EXPLORING DIGITAL TECHNOLOGY INDUSTRY CLUSTERS USING ADMINISTRATIVE AND FRONTIER DATA

Max Nathan and Anna Rosso

Abstract

Industrial clusters help firms and workers become more productive, so are of great interest to researchers and policymakers. However, exploring and analysing such clusters is challenging. Big data and data-driven approaches can help. We study the emerging digital sector using big data obtained from administrative datasets but also from data science routines to develop modelled firm variables and firms' activities. These matched datasets allow us to have new insights on the importance of digital technology sectors in the UK, their structure and their co-location patterns.

Keywords: ICT, clusters, agglomeration, innovation, big data, data science

[3990 words including tables, footnotes and references]

1/ Introduction

The study of industrial clusters or milieux dates back to Alfred Marshall (1918). Such physical co-location helps firms and workers become more productive through a mix of 'matching', 'sharing' and 'learning' economies (Duranton and Puga 2014). Globalisation and new technologies appear to have reduced the salience of some of these forces while leaving others more important (Glaeser 2011, Moretti 2012)., For these reasons, industrial clusters are of great interest to researchers and to policymakers.

'Frontier' datasets and data science techniques (Feldman et al 2015) can contribute to the study of industrial clusters: defining emerging industries, improving co-location measurement, and potentially, identifying relationships between cluster protagonists. This chapter proposes an approach to studying industrial clusters that combines new data sources with high quality administrative microdata. We argue that this layered approach is a promising way forward. Specifically, we start with a novel dataset developed by the data science firm Growth Intelligence: this uses machine learning routines on company website content and media content to model firm characteristics and activity. We then match this dataset to high quality UK administrative firm level microdata. Combining administrative and frontier data in this way extends our view of firm activity beyond what is normally possible; importantly, it also provides a natural validation setting for the big data component.,

In particular, we show that using big data we have a more detailed and better measure of some of the most dynamic sectors in the economy, in this case, digital technology.

We build a firm-year panel that covers the financial years 1997 to 2013, with 1.29m firms and a total of more than 10m observations. The variables modelled by Growth Intelligence create alternative industry definitions, which we use to measure the importance and location of the digital-tech economy in the UK. We find that this sector is 11.5 percentage points larger than the official estimates, in terms of firms: they are more likely to be SMEs and to have at least one employee. In the co-location analysis we find that location quotients are greater outside London, while it is there that we find the highest number of startups and high jobs growth firms.

2/ Challenges for industrial clusters research

The idea of industrial clusters has its roots in Alfred Marshall's pioneering work on 'industrial districts' (Marshall 1918) and Jane Jacobs' analysis of ideas-driven urban economic change (Jacobs 1969), as well as a large body of empirical work (Scott 1988, Saxenian 1994, Storper 1997, Hall 2000). However, there is little agreement about defining clusters, or the usefulness of 'cluster policy' (Martin and Sunley 2003, Duranton 2011, Nathan and Overman 2013).

There are also real challenges in measuring and mapping clusters: many of these challenges may be amenable to big data and/or data-science driven solutions. First, studies tend to proxy co-location patterns using standard administrative units. However, such standardised spatial units may not capture actual co-location patterns well.[1] Microdata sets with fine-grained spatial identifiers offer the chance to work in a

[1] The Modifiable Unit Area Problem, or MUAP.

more detailed and flexible way (Duranton and Overman 2005). Second, and relatedly, to date researchers have worked with standardised industry codes. Even at a high level of detail such codes may not capture emerging economic activities of interest to policymakers. SICs are necessarily backward looking and lag real-world industrial and technological change (Nathan and Rosso 2015). New insights from big data can shed light on emerging industrial clusters that current SICs cannot see. Finally, in theory big data can help to study the 'map' of institutional component of clusters (universities, public-private partnerships, key firms and so on) and also the functional relationships within and centred on physical clusters (firm-firm linkages, for example).

3/ What big data can offer

'Big data' comes in three main flavours (Arribas-Bel 2014). These are: data from sensor networks and other sources 'in the wild'; corporate datasets, either internal business data or online sources (from search, social networks or company websites); and administrative datasets, especially microdata. These latter may be online and open; or available through resources such as the UK Data Service (UKDS).

To date, big data and data science techniques have been applied in a small number of cluster analyses. For example, Catini and colleagues (2015) develop a bibliometric approach to trace cluster boundaries, using the institutional address fields of researchers publishing in biomedical science journals. In their study of the computer games industry, Mateos-Garcia and Bakhshi (2014) use information from online

games directories, review sites and industry wikis to develop a detailed list of gaming firms and their locations, which they match to Companies House information. The 2016 Tech Nation report (Tech City UK and NESTA 2016) develops a multi-angle take on the tech economy, using Growth Intelligence and Companies House data alongside a number of other commercial, unstructured sources including online job ads and meetups. Bernini et al (2016) use a combination of commercial companies data from FAME and company website information to develop alternative estimates and location patterns for the digital health, finance and processing sectors in England.

Very few studies have attempted to get a handle on relational aspects of industry clusters. (Williams and Currid-Halkett (2014) use Foursquare data to track the physical movements of fashion designers in Manhattan during a two-week period, with hourly 'check-in' data recording visits frequency and type. The London and Cambridge Tech Maps[2] use live Twitter data to show mentions and retweets of local firms. Mateos-Garcia and Bakhshi (ibid) highlight a number of suggestive relational and institutional findings for computer games hubs.

3.1 / Pros and cons

'Big data' is generally defined in terms of the Four V's: volume (massive datasets, with millions or billions of observations); velocity (data which may be available at real time or close to it) and variety (a wide range of sources which help us observe, or model, phenomena previously hard to observe). The fourth V is veracity – which throws up a number of analytical challenges for those using many frontier datasets,

[2] http://www.techcitymap.com, http://www.camclustermap.com (accessed 13 April 2016).

especially data from the web, from social networks or from internal corporate sources (Einav and Levin 2014).

These challenges include dealing with raw data that is often unstructured, and may need substantial cleaning. Many commercial datasets have an unclear sampling frame (for example, web-scraped data will miss firms without websites, or who have non-scrapable sites). Metadata is either minimal or non-existent. In many cases, then, working with such datasets requires substantial additional research time for cleaning, testing and understanding – something not typically required with 'conventional', structured resources. For these reasons, combining administrative and commercial frontier datasets can be a promising way forward.

The four V's framework is useful to describe the dataset that we use. GI's data is based on Companies House, an open register of over 3m companies active in the UK. The additional data GI develop are available in close to real-time, being built on raw information from online content as well as existing Companies House variables. However, because not all companies have websites or scrapable websites, and because GI develops modelled variables, there are a range of veracity issues (see Nathan and Rosso 2015) for a discussion.).

4/ Our approach

Over the past couple of years we have been working with a very large panel dataset that combines open administrative data from the UK with raw and modelled,

machine-learnt variables developed by the data science firm Growth Intelligence. We use this layered approach because each data layer adds richness and detail. For example, the administrative data provides postcode level location information, which is unavailable in the modelled data; conversely, the modelled data offers significantly more up to date product and sector information than the administrative dataset. The administrative layers also allow for effective validation of the more experimental layers, as we explain below. The dataset allows new insight on emerging industries and product sets, especially in the 'digital economy', as well as allowing for very detailed co-location analysis.

The 'base' layer of our dataset is the Business Structure Database (BSD), which provides plant and firm-level information for 99% of UK firms, including age, sector, location, employment, and revenue, in a series of linked cross-sections (Office of National Statistics 2016). Specifically, firms are included if their annual turnover is high enough to incur UK sales tax, they have at least one employee, or both. Firms enter the BSD when these conditions are met, and temporarily leave the BSD if neither of these conditions hold. The other administrative layer is company-level information from Companies House (CH), a UK-wide government agency that includes company formation, financials and corporate structure information for all publicly listed companies.[3] Our CH data includes all companies active in the UK as of August 2012.

Companies House data is now fully open and accessible through an API (Application Programming Interface). Growth Intelligence (GI) use this functionality to develop a

[3] www.companieshouse.gov.uk, accessed 24 May 2016. Note that Companies House does not cover sole traders or certain forms of partnership, where the partners are all registered as individually self-employed.

close-to-comprehensive database of UK company information. They combine this with data from a range of sources, including text scraped from company websites, news media sources, industry forums and professional networking platforms. GI then uses a range of data science routines (in particular, feature extraction and supervised learning) to develop modelled variables including company sector; principle product and customer type; and lifecycle events such as new product launches, mergers/acquisitions and joint ventures. See Nathan and Rosso (2015) for detail.

4.1 / Build

We build a firm-year panel which covers the financial years 1997 to 2013 in its current form, and contains over 10m observations for 1.26m firms. The construction process is complex due to data access restrictions, as well as the underlying challenges in matching companies (legal structures) to real-world firms.

First, BSD data is only available through the UK Data Service Secure Lab, and is anonymised. We pre-clean and anonymise Companies House data, removing inactive and dormant companies, as well as using shareholder and reported revenue information to control for company group structure (see Nathan and Rosso, (2015) for details). This data is then matched to the BSD by the Secure Lab staff: for the 2013 BSD, the raw matching rate is 61.1%, or 75.7% against our pre-cleaned data.[4] Second, we run further cleaning routines, removing public sector organisations and those firms who left the BSD pre-August 2012. We then run further cleaning routines to simplify company group structure, for around 1.6% of observations where there is a 1:n

[4] Each year of BSD data represents a financial year, so the 2013 BSD cross-section covers the period April 2012-April 2013. This is the best fit for our CH data, which is a grab from August 2012.

company:firm match. Because our data are anonymised, we use heuristics based on company formation year, reported revenue and revenue levels to keep the most established and highest-revenue reporting companies in a linked group.[5] We then shuffle out a very small number of remaining duplicates.[6]

The final panel has two important features. First, it is unbalanced, as firms enter at different times. We interpolate observations for the small number of firms that leave the dataset temporarily due to not meeting turnover or employment thresholds. Second, the panel contains all and only those firms active as of August 2012. On the one hand, we do not see firms who enter after August 2012; conversely, we do not observe firms who entered and died before this date. In its current state, then, the panel is weighted towards newer businesses, and towards older 'survivors'.

5/ Descriptive analysis

5.1/ Re-framing 'digital tech' firms

We use GI's modelled sector and product variables to generate alternative definitions of digital technology firms. We compare these to 'official' estimates for firms with

[5] Specifically, these firms are older than average (mean incorporation year is 1990 vs 2002); enter the BSD earlier (1984 vs 2001); have a lot more plants (94 vs 6); have much higher employment (3096 vs 187) and employees (3095 vs 187); have much higher annual turnover (£1,200,313 vs £70,983); are more likely to file revenue to Companies House; and report higher 2010-2013 revenue to Companies House (average £12.4bn vs 2£.53bn). We use these characteristics to inform the heuristics we develop.

[6] As a sensitivity check we then correlate the characteristics of the retained observations against the modal values of group of linked companies. We find a 0.67 correlation between the incorporation years; a 0.86 correlation between GI sectors; a 0.86 correlation between GI products; and a 0.82 correlation between SIC5 codes. All but one of these is significant at 1%. Overall, we conclude that our cleaning rules do not systematically misrepresent underlying corporate structure.

official 'digital technology' SIC codes (Harris 2015). Specifically, we want to improve on SIC-based counts by removing false positives: firms in ‘digital’ industries that do not produce relevant products / services (for example, excluding mobile phone shops from a hypothetical ‘mobile telephony’ industry). Conversely, we also want to remove false negatives: firms in ‘non-digital’ sectors who offer products / services built on ICT, software or related digital content. To do this, we improve the mapping technique developed in Nathan and Rosso (2015).

First, for firms with ONS ‘digital technologies’ SICs, we extract the corresponding 38 GI product and 134 sector categories. Second, we cut off ‘sparse’ products and sectors which account for less than 0.2% of all observations, leaving us with 17 products and 31 sectors.[7] Third, we recover ‘sparse but relevant’ GI products, and drop ‘irrelevant’ products by comparing GI descriptors with OECD-UN descriptors developed for a recent digital economy mapping exercise (OECD 2011). For example, we recover the product category ‘peer to peer communications’, but drop ‘printing services’. This gives us a core list of 16 digital technology products and services. Fourth, we denote any GI sector as ‘digital’ if over 50% of the firms in that sector produce some digital product or service. The idea here is to provide a more natural representation of where digital tech products/services are being generated across the economy (e.g. financial services, medicine, or engineering). This gives us 26 sector categories. Finally, we generate product times sector dummies equal to one if a firm is in both a GI digital sector and whose principal product / service is a GI-digital product.[8]

[7] In Nathan and Rosso (2015) we experiment with variations on this threshold rule without materially changing our results.
[8] We disallow the non-relevant GI product-sector cell consultancy*management consultancy. Results on different steps upon request.

5.2 / Counts and shares

Table 1 shows counts and shares of digital tech firms in the panel, defined first by ONS 'digital technology' SICs and then by Growth Intelligence sector-product categories. As defined by GI's data science-driven approach, the digital tech sector is substantially larger that what official definition using SIC codes would imply: 19.29% vs 7.7% of firms. We also repeat the analysis for sub-samples of firms: digital tech firms are more likely to be SMEs and to have at least one employee.[9]

Table 1 about here

5.3 / Internal structure

Table 2 sets out the 'digital' GI products we identify in section 6.1.

Table 2 about here

Table 3 decomposes firms in the GI digital technology set by GI sector category. As expected, we see a mixture of conventional 'ICT' sectors (information technology, telecoms, computer software), plus a number of industries not historically seen as tech, but where our data indicates extensive adoption of a digital product set, such as financial services, healthcare and mechanical/industrial engineering. This aligns with qualitative evidence on the growth of fintech and digital health spaces, as well as the increasing role of software and automation in advanced engineering.

[9] Tables with GI sector and product typology breakdown can be provided upon request.

Table 3 about here

5.4/ Co-location patterns

Tables 4 and 5 show co-location patterns at respectively, Travel to Work Area level (the UK has 243 TTWAs, roughly corresponding to labour markets) and postcode district level (the UK has about 3,000 of these, and 150 in London alone). For concision we restrict the results to the 'top 10' locations in each case (full results available on request). In each table, the left hand column shows location quotients for digital technology firms; the central column shows counts of start-ups (defined as firms that are three years old or less; the right hand column shows counts of employment gazelles, defined as above using the OECD definition.

Table 4 about here

The TTWA-level analysis shows two striking findings. First, location quotients for digital technology firms are greatest outside London as a whole, a result presumably driven by the capital's economic diversity in comparison to other metros. Second, raw counts of startups and high jobs growth firms are much higher in London than in the rest of the UK.

Table 5 about here

Table 5 shows the postcode district level results, which adds further richness to the picture. We can see that, as with the TTWA level analysis, digital technology firms are most densely co-located outside London, although many local-level clustering takes place in the Greater South East; the top 10 PCDs are WA1 (Warrington); RG6 (Reading /Wokingham); MK4 / MK5 / MK8 (Milton Keynes); TS20 (Stockton on Tees / Middlesbrough); SN5 (Swindon); AB22 (Aberdeen); and GU22 (Guildford). By contrast, startup counts and counts of high-growth firms are highest in London, and tend to be concentrated in inner London locations such as Canary Wharf, Islington, Hoxton, Spitalfields, the City of London and Camden. The only non-London locations to feature in these top 10s are IG1 (Redbridge); TW3 (Hounslow); RG1 (Reading), CR0 (Croydon) and BN1 (Brighton).

6/ Conclusions

In this chapter we develop a new, data-driven approach to studying industrial clusters. Specifically, we match administrative firm-level data for the UK with a novel dataset developed by a data science company, which exploits content found on company websites and in media sources.

We use this information to measure the location and importance of the digital technology firms and clusters in the UK. Our estimates show that the digital technology sector is much larger than the estimates obtained SIC codes; digital technology companies are spread out in most of the country, with a large proportion of startups and high growth firms located in London. Our analysis highlights the fact

that frontier datasets, used in combination with administrative data, provides new and high quality findings about industrial clusters. In turn, this can be used both to deal with some challenges of cluster research, and to inform public policy.

REFERENCES

Arribas-Bel, D. (2014). "Accidental, open and everywhere: Emerging data sources for the understanding of cities." Applied Geography **49**: 45-53.

Bernini, M., M. Barbera, S. Addison, R. Mulhall, M. Nathan, P. Ramirez and N. Sambin (2016). Industrial Clusters in England. Report for BIS. London, NIESR.

Catini, R., D. Karamshuk, O. Penner and M. Riccaboni (2015). "Identifying geographic clusters: A network analytic approach." Research Policy **44**(9): 1749-1762.

Duranton, G. (2011). "California Dreamin': The feeble case for cluster policies." Review of Economic Analysis **3**(1): 3-45.

Duranton, G. and H. G. Overman (2005). "Testing for Localization Using Micro-Geographic Data." The Review of Economic Studies **72**(4): 1077-1106.

Duranton, G. and D. Puga (2014). The Growth of Cities. Handbook of Economic Growth. **2:** 781-853.

Glaeser, E. (2011). The Triumph of the City. London, Pan Macmillan.

Hall, P. (2000). "Creative Cities and Economic Development." Urban Studies **37**(4): 639-649.

Harris, J. (2015). Identifying Science and Technology Businesses in Official Statistics. London, ONS.

Jacobs, J. (1969). The Economy of Cities. London, Vintage.

Marshall, A. (1918). Principles of Economics. New York, Macmillan.

Martin, R. and P. Sunley (2003). "Deconstructing clusters: chaotic concept or policy panacea?" Journal of Economic Geography **3**(1): 5-35.

Mateos-Garcia, J., H. Bakhshi and M. Lenel (2014). A Map of the UK Games Industry. London, NESTA.

Moretti, E. (2012). The New Geography of Jobs. Boston, Haughton Mifflin Harcourt.

Nathan, M. and H. Overman (2013). "Agglomeration, clusters, and industrial policy." Oxford Review of Economic Policy **29**(2): 383-404.

Nathan, M. and A. Rosso (2015). "Mapping digital businesses with Big Data: some early findings from the UK " Research Policy **44**(9): 1714-1733.

OECD (2011). OECD Guide to Measuring the Information Society 2011. Paris, OECD.

Office of National Statistics (2016). Business Structure Database, 1997-2015 SN: 6697. Secure Data Service Access [computer file]. Colchester, UK Data Archive.

Saxenian, A.-L. (1994). Regional Advantage: Culture and Competition in Silicon Valley and Route 128 Cambridge, MA, Harvard University Press.

Scott, A. (1988). New industrial spaces: Flexible production organization and regional development in North America and Western Europe. London, Pion.

Storper, M. (1997). The Regional World: Territorial Development in a Global Economy. New York, Guilford.

Tech City UK and NESTA (2016). Tech Nation 2016: Transforming UK industries. London, TCUK.

TABLES

Table 1. Digital technology firms: counts and shares, 1997-2013.

Category	Freq.	Percent
Other	9,362,261	92.45
Digital Tech SIC	764,110	7.55
Other	8,173,448	80.71
Digital Tech GI	1,952,923	19.29
Total	*10,126,371*	*100*

Source: BSD / Companies House / Growth Intelligence.

Table 2. Product breakdown for GI digital technology firms.

GI product	Freq.	Percent	Cumulative
consultancy	1,276,928	65.39	65.39
care_or_maintenance	306,763	15.71	81.09
electronics	145,160	7.43	88.53
custom_software_development	113,763	5.83	94.35
broadband_services	38,672	1.98	96.33
web_hosting	28,808	1.48	97.81
software_desktop_or_server	21,096	1.08	98.89
advertising_network	15,638	0.8	99.69
peer_to_peer_communications	5,204	0.27	99.95
software_web_application	331	0.02	99.97
digital_media	305	0.02	99.99
software_mobile_application	255	0.01	100
Total	*1,952,923*	*100*	

Source: BSD / Companies House / Growth Intelligence.

Table 3. Sectoral breakdown for GI digital technology firms.

GI sector	Freq.	Percent	Cumulative
information_technology	477,939	24.47	24.47
mechanical_or_industrial_engineering	173,903	8.9	33.38
financial_services	161,377	8.26	41.64
accounting	155,112	7.94	49.58
biotechnology_greentech	149,561	7.66	57.24
hospital_and_health_care	144,391	7.39	64.64
electrical_electronic_manufacturing	142,143	7.28	71.91
computer_software	131,989	6.76	78.67
telecommunications	87,876	4.5	83.17
design	65,921	3.38	86.55
marketing_advertising	56,997	2.92	89.47
security_and_investigations	48,295	2.47	91.94
medical_practice	42,637	2.18	94.12
computer_networking	27,636	1.42	95.54
internet	24,726	1.27	96.8
computer_hardware	22,097	1.13	97.94
consumer_electronics	14,098	0.72	98.66
computer_games	10,100	0.52	99.17
information_services	5,276	0.27	99.44
industrial_automation	5,146	0.26	99.71
computer_network_security	2,113	0.11	99.82
semiconductors	2,001	0.1	99.92
e_learning	951	0.05	99.97
wireless	352	0.02	99.99
management_consulting	169	0.01	99.99
nanotechnology	117	0.01	100
Total	*1,952,923*	*100*	

Source: BSD / Companies House / Growth Intelligence.

Table 4. TTWAs location patterns, 2013. Top 10s.

TTWA	LQ	TTWA	Start-ups	TTWA	Gazelle firms
Basingstoke	1.56	London	16801	London	1936
Reading & Bracknell	1.49	Manchester	1852	Manchester	286
Warrington & Wigan	1.478	Guildford & Aldershot	1335	Birmingham	187
Guildford & Aldershot	1.331	Birmingham	1270	Reading & Bracknell	139
Newbury	1.31	Reading & Bracknell	1097	Guildford & Aldershot	139
Aberdeen	1.255	Luton & Watford	1047	Leeds	138
Milton Keynes & Aylesbury	1.251	Wycombe & Slough	1000	Bristol	129
Middlesbrough & Stockton	1.244	Bristol	898	Wycombe & Slough	115
Wycombe & Slough	1.232	Crawley	760	Luton & Watford	103
Luton & Watford	1.231	Glasgow	751	Glasgow	100

Source: BSD / Companies House / Growth Intelligence.

Table 5. Postcode district colocation patterns, 2013. Top 10s.

Postcode district	LQ	Postcode district	Start-ups	Postcode district	Gazelle firms
WA1	2.782	E14	323	SE1	96
RG6	1.993	SW19	291	EC2A	44
MK4	1.889	CR0	264	N1	41
TS20	1.887	SE1	240	E1	39
LL29	1.883	N1	215	W1W	37
SN5	1.876	SW6	190	E14	36
MK5	1.834	SW15	178	W1T	33
MK8	1.823	IG1	177	BN1	31
AB22	1.768	TW3	177	EC1V	30
GU22	1.763	RG1	167	NW1	28

Source: BSD / Companies House / Growth Intelligence.