Public Procurement from the Private Sector

Charles Rahal
Department of Sociology and Nuffield College
University of Oxford

ADRN Conference
Queens University, 22nd June, 2018

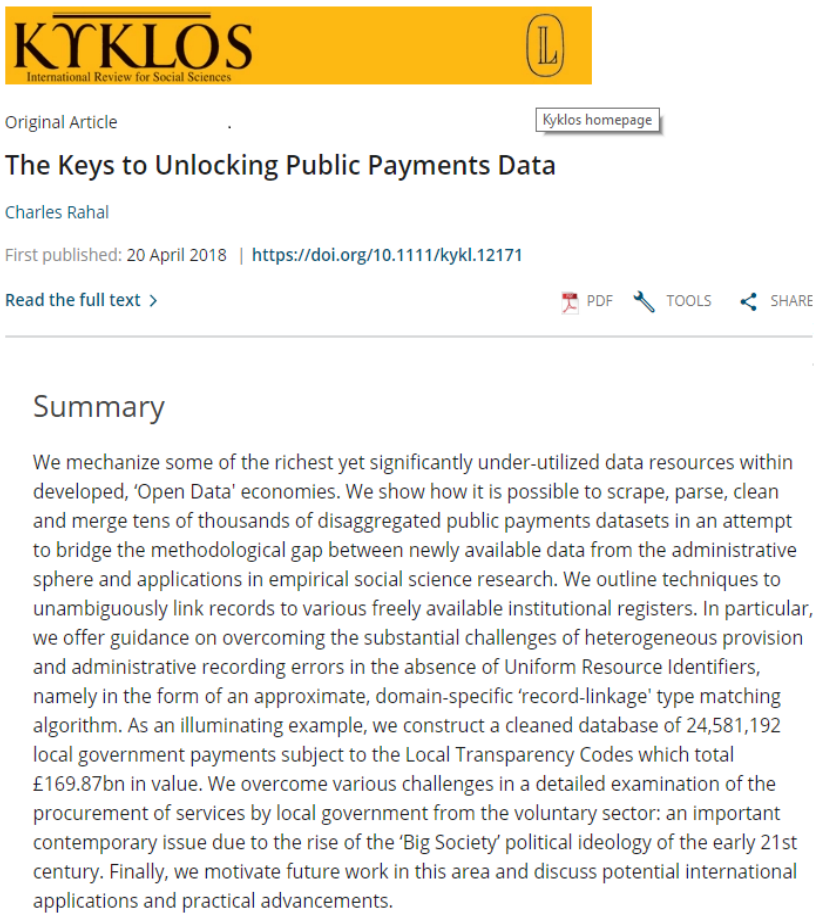
Open Source (MIT): https://github.com/crahal/centgovspend

Please note this is extremely preliminary work!

Introducting a Lesser Known Administrative Data Type

- Today we'll talk about transparency and public payments data ('Open Data').
- Coalition (31/05/2010) enforced accountability and aimed to reduce costs.
- Each department now has a Minister responsible for transparency issues.
- Range of measures: £10k contracts, gifts, gender pay gaps, hospitality, etc.
- We're going to focus on central government spending over £25k.
    - In particular: transaction level data across ministerial and non-ministerial bodies.
- This policy relates to all central govermnent including, NHS, NDPBs, etc.
- Datasource largely untapped: **huge** in magnitude – UK spends £800bn p.a.
- It would be great to know where that money goes, right?

Mostly NGO and With Limited Academic Contributions



KYKLOS
International Review for Social Sciences

Original Article

Kyklos homepage

The Keys to Unlocking Public Payments Data

Charles Rahal

First published: 20 April 2018 | https://doi.org/10.1111/kykl.12171

Read the full text > PDF TOOLS SHARE

Summary

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 of 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.

- Academic work limited to this one paper (?)
- Todays content builds on an earlier prototype which creates payments databases from LA's subject to the LATC.
- Most civic tech projects ran out of steam.
- OCP are an integral player internationally.
- Some collaboration with TI.
- Spend Network: excellent open data business with a pricing structure.
- This project aims to make software (`centgovspend`) to produce a bulk download for academic analysis possible.

Data

- Data origination is convoluted for multiple reasons:
  1. Fragmented: hosted on gov.uk, data.gov.uk *or* their department sites.
  2. Untimely: sporadically updated (at best) – *should be* within one month
  3. Heterogeous: despite guidelines, providers supply in different formats.
  4. Messy: requires substantial cleaning.
  5. User-input: requires verification (boo) or automated reconciliation (yay).
- In theory: 45 (non-)ministerial departments, 8 years, monthly = 4,320 files.
- Most commonly available in `.csv`, .xls, `.xlsx` ('3*' and '2*' files): none of the .pdfs ('1*') which plague NHS and LA data! This leads to a clean tool.
- Spending data involved made available under an Open Government License.

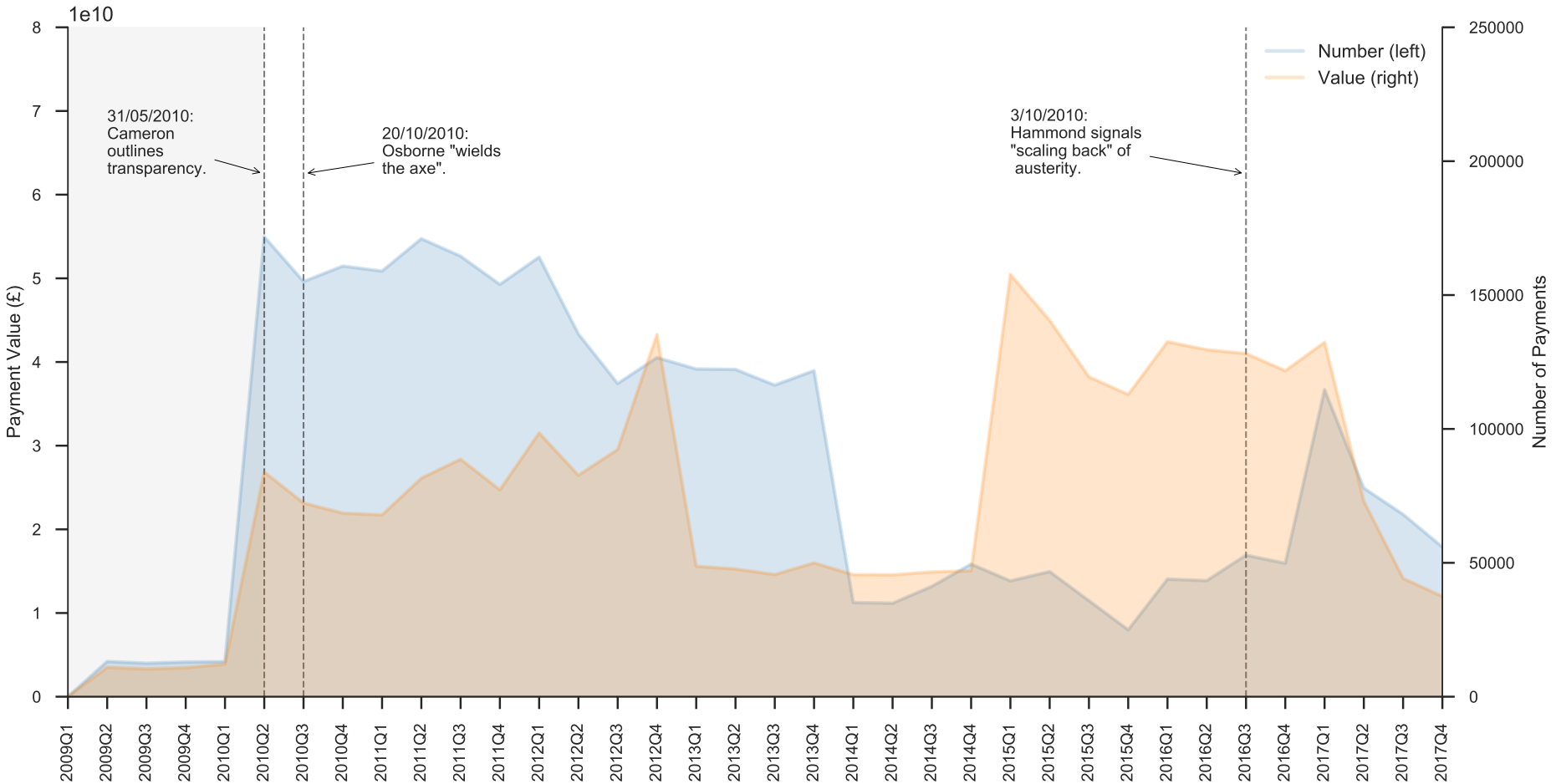
Introducing centgovspend

- `centgovspend`: a modular open source library available on GitHub.
- Updated quarterly (next: 01/07/2018) for new sources and edge cases.
- Different modules and functions...:
  1. Scrape all ministerial and non-ministerial locations for data.
  2. Clean files generally (e.g. dupes, blank rows) and harmonizes headers.
  3. Parse into a database of cleaned payments, retaining seven key fields.
  4. Remove problematic suppliers (e.g. 'redacted', 'various' and variants).
  5. Evaluates the scraping procedure.
  6. Moves onto reconciliation...

Evaluating the Scrape and Parse

```
** Evaluating the merged Dataset!**
** We have 3291682 total rows of data to begin.
We lose 11412 rows due to nulls in supplier\amount\bad dates.
Dropped 36361 redacted payments
Dropped redacted payments worth 201.08bn
We identified 247 redacted Suppliers
Dropped 14 "various" payments
Dropped "various" payments worth 50.05bn
We identified 1 "various" Suppliers
Dropping 309653 potential duplicates
** This spending totals 998.68bn.
** This is from a total of 2,719,026 payments.
** We merge from across 38 departments.
** This data comes from: 2179 files.
** We have: 60,597 unique supplier strings.
Cleaned file at: output\All_Merged_Unmatched.csv
```

Results vis a vis Policy timeline

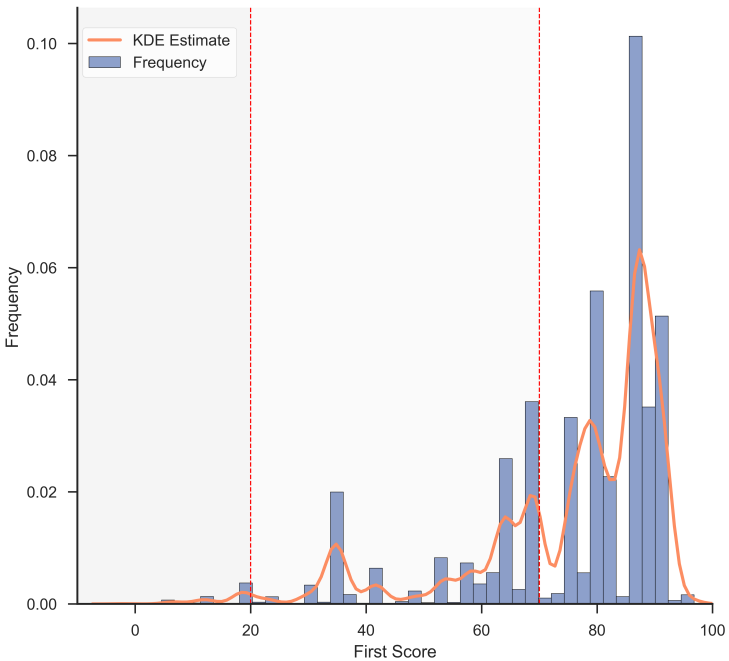


Methods: Reconciliation

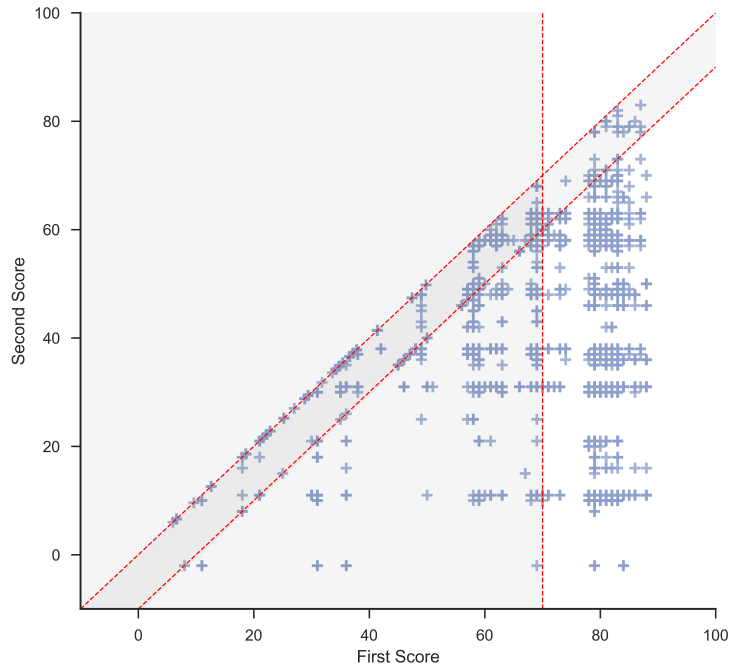
- Before analysis, we need an algorithm for reconciling supplier names.
- With a simple exact match strategy for groupings, how to know '23RED' and '23RED LTD' both relate to a '23RED LIMITED'?
- We design modules to reconcile via the OpenCorporates API
  - We write code to bypass OpenRefine and hit the OC REST API directly
  - OC API is based on a normalized supplier name with Elastic Search
- For each unique company reconciled, request from Companies House API:
  1. Basic company details.
  2. Information on all company officers.
  3. Information on persons of signifciant control (PSC).

Manual Verification and Automated Safematch Algorithms

(a) Manual Verification



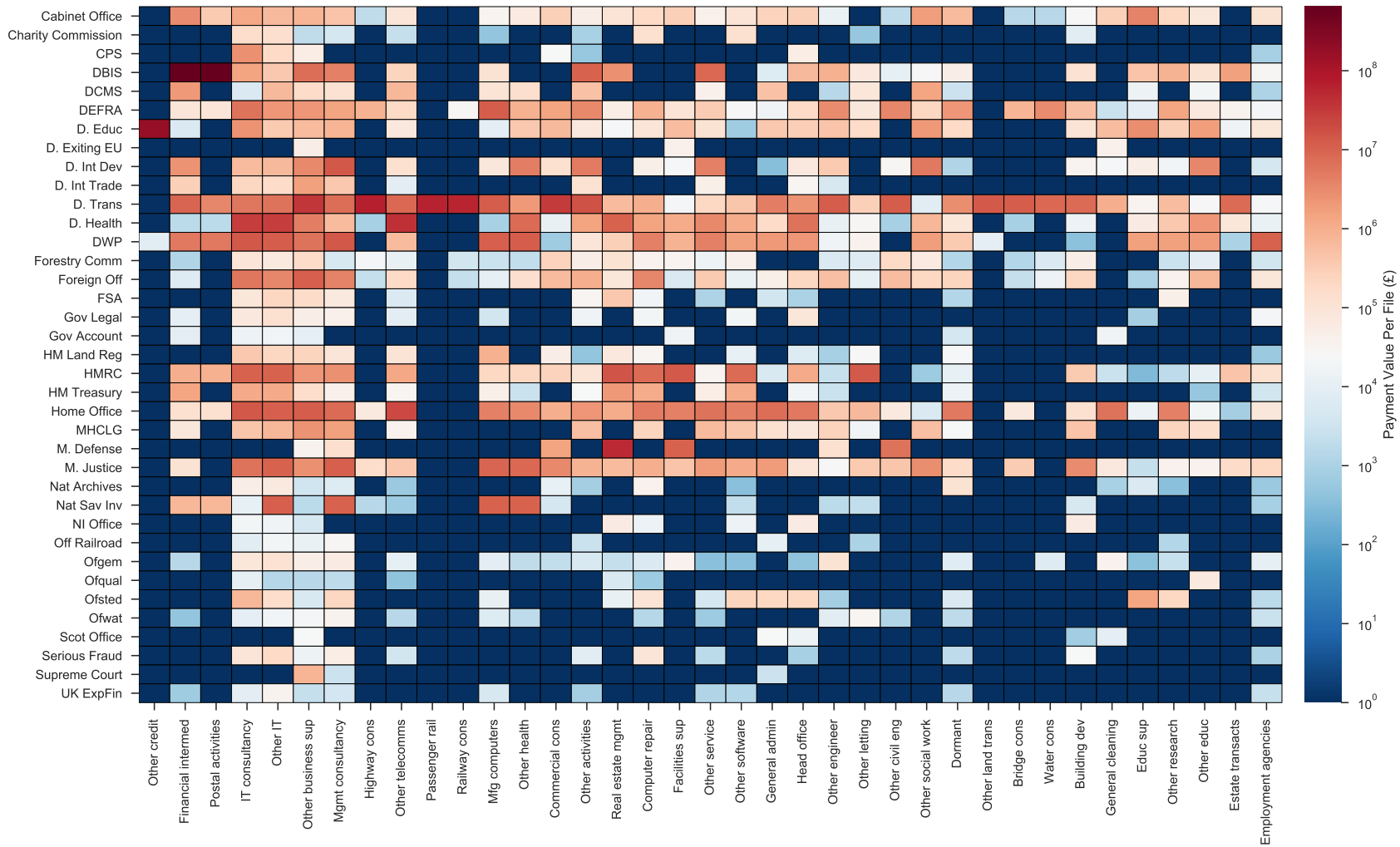
(b) Automated Safematch



We provide two functions for verifying the OC matches. From `automated safematch`:

```
We matched 1443852 out of 2719026 payments in total (53.1%).
We matched 11886981782 out of 998678528765 value in total 11%).
We matched 18963 out of 60597 unique suppliers in total (31.29%).
```

Aggregating Spend Across Standard Industry Classifiers



Highest Value Private Sector Suppliers (top 20 by value)

Best Match	Value (£bn)	Number	Postcode	SIC	Type
Student Loans Company	16.36	27	DL1 1RW	64929	ltd
British Broadcasting Corporation	11.98	178	N/A	N/A	Royal-charter
Post Office	9.73	1086	EC2Y 9AQ	Multiple	ltd
Crossrail Limited	4.72	27	E14 5LQ	42120	ltd
Atos It Services UK	2.37	10763	WC1V 6EA	Multiple	ltd
Airwave Solutions	2.13	2004	SW1E 5LB	61900	ltd
Connect Plus (M25)	2.01	2662	EN6 3NP	42110	ltd
Ibm United Kingdom	1.87	4663	PO6 3AU	26200	ltd
The Arts Council Of England	1.68	61	N/A	N/A	royal-charter
London & South Eastern Railway	1.61	242	NE1 6EE	49100	ltd
BT Group	1.57	732	EC1A 7AJ	61900	plc
Fujitsu Services	1.47	4468	W1U 3BW	Multiple	ltd
Balfour Beatty Civil Engineering	1.46	1807	E14 5HU	41201	ltd
Capita Business Services	1.31	26121	SW1H 0XA	62020	ltd
Northern Rail	1.31	377	RG27 9UY	82990	ltd
Csc Computer Sciences	1.28	1083	GU11 1PZ	Multiple	ltd
Mapeley Steps Contractor	1.15	2779	WD17 1HP	Multiple	ltd
EPSRC	1.13	61	N/A	N/A	royal-charter
High Speed Two (HS2)	0.98	140	B4 6GA	Multiple	plg-nsc
MRC	0.86	129	N/A	N/A	royal-charter
⋮	⋮	⋮	⋮	⋮	⋮

Grouped by Departments (top 20 by value)

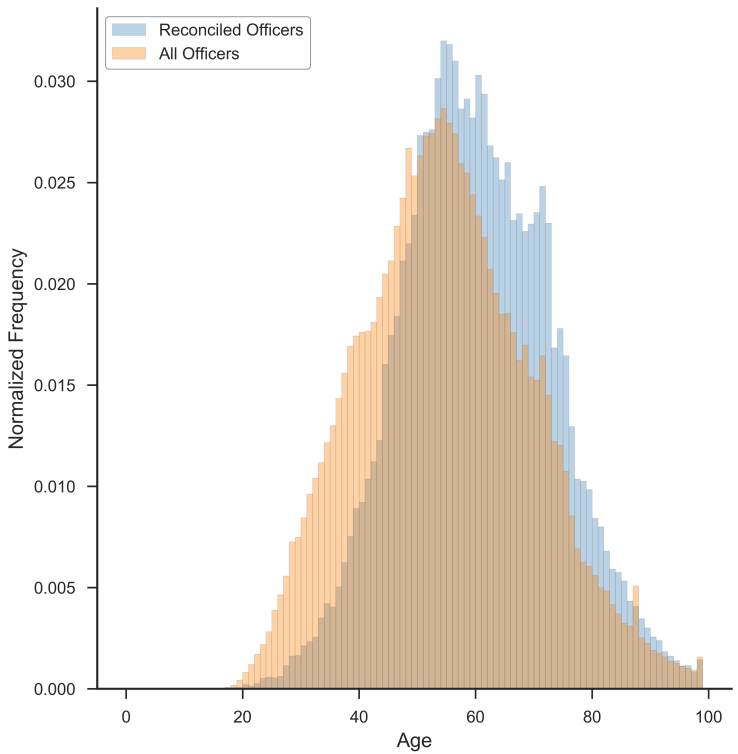
Department	Files	Value (£m)	Number	% to Prviate	Most Frequent PS Supplier
D. Health	47	417903	60873	1.4	EXPOTEL HOTEL RESERVATIONS LIMITED
D. Trans	84	149733	669722	20.62	INCHCAPE FLEET SOLUTIONS LIMITED
D. Educ	93	145867	89067	12.74	REDFERN TRAVEL LIMITED
Home Office	112	96104.4	113659	10.43	SPECIALIST COMPUTER CENTRES PLC
D. Int Dev	88	50030.8	144439	9.56	ADAM SMITH INTERNATIONAL LTD
DBIS	14	41388.9	27670	32.67	CAPITA BUSINESS SERVICES LTD
DWP	63	23509.7	1363930	27.14	XEROX (UK) LIMITED
DCMS	44	18271.4	4614	78.58	BRITISH BROADCASTING CORPORATION
DEFRA	62	18082.2	36603	19.34	IBM UNITED KINGDOM LIMITED
HMRC	95	12996.6	43359	28.41	TNT UK LIMITED
M. Justice	42	9396.63	24155	17.13	ATOS IT SERVICES UK LIMITED
Foreign Off	24	4685.37	16892	28.79	FCO SERVICES LTD
Cabinet Office	100	3308.6	14643	48.72	CAPITA BUSINESS SERVICES LTD
MHCLG	2	2210.53	8553	1.2	REDFERN TRAVEL LIMITED
Nat Sav Inv	93	1288.33	2532	83.84	ATOS IT SERVICES UK LIMITED
HM Treasury	85	1236.09	3973	40.13	FUJITSU SERVICES LIMITED
D. Int Trade	20	427.29	1966	35.74	GRANT THORNTON UK LLP
Forestry Comm	47	405.3	8865	32.86	EDF ENERGY 1 LIMITED
CPS	49	400.63	6924	38.45	CGI IT UK LIMITED
HM Land Reg	62	341.17	36201	70.13	CARILLION SERVICES LIMITED
⋮	⋮	⋮	⋮	⋮	⋮

A Social Stratification Style Application

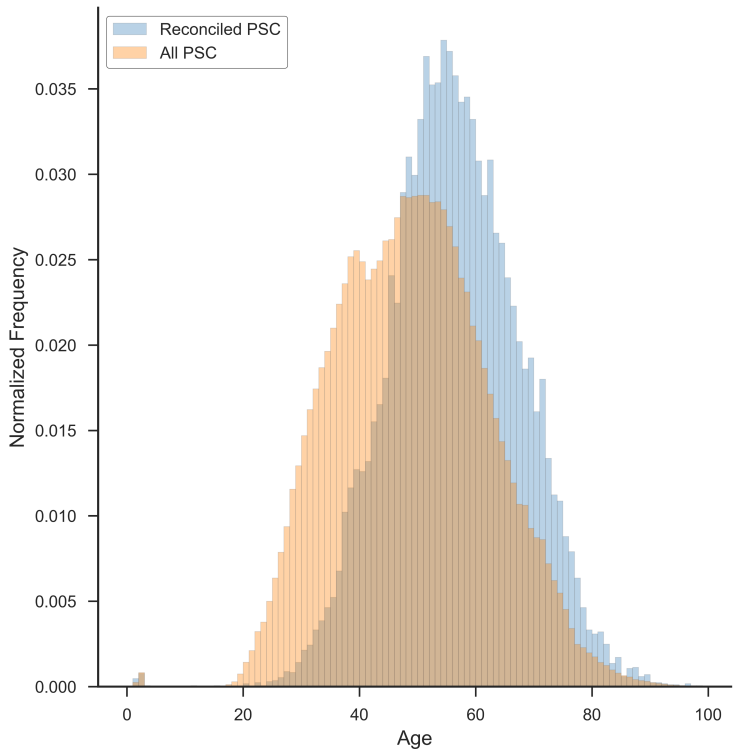
- This is where the talk about `centgovspend` ends - hopefully there will be multiple uses for it. At this point we have:
  1. Cleaned datbase of reonciled government spending.
  2. Database of these suppliers and a confident subset of private sector companies and their characteristics.
  3. A list of officers and PSC associated with all these companies.
- Moving to a **brief** example: a social stratification style application of officers and PSC which are supplying central government relating to:
  1. Age
  2. Gender
  3. Nationality and Country of Residence
  4. Occupation
- We can use the CH PSC flatfile, but had to set up a huge scrape of CH API to build an officers flatfile.

Age of Officers and Persons of Significant Control

(a) Officers



(b) Persons of Significant Control

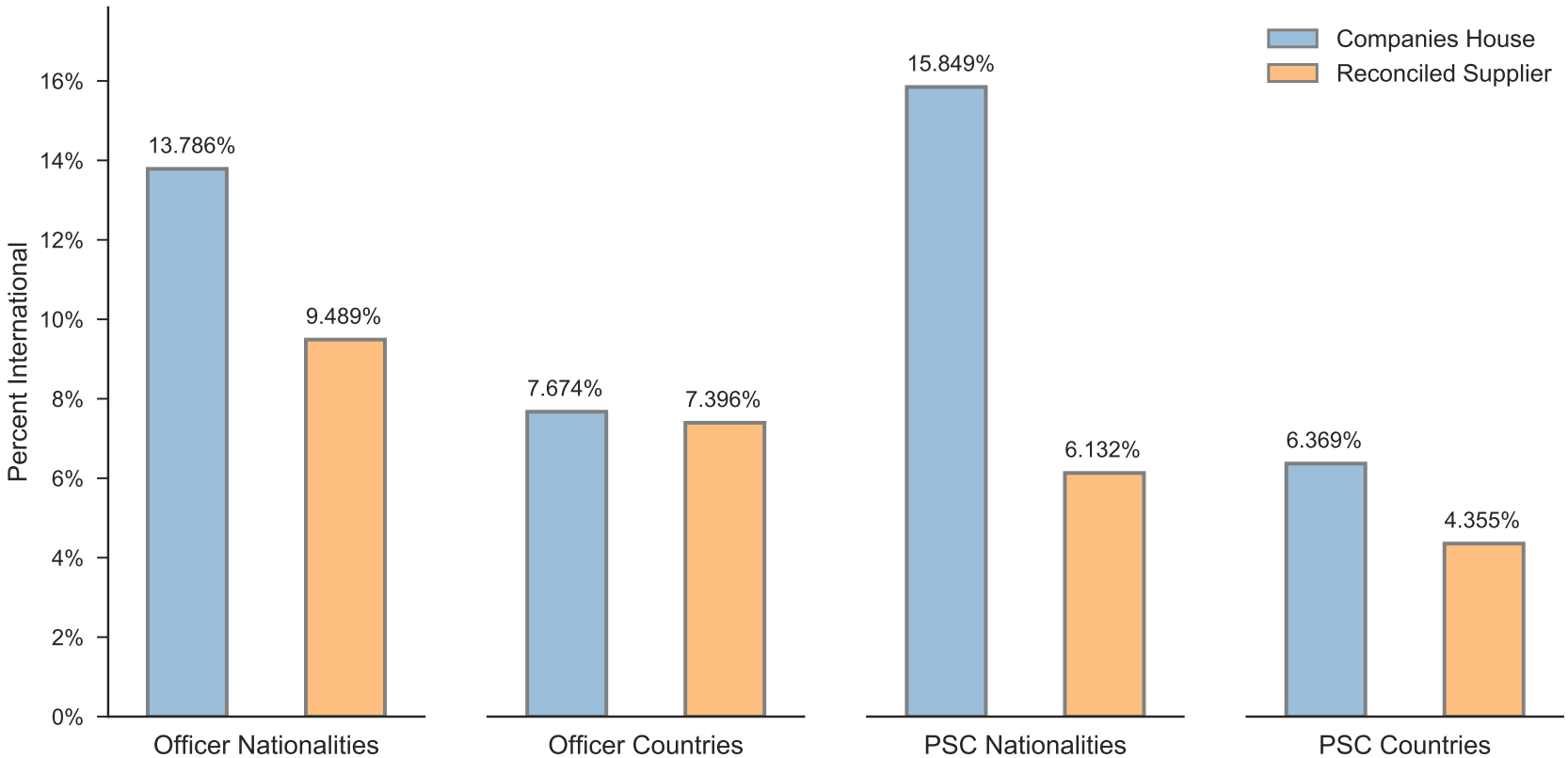


- Full CH Officers have an average age of 58: government supplier subset 64.
- CH PSC have an average age of 56: government supplier subset 62.

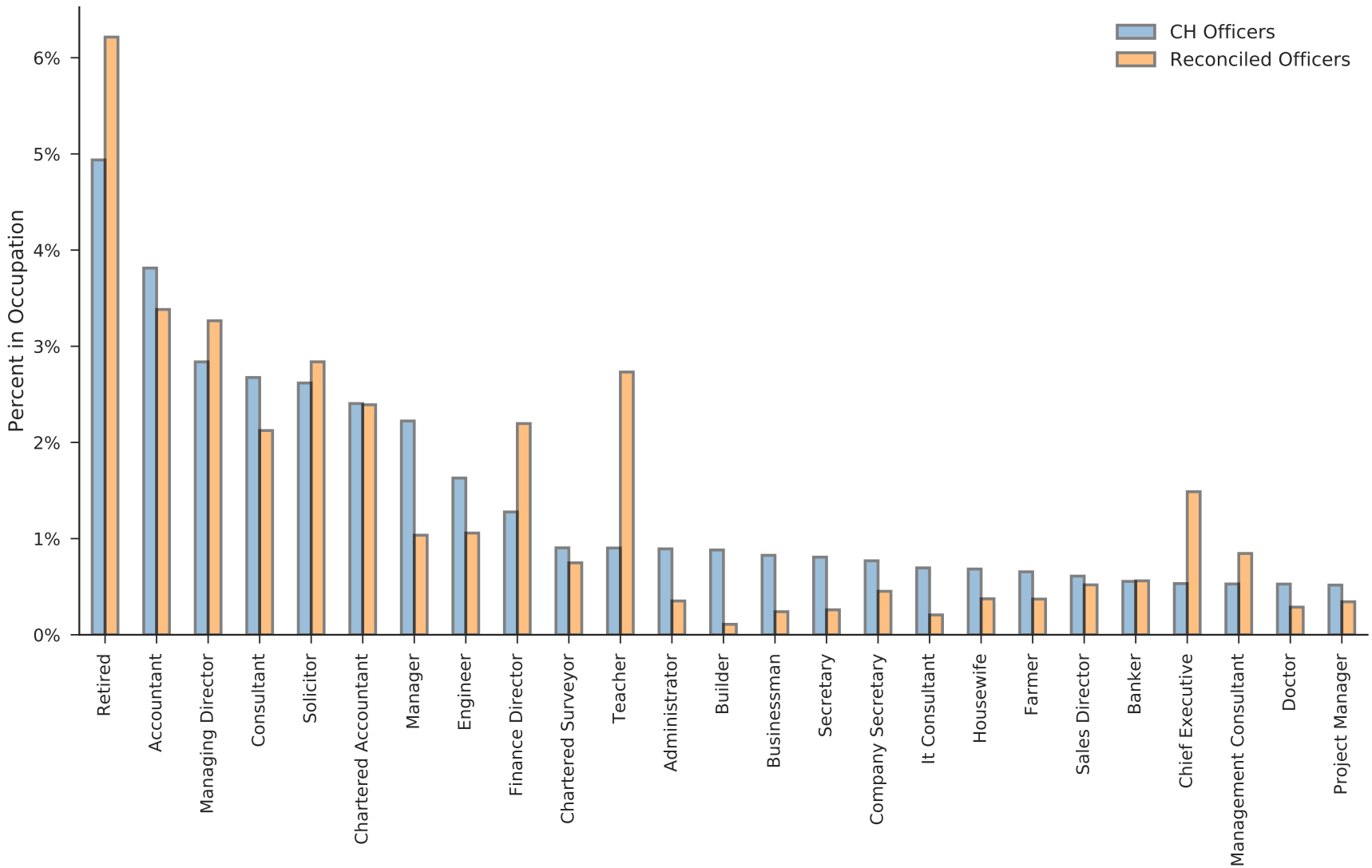
Gender

- We can also attempt to estimate gender ratios...
- for officers and PSC in whole of CH and subset of government suppliers.
- Uses a carefully prepared international name dictionary (`gender_guesser`).
- Generate forenames from splitting names on whitespace, ignoring titles.
- Drop any returns to androgynous names.
- Naively appropriate 'mostly_male/female' to 'male/female'.
- Gender ratios are fairly constant between CH and reconciled subset:
  - 29.5% female officers in entire CH, 28.1% in reconciled suppliers.
  - 27.8% female PSC in entire CH, 27.9% in reconciled suppliers.
- Perhaps most striking here is the global distance from 51% females.

Internationality of Officers and PSC



Occupation


CH Officers
Reconciled Officers
Percent in Occupation
0%
1%
2%
3%
4%
5%
6%
Retired
Accountant
Managing Director
Consultant
Solicitor
Chartered Accountant
Manager
Engineer
Finance Director
Chartered Surveyor
Teacher
Administrator
Builder
Businessman
Secretary
Company Secretary
It Consultant
Housewife
Farmer
Sales Director
Banker
Chief Executive
Management Consultant
Doctor
Project Manager

Conclusion

- This is **very** much a work in progress! Developments to come:
  1. Brush off JavaScript skills and build an online interactive dashboard.
  2. Write ML based algorthm for suggesting better, more consistent redactions.
  3. Develop elastic-search based API for reconciling other (e.g. NHS Digital) registers.
  4. Validate through merges with annual accounting budgets.
  5. Add a third – national – dimension of comparisons to officer & PSC characteristics.
  6. Tidy up the library, unit test, and turn it into a working paper.
- This is hopefully a valueable tool for academic researchers and is still in its inception – please don't hesitate to get in touch with comments, suggestions or requests for bulk downloads.
- Next up: `NHSspend` (for which this was largely a prototype of).