The Orders of St. John Care Trust (registered charity number 1048355). The charity which receives the largest average payment is Keelman Homes Limited (averaging £1,500,000 over three payments): a charity (housing association) responsible for the introduction of new housing stock within the Kibblesworth area.

## 5.2 Payments against Charity Size (Income)

The next application is to cross-reference the payments received against the incomes provided by the charity’s record of submitting accounts, Annual Returns and/or Annual Updates for the last 5 years to the Charity Commission. The string-matched payment recipient names are reconciled with their appropriate registration numbers within the Charity Commission bulk data download, and then merged with the appropriate look-up table on financial information in order to match charities receiving LA money with their respective income.

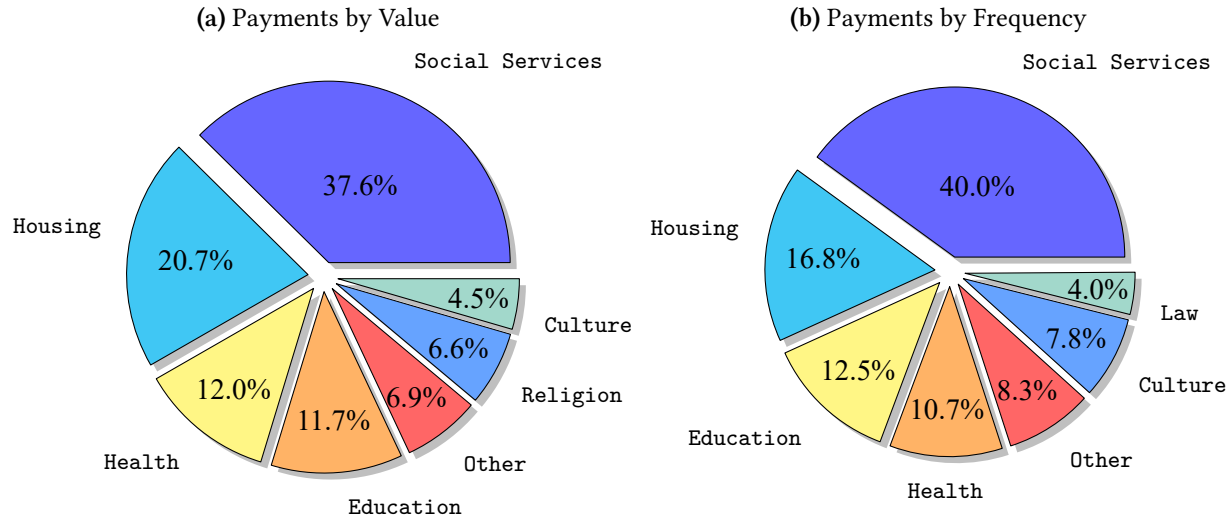
Analyzing the ‘concentration’ of charity financing is important given the current debate on the potential dominance of bigger charities. [Duncan-Smith \(2005\)](#) claimed ‘There now seems to be an established trend towards the concentration of income in the very richest charities... with a small minority of large charities becoming ever more dominant’, calling it the ‘Tesco-isation’ of the charity sector. We analyze across quantile group shares defined as the charities with highest incomes of all those receiving LA funding. Table 4 presents the payment value shares of the top 1%, 10%, 25% and 50% charities by income, similar to [Backus and Clifford \(2013\)](#), and what percent of the LA payments they receive by value and frequency. These results lend some significant support to this ‘Tesco-isation’ hypothesis, showing that there is a considerable bias towards the larger charities in our sample, where further work would decompose this as a concentration of the income within each of these charities.

**Table 4:** Concentration of Activity by Income

Percentiles of Charities by Income	Top 1%	Top 10%	Top 25%	Top 50%
Percent of Payments by Value	18.93	62.12	79.40	92.35
Percent of Payments by Frequency	24.68	59.49	78.04	89.94

<sup>12</sup>Please recall that the Total Value (£m) column is calculated over approximately six years (2010-2015) worth of LA payments data.

**Figure 9: Payments by ICNPO Group**



### 5.3 Distribution Across ICNPO Numbers

We use registered charity numbers to merge with the look-up tables of International Classification of Nonprofit Organizations (ICNPO) categories provided by the National Council for Voluntary Organizations (NCVO). The ICNPO system was designed by the Center for Civil Society Studies at Johns Hopkins University in the United States as part of efforts to draw up a UN Satellite Account for the nonprofit sector (and outlined in [Salamon and Anheier, 1996](#)). The NCVO categorization is based on a largely automated approach which includes keyword searches, matching to other registers and looking at individual sources, and is discussed in [Kane \(2008\)](#). The majority of charities were classified using common phrases in their charitable objects, precise keywords in their name or using the CASCOT program. Our analysis (in Figure 9 and Table 5) shows that charities within the ‘Social Services’ ICNPO grouping receive the largest amount of payments by both frequency and value, followed by Housing, Health and Education. This is consistent with [Kane \(2008\)](#) who most frequently classifies charities within the ‘Social Services’ category (24,242 out of 169,224 in their full sample of charities) and [Clifford \*et al.\* \(2013\)](#) who show that ‘organisations serving the personally or socially disadvantaged are most likely to be publicly funded’.

### 5.4 Start Date

It is also possible to examine the relationship between the age of the charitable organizations supplying the LA and the amount which they receive. One competing theory might suggest that more established charities are more likely to receive more lucrative contracts to supply LAs due to their existing relationships with the public sector, or due to their reputational capital.

**Table 5: Local Authority Payments Across ICNPO Subgroups**

ICNPO	Group	Subgroup	NCVO Category	# Payments	Amount (£m)
1100	Culture & Recreation	Culture and Arts	Culture & recreation	27236	262.2
1200	Culture & Recreation	Sports	Culture & recreation	18370	209.0
1300	Culture & Recreation	Other Recreation & Social Clubs	Culture & recreation	888	7.0
2100	Education & Research	Primary & Secondary Educ.	Education	29859	311.3
2110	Education & Research	Parent Teacher Associations	Parent Teacher Assoc.	2026	15.8
2120	Education & Research	Educational Foundations	Education	1121	11.1
2130	Education & Research	Playgroups and nurseries	Playgroups & nurseries	44724	218.6
2200	Education & Research	Higher Education	Education	2194	9.0
2300	Education & Research	Other Education	Education	25133	107.5
2400	Education & Research	Research	Research	14685	88.0
2410	Education & Research	Medical Research	Research	916	3.1
3100	Health	Hospitals and Rehabilitation	Health	11148	125.6
3200	Health	Nursing Homes	Health	70064	273.1
3300	Health	Mental Health & Crisis Intervention	Health	34634	193.2
3400	Health	Other Health Services	Health	8328	62.4
4100	Social Services	Social Services	Social Services	368918	2338.4
4110	Social Services	Scout groups & youth clubs	Scout groups/youth clubs	3111	37.3
4200	Social Services	Emergency and Relief	Social Services	7183	36.9
4300	Social Services	Income Support & Maintenance	Social Services	9313	35.8
5100	Environment	Environment	Environment	14187	150.8
5200	Environment	Animal Protection	Environment	3731	20.8
6100	Development & Housing	Economic, Social/Comm. Dev.	Development	32208	193.0
6110	Development & Housing	Village Halls	Village Halls	3743	6.5
6200	Development & Housing	Housing	Housing	159960	748.8
6300	Development & Housing	Employment and Training	Employment & training	18479	79.2
7100	Law, Advocacy & Politics	Civic and Advocacy Organizations	Law & advocacy	11653	100.6
7200	Law, Advocacy & Politics	Law and Legal Services	Law & advocacy	10083	142.5
7300	Law, Advocacy & Politics	Political Organizations	Law & advocacy	10	0.0
8100	Philanthropics	Grant-making Foundations	Grant-makers	5349	34.8
8200	Philanthropics	Other Philanthropic	Umbrella bodies	13625	75.3
9100	International	International activities	International	5634	14.6
10100	Religion	Religious congregations	Religion	67999	198.5
11100	Business/Prof. Assoc.	Business associations	Other	2081	4.6
11200	Business/Prof. Assoc.	Professional associations	Other	4714	7.4
11300	Business/Prof. Assoc.	Labour Unions	Other	1	0.0
12100	Not Elsewhere Classified	Not classified	Other	151	0.1

Another theory might suggest that newer charities might originate with the intention of specifically providing services in response to unmet demands. To analyze this, we plot the percent of payments by volume and value to reconciled charitable services which are then matched with the registration date on the Charity Commission (Figure 10). Our results can be compared with (Clifford *et al.*, 2010, p.11), which claims: ‘Charities registered in the 2000-2008 period are 2.5 times more likely to regard the public sector as their most important source of income than charities registered prior to 1970’.

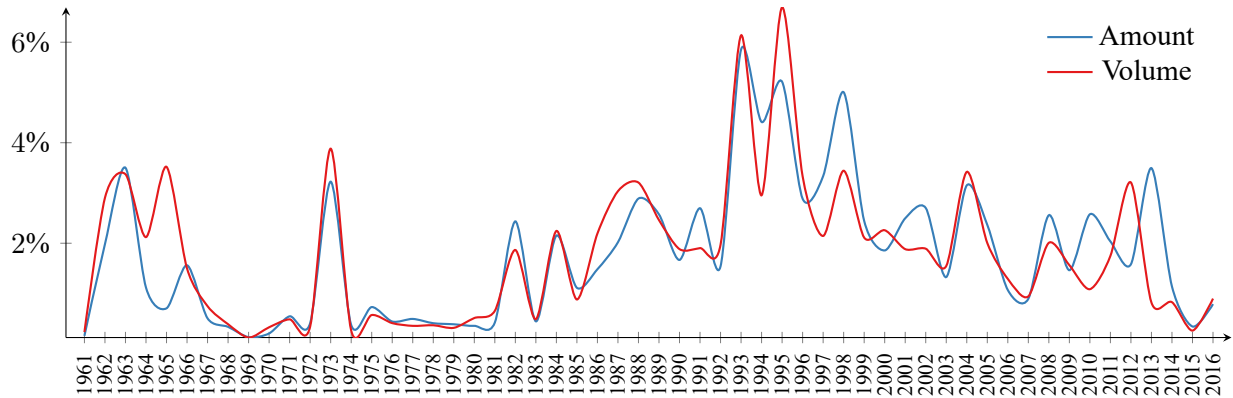
## 5.5 Geographic Variation

### 5.5.1 LAs which Procure the Most from the Third Sector

The geographic variation in third sector activity is well-studied, including questions on where a charity operates, the scope of their activities and where their beneficiaries live. However, the issue is plagued by what Kane and Clark (2009) call the ‘head-quarters’ problem, whereby



**Figure 10: Payments and Charity Inception**



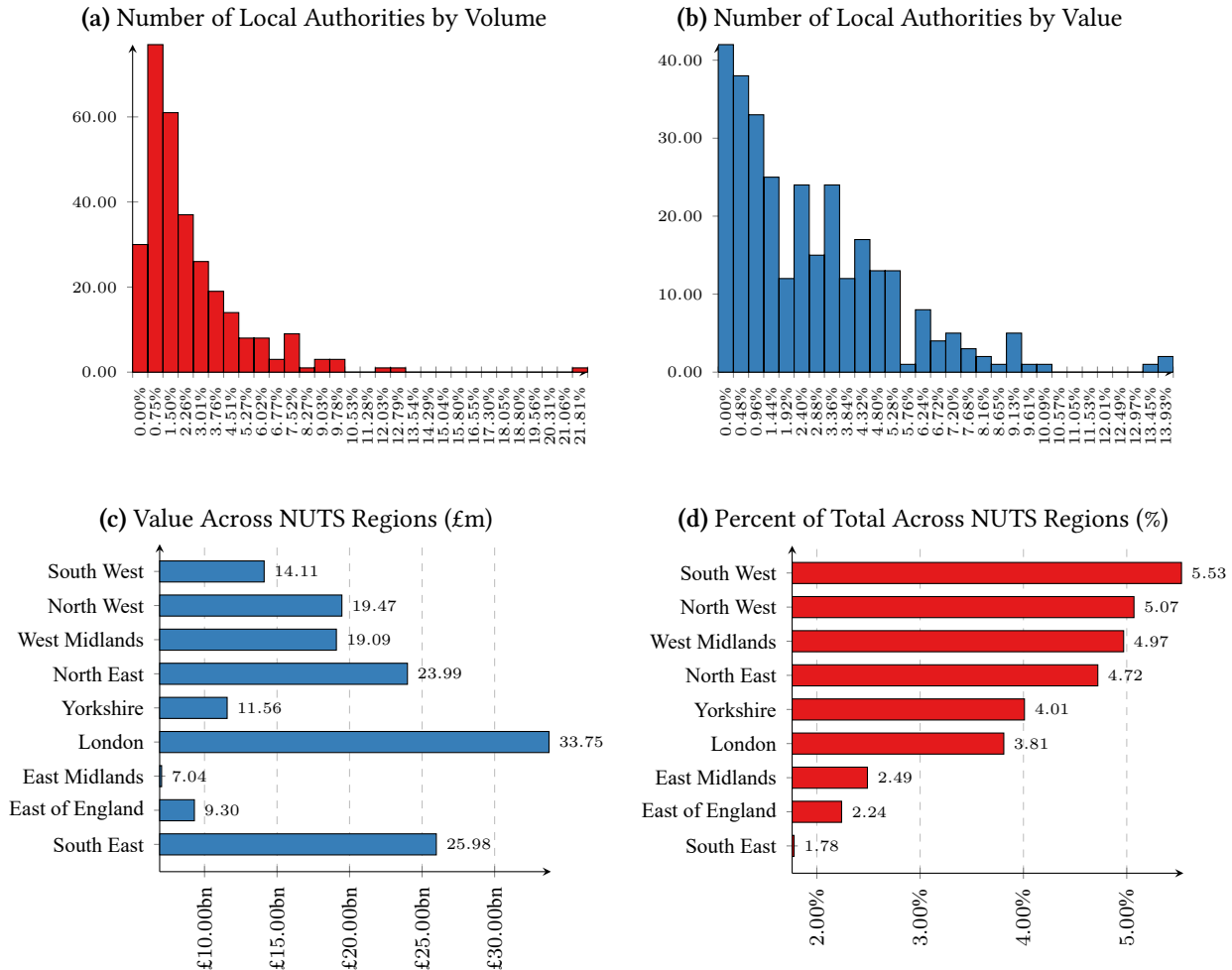
charitable expenditure is attached to a geographical unit based on their registered address. This can be especially distortive for the analysis of small-scale or local charities who have no specific headquarters, but instead register under the home address of a nominated trustee. [Kane and Clark \(2009\)](#) claim ‘the total expenditure of charities in the UK is £31.2 billion; out of this amount £12.9 billion (41%) is spent by charities which are registered in London’. Analyzing a dataset of public payments originating from a range of specific geographic points allows us to consider a new dimension to this problem.

Variations in expenditure by LA are likely to reflect two things. The first is the mix of third sector organizations operating within that area. The second is that the procurement practice across LAs is likely to be variable. Table 6 shows the ten LAs which procure the most from the third sector (by percentage), as well as how much they spend with Charity Commission registered institutions in

**Table 6: Local Authorities which Procure the Most from Third Sector**

Local Authority	Charity Value (£m)	% to Charities	Most Common Charity
Forest Heath	31.27	14.41	Anglia Community Leisure
Rochdale	346.73	14.15	Child Action Northwest
Burnley	9.58	13.79	Burnley Leisure
Southhams	67.19	10.55	Tor Homes
Solihull	1152.26	9.81	Family Care Trust
Carlisle	79.91	9.41	Tullie House Museum
Windsor & Maidenhead	100.09	9.37	People to Places
Gateshead	825.23	9.33	Anchor Trust
Coventry	1618.66	9.18	Life Path Trust
Wiltshire	1843.69	9.13	Orders of St John Trust

**Figure 11: The Distributions Across all Local Authorities**



total and the charity which they procure from most frequently.<sup>13</sup> The first two sub-figures of Figure 11 expand this analysis to encompass each of the 302 LAs within our sample of payments, showing histograms of percentages of payments by value and frequency to the third sector. Furthermore, the average percent of payments to third sector institutions is 2.80%, and the average value of all spending is 2.93%.

Finally, the latter two subfigures of 11 aggregates the spend into Nomenclature of Territorial Units for Statistics (NUTS) statistical regions based on the geographic locale of each LA. This enables comparison with other existing studies such as NCVO (2016), which finds substantial regional variation. This analysis shows that the highest total amount of spending to Charity Commission registered institutions originates from LAs within the London area, but that it is LAs within the

<sup>13</sup>The fact that Tor Homes was removed from Charity Commission in December 2011 poses an interesting question of how to refine the reconciliation approach, which is left for further work.

South West and North West which procure the highest value by percentage.<sup>14</sup> A natural extension to this work would be to supplement these findings with deprivation data from the Office for National Statistics at the lower super output area (LSOA) level, as per Clifford *et al.* (2013).

## 6 Extensions

While our applications focused uniquely on LA level data in England, there are countless other potential case-studies both within the United Kingdom and abroad which can utilize and expand on the tools developed in this paper. For example: on December 17th, 2014, the Government issued the Transparency Code for Smaller Authorities under Section 2 of the 1980 Act. It applies to parish councils, charter trustees, internal drainage boards and port health authorities with a ‘turnover’ of under £25,000, requiring publication of information on all items of expenditure above £100. A parallel transparency requirement also binds central government (ministerial and non-ministerial) departments to make data on payments above £25,000 available.

In the United States, the Federal Funding Accountability and Transparency Act of 2006 was signed into law on September 26, 2006 and mandated the creation of a publicly available, searchable website similar to the data provided within the U.K. The data is hosted on [USAspending.gov](http://USAspending.gov), with an API which makes available information on each contract such as the vendorname, dollarsobligated and the signeddate. Within the European Union, a subset of Tenders Electronic Daily (TED) data covering public procurement for the European Economic Area, Switzerland, and the former Yugoslav Republic of Macedonia is made available in aggregated comma separated value format. Included are fields such as WIN\_NAME, AWARD\_VALUE\_EURO, and DT\_DISPATCH.<sup>15</sup> One of the main advantages of obtaining data from centralized sources such as this is that not only is the data in a consistently structured format, but that there are a large number of supplemental fields.

The type of analysis is not limited to a deeply descriptive approach as undertaken in this paper. Future work, for example, might take more explicitly geo-spatial approaches based on Geographic Information Systems (GIS). They might also consider not only the network science of subsidiaries of suppliers within the procurement chain, but also the network of officers involved in each subsidiary themselves. Supervised machine learning algorithms can be trained to spot potentially sensitive material and suggest redactions to LAs at the data origination stage, much in the same way that fraudulent bank transfers are detected. Another project could involve a systematic evaluation of the success of various automated cut-off values for approximate matching on supplier names, or a more rigorous comparison of different linkage algorithms altogether (comparing, for example, Jaro-

---

<sup>14</sup>There are substantial further applications based on standard GIS techniques, which are left for further work.

<sup>15</sup>Data is available in a range of other countries, such as AusTender and [data.gov.au](http://data.gov.au) in Australia.

Winkler, Hamming or other distance metrics found in Python implementations such as *jellyfish* with proprietary algorithms such as Big Match for Hadoop by IBM). This might be especially useful for private sector firms looking to predict future areas of public procurement. There is also a substantial, yet unexamined political element to this paper: what about the characteristics of the LAs? Do Conservative controlled councils behave substantially different (with regards to voluntary income or otherwise) to their Labour or Liberal Democrat counterparts? From a logistical, data provision perspective, further work and infrastructure investment is required to create centralized interfaces and APIs which are able to query individual public office payments at an international level.

## 7 Conclusion

In this paper we have shown how to operationalize one of the largest sources of currently unexamined data in developed, ‘Open Data’ economies. We develop various tools to aggregate disparate data and create previously unobtainable aggregated databases of public spending. The database which we construct to document our approach contains over 24 million unique transactions, and contains approximately £170bn of LA spending in England. We then provide an example of what this approach can make possible in the form of a deep descriptive analysis of local government procurement from the charitable sector.

We hope that this work will help researchers overcome the challenges involved in realizing the full value of such an important set of resources until an aggregated, standardized provision is one day possible. The next step is to utilize classification and clustering based algorithms to assist in the reconciliation task, improving on the Dice Coefficient based approach discussed herein. To this end, the exploratory framework described above can be thought of as modular, in that different steps can be replaced with others as the open-source code is refined further (such as a methodological advancement similar to [Enamorado \*et al.\*, 2017](#)). While research involving increasingly bigger, innovative types of open data will only become more important over time (within this domain and others), data originators too have an important role to play in minimizing the costs associated with aggregating and cleaning the data at source.

## References

Alvarez, J.M., Labra, J.E., Cifuentes, F. *et al.* (2012) Towards a pan-european e-procurement platform to aggregate, publish and search public procurement notices powered by linked open data: The moldeas approach. *International Journal of Software Engineering and Knowledge Engineering*, 22 (03): 365–

383. URL <http://www.worldscientific.com/doi/abs/10.1142/S0218194012400086>.
- Alvarez-Rodríguez, J.M., Vafopoulos, M. and Llorensm, J. (2015) Enabling policy making processes by unifying and reconciling corporate names in public procurement data. The CORFU technique. **Computer Standards and Interfaces**, 41 (1): 28–38. URL <http://dx.doi.org/10.1016/j.csi.2015.02.009>.
- Backus, P. and Clifford, D. (2013) Are big charities becoming more dominant?: cross-sectional and longitudinal perspectives. **Journal of the Royal Statistical Society: Series A (Statistics in Society)**, 176 (3): 761–776. URL <http://dx.doi.org/10.1111/j.1467-985X.2012.01057.x>.
- Barro, R.J. (1991) Economic growth in a cross section of countries. **The Quarterly Journal of Economics**, 106 (2): 407–443. URL <http://qje.oxfordjournals.org/content/106/2/407.abstract>.
- Berners-Lee, T. (2016) Linked data. <https://www.w3.org/DesignIssues/LinkedData.html>.
- Besley, T. and Coate, S. (2003) Centralized versus decentralized provision of local public goods: a political economy approach. **Journal of Public Economics**, 87 (12): 2611 – 2637. URL <http://www.sciencedirect.com/science/article/pii/S004727270200141X>.
- Blankart, C.B. and Borck, R. (2005) **Handbook of Public Finance**. Boston, MA: Springer US. chap. Local Public Finance, pp. 441–476
- Brueckner, J.K. (1982) A test for allocative efficiency in the local public sector. **Journal of Public Economics**, 19 (3): 311–331. URL <https://ideas.repec.org/a/eee/pubeco/v19y1982i3p311-331.html>.
- Christen, P., Hegland, M., Roberts, S. *et al.* (2002) Parallel computing techniques for high-performance probabilistic record linkage. **Parallel computing techniques for high-performance probabilistic record linkage**.
- Clifford, D., Geyne-Rahme, F. and Mohan, J. (2010) How dependent is the third sector on public funding? Evidence from the National Survey of Third Sector Organisations Contents. **Third Sector Research Centre Working Paper**, 35.
- Clifford, D., Geyne-Rahme, F. and Mohan, J. (2013) Variations between organisations and localities in government funding of third-sector activity: Evidence from the national survey of third-sector organisations in england. **Urban Studies**, 50 (5): 959–976. URL <http://dx.doi.org/10.1177/0042098012458550>.
- Clifford, D. and Mohan, J. (2016) The sources of income of English and Welsh charities: An organisation-level perspective. **VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations**, 27 (1): 487–508. URL <http://dx.doi.org/10.1007/s11266-015-9628-5>.

- Duncan-Smith, I. (2005) Breaking the Big State Big Charity duopoly. Available from <http://www.iainduncansmith.org.uk/>.
- Einav, L. and Levin, J. (2013) **The Data Revolution and Economic Analysis**. University of Chicago Press. pp. 1–24. URL <http://www.nber.org/chapters/c12942>.
- Enamorado, T., Fifield, B. and Imai, K. (2017) Using a Probabilistic Model to Assist Merging of Large-scale Administrative Records. **Working paper**, available at: <http://imai.princeton.edu/research/files/linkage.pdf>.
- Fellegi, I.P. and Sunter, A.B. (1969) A theory for record linkage. **Journal of the American Statistical Association**, 64 (328): 1183–1210. URL <http://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10501049>.
- Gyourko, J. and Tracy, J. (1991) The structure of local public finance and the quality of life. **Journal of Political Economy**, 99 (4): 774–806. URL <https://ideas.repec.org/a/ucp/jpolec/v99y1991i4p774-806.html>.
- Jaro, M.A. (1995) Probabilistic linkage of large public health data files. **Statistics in Medicine**, 14 (5-7): 491–498. URL <http://dx.doi.org/10.1002/sim.4780140510>.
- Jin, L., Li, C. and Mehrotra, S. (2003) “Efficient record linkage in large data sets.” In **Eighth International Conference on Database Systems for Advanced Applications, 2003. (DASFAA 2003)**. Proceedings. pp. 137–146
- Kane, D. (2008) Classification of charities in England and Wales. **National Council for Voluntary Organizations**.
- Kane, D. and Clark, J. (2009) The regional distribution of charitable expenditure. **NCVO/VSSN Researching the Voluntary Sector Conference**, Warwick.
- Koudas, N., Sarawagi, S. and Srivastava, D. (2006) “Record linkage: Similarity measures and algorithms.” In **Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data**. SIGMOD ’06, New York, NY, USA: ACM. pp. 802–803. URL <http://doi.acm.org/10.1145/1142473.1142599>.
- Morgan, G.G. (2012) Public benefit and charitable status: Assessing a 20-year process of reforming the primary legal framework for voluntary activity in the UK. **Voluntary Sector Review**, 3 (1).
- NCVO (2016) **The UK Civil Society Almanac**. NCVO. URL <https://data.ncvo.org.uk/category/almanac/>.
- Newcombe, H.B., Kennedy, J.M., Axford, S.J. *et al.* (1959) Automatic linkage of vital records. **Science**, 130 (3381): 954–959. URL <http://science.sciencemag.org/content/130/3381/954>.

- Oates, W.E. (1969) The Effects of Property Taxes and Local Public Spending on Property Values: An Empirical Study of Tax Capitalization and the Tiebout Hypothesis. **Journal of Political Economy**, 77 (6): 957–71. URL <https://ideas.repec.org/a/ucp/jpolec/v77y1969i6p957-71.html>.
- OpenCorporates (2017) API Reference: version 0.4.6. Accessed: 2017-01-01. <https://api.opencorporates.com/documentation/API-Reference>.
- Osborne, S. and Hems, L. (1995) The economic structure of the charitable sector in the United Kingdom. **Nonprofit and Voluntary Sector Quarterly**, (24): 321–335
- Porter, E.H., Winkler, W.E., Census, B.O.T. *et al.* (1997) “Approximate string comparison and its effect on an advanced record linkage system.” In **Advanced Record Linkage System**. U.S. Bureau of the Census, **Research Report**. pp. 190–199
- Posnett, J. (1990) The resources of registered charities in England and Wales. **Researching the Voluntary Sector**, Charities Aid Foundation.
- Salamon, L.M. and Anheier, H.K. (1996) **The international classification of nonprofit organizations: ICNPO-Revision 1, 1996**.
- Schaltegger, C.A. and Torgler, B. (2004) Growth effects of public expenditure on the state and local level: Evidence from a sample of rich governments. CREMA Working Paper Series 2004-16, Center for Research in Economics, Management and the Arts (CREMA). URL <https://ideas.repec.org/p/cra/wpaper/2004-16.html>.
- Solé-Ollé, A. (2006) Expenditure spillovers and fiscal interactions: Empirical evidence from local governments in Spain. **Journal of Urban Economics**, 59 (1): 32 – 53. URL <http://www.sciencedirect.com/science/article/pii/S0094119005000604>.
- Svátek, V., Mynarz, J., Węcel, K. *et al.* (2014) **Linked Open Data – Creating Knowledge Out of Interlinked Data: Results of the LOD2 Project**. Cham: Springer International Publishing. chap. Linked Open Data for Public Procurement, pp. 196–213. URL [http://dx.doi.org/10.1007/978-3-319-09846-3\\_10](http://dx.doi.org/10.1007/978-3-319-09846-3_10).
- Taylor, L., Schroeder, R. and Meyer, E. (2014) Emerging practices and perspectives on big data analysis in economics: Bigger and better or more of the same? **Big Data and Society**, 1 (2).
- Tiebout, C.M. (1956) A pure theory of local expenditures. **Journal of Political Economy**, 64. URL <http://EconPapers.repec.org/RePEc:ucp:jpolec:v:64:y:1956:p:416>.
- Yinger, J. (1982) Capitalization and the theory of local public finance. **Journal of Political Economy**, 90 (5): 917–943. URL <http://www.jstor.org/stable/1837126>.