Applied Network Science

RESEARCH Open Access

The nested structure of urban business clusters

Clémentine Cottineau[1,2*] and Elsa Arcaute[1]

*Correspondence:
clementine.cottineau@ens.fr
[1]Centre for Advanced Spatial Analysis, University College London, 90 Tottenham Court Road, London, W1T 4TJ, UK
[2]Centre National de la Recherche Scientifique (CNRS), UMR 8097 Centre Maurice Halbwachs, 48 Boulevard Jourdan, 75014 Paris, France

Abstract

Although the cluster theory literature is bountiful in economics and regional science, there is still a lack of understanding of how the geographical scales of analysis (neighbourhood, city, region) relate to one another and impact the observed phenomenon, and to which extent the clusters are industrially coherent or geographically consistent. In this paper, we cluster spatial economic activities through a multi-scalar approach making use of percolation theory. We consider both the industrial similarity and the geographical proximity between firms, through their joint probability function which is constructed as a copula. This gives rise to an emergent nested hierarchy of geoindustrial clusters, which enables us to analyse the relationships between the different scales, and specific industrial sectors. Using longitudinal business microdata from the Office for National Statistics, we look at the evolution of clusters which spans from very local groups of businesses to the metropolitan level, in 2007 and in 2014, so that the changes stemming from the financial crisis can be observed.

Keywords: Geoindustrial clusters, Multi-scalar analysis, Business, Greater London, Microdata, Percolation theory

Introduction

According to (Malmberg and Maskell (2002), p.430-1), "*there are several reasons to take the issue of spatial clusters seriously. One is that spatial clustering is at the very core of what research in economic geography is all about. [...] There is a lot to learn about the role of proximity and place in economic processes by trying to pinpoint the driving forces that make for the agglomeration in space of similar and related economic activities [...] Second, this task has obvious policy relevance today*". Interestingly though, the economic drivers and rationale behind the clustering of businesses might be at odds with the policy incentives to promote particular locations for institutionalised clusters. For example, the eastern Fringe of the City in London has witnessed the rapid clustering of start-ups and businesses from the 'digital creative', tech and advertisement industries[1] in the aftermath of the 2008 crisis, around Shoreditch and Old Street (Foord 2013). However, from the moment the digital cluster was recognised, labelled and institutionalised as 'Tech City' by local actors and eventually by the government (in 2011), the hype and investments by big players of the sector (Google, Cisco, Vodafone) contributed to push away the endogenous small actors of the cluster, who started relocating to (cheaper) neighbouring locations (Nathan and Vandore 2014), following the spatial development of key amenities such as

[1]A set of industries also identified as the "flat white economy" (McWilliams 2015).

semi-public spaces and a diverse mix of building types and empty sites (Martins 2015). Moreover, if the *"current vitality emerges from the risky experimentation across co-located sectors in which hitherto unrelated knowledge and activities (for example, software and advertising) are being combined"* (Foord (2013), p.52), it suggests that any successful sectoral combination at present might not be so successful in the future, which instead should benefit newer risky combinations. This highlights the need for a better understanding of the inner (industrial and spatial) dynamics of clusters and the overarching organisation of urban economies driving individual firms' relocation strategies, for analytic purposes as well as for policy efficiency.

Identifying clusters and drawing cluster policies have become mainstream since the influential contribution of Michael Porter in the 1990s (Porter 1998). Nevertheless, there is no unique way to define a cluster, and the fuzziness of its original definition has made it "confusing" (Martin and Sunley 2003). Within the literature, the term is used to refer to very local phenomena (e.g. eastern Fringe of the City in London) as well as their regional counterparts (e.g. the South-East of the UK, which includes Greater London and the surrounding local authorities). On the one hand, there is either no consensual way to define clusters, and on the other, the methods employed might contain hidden assumptions. Our contribution thus aims at rendering explicit and transparent the cluster delineation process, but also at introducing other tools outside the traditional ones, such as percolation[2] and network theory. Within the framework of the economic geography literature, one can identify two recurrent elements referring to the definition of clusters listed below.

The first one refers to considering clusters as networks of inter-dependent firms and industries. (Iammarino and McCann (2016), p.1023) summarize this idea by stating that *"industrial clusters are distinguished in terms of the nature of firms in the clusters and of their relations and transactions undertaken within clusters"*. In more classical definitions, we find similar descriptions of networks of firms. For example, (Porter (1998), p.199) mentions *interconnected companies* and associated institutions as the core of clusters. (Rosenfeld (1997), p.4) talks about the *interdependence* of firms, (Feser (1998), p.26) and (Swann et al. (1998), p.139) talk about their *relatedness*. (Simmie and Sennett (1999), p.51) insist on service companies being *interconnected*, while (Roelandt et al. (1999), p.9) and (Van den Berg et al. (2001), p.187) use the figure of the *network* to define clusters, even though they refer to producing firms in the first case (Roelandt et al. 1999) and to specialised organisations in the second case (Van den Berg et al. 2001). All in all, the element of networked firms of similar or interrelated industries is a constant of most definitions of industrial clusters.

The second broad element that we find in most definitions of clusters is a spatial reference. However, the concrete specification of this spatial reference is all but precise and homogeneous across authors. For example, some definitions of clusters mention *geographical proximity* of the connected firms (Porter 1998; Rosenfeld 1997; Enright 1996), or the fact that they are *closely located* (Crouch and Farrell (2001), p.163). (Swann et al. (1998), p.139) define clusters as "a large group of firms in related industries, *at a particular*

[2]Percolation theory studies the propagation of a phenomenon in a medium, say the spread of a fire in a forest or of oil in a porous rock. It looks at the probability p that neighbouring sites are connected. At a specific value of p, there is a *critical* transition at which the phenomenon has spread say, *from top to bottom*. Studying the phenomenon at different values of p, can be seen as the study of connected clusters in graphs. Or more relevant for resilience in network theory, the study of the probability that links get disconnected through say failure, giving rise to disconnected components in the network.

location", thus avoiding any precision about the scale and spatial extent of this agglomeration. Finally, the question of scale is also avoided by (Van den Berg et al. (2001), p.187), as they allow networks of firms to have a "*local or regional* dimension". To get the picture a little more confused, (Bergman and Feser (1999), p.2) "make a key distinction between clusters in economic space and clusters in geographic space". However, in the dominant majority, spatial industrial clusters tend to be identified first by the co-location of a set of interdependent firms or activities of a given industry, and second by the enclosing geographical unit in which they are located. More precisely, if this network happens to correspond to a territorial entity, the cluster becomes a local or regional cluster, otherwise it is left to other branches of economics to study. Unfortunately, these practices are not systematic and do not always use reproducible methods.

Park et al. (2019) propose an ambitious systematic approach to look at hierarchical firm clustering, using labour flows estimated by LinkedIn profiles over the past twenty years in the US. They are thus able to compare the geographical and industrial aspects of firm clustering through labour flows. In general, they show that homogeneity regarding the dominant industrial specialisation of firms tends to be stronger than their dominant geographical location, although both are significant. They can also match market capitalisation at the firm level and skills at the individual level to assess the profiles of dynamic clusters. However, what they call "geoindustrial clusters" diverges from our own definition, since these refer to network communities of firms that have geographical and industrial attributes attached to them, whereas we call "geoindustrial clusters" a group of geographical units which are close in terms of industry mix as well as in terms of travel proximity. In addition, their approach focuses on networks of firms given by the labour transitions, which defines a very precise subset of firms[3], whereas we are interested in the spatial evolution of economic activities. Hence, we consider all firms, including their constitutive units (or plants).

We use exhaustive administrative data at the business unit level, allowing us to refine the industrial description of economic activity in London. The data is gathered at the local level of businesses, ie. plants or establishments, referred to as 'business units' thereafter. Businesses can cover a large diversity of activities, and therefore it makes more sense to look at their constitutive units, described by their dominant industry, rather than working at the level of the firm (business). This paper also differs from (Park et al. 2019) with respect to the way clusters are identified. Instead of using community detection methods, which produce a partition of the network into communities, we choose a percolation algorithm which identifies tightly knit clusters rather than allocate all the nodes to communities. This means that some nodes are not clustered, and hence small areas might be excluded from our analysis, as they represent areas which are very distant and different from others[4]. This characteristic makes sense in our case, since we understand geoindustrial clusters as peculiar agglomerations of indutries in cities rather than the standard distribution of economic activity. Finally, we focus on the multi-scalar organisation of clusters by varying the value of the clustering parameter. This focus is independent from the clustering method chosen: for example, traditional hierarchical clustering methods already give rise to a hierarchical structure which can

[3] the ones with a positive turnover of workers who have reported working in those firms through their LinkedIn profile, thus excluding firms where current and former employees do not use the platform, firms with a stable workforce, etc.
[4] This could be achieved with community detection by selecting only "significant communities", although percolation additionally brings interesting properties with respect to phase transitions for example.

be visualised through the resulting dendogram; in addition, any community detection method could be used in a multi-scalar way by applying the algorithm recursively to each of the identified clusters or by adding a scale parameter, as shown respectively in Park et al. (2019) and Adam et al. (2018) with the Louvain method. These methods might not be entirely appropriate when uncovering modularity at different scales, and specific multi-scalar community detection techniques have been developed for this purpose, for example (Lambiotte 2010) and (Liu and Barahona 2018) for dynamical systems.

Among the approaches to clustering of businesses within network theory, Catini et al. (2015) suggest a graph-based method of cluster definition which "takes into account the relational patterns among co-located activities". Using geolocated PubMed scientific publications as an indication of activity for the biomedical sector, they apply the City Clustering Algorithm (CCA) (Rozenfeld et al. 2008) at a fixed distance of 1km, and then identify clusters through k-shell decomposition. This amounts to a percolation process based on a single dimension (physical distance) between firms of a single industry (the biomedical sector), using a single threshold (1km). In this paper, we present a network-based method which is similar to this framework but which extends it to all sectors and all relevant thresholds for a multi-scalar approach. The percolation process is applied to small geographical units based on the copula of two distances: the travel time distance and the similarity of the industrial composition between each pair of geographical units. In a nutshell, our approach consists in the aggregation of business data by industrial sector into geographical small areas. Each pair of small areas is described by the geographical distance between them and the similarity of their distributions of business units by industry. These two measures are reduced to one joint probability function using a copula. The full network is defined by nodes given by the small geographical areas, and by links which are the cumulative probabilities obtained through the copula. The percolation process then consists in selecting the links whose weights are above a certain threshold, decomposing the network and producing clusters given by the subgraphs[5]. Some nodes might not be linked to any other node and thus do not belong to any cluster. The operation is repeated for different threshold values to produce subsequent clusters of different scales, which can be represented in a hierarchical tree. This approach allows us to consider different resolutions of clusters which give rise to a nested structure across geographical and industrial similarity scales. Our approach provides a powerful insight on the relation between different cluster scales, which can ease the process of understanding spillover effects for policy making. It also bridges the various cluster studies produced at different scales. This piece of research is developed using longitudinal business microdata for London (see "Materials and methods" section), focusing on two years (2007 and 20014): before and after the financial crisis of 2008.

Materials and methods

London's microdata and economic geography

According to the Greater London Authority, i.e. the metropolitan institution which comprises the City of London, 13 inner boroughs and 19 outer boroughs, there were 8.825

[5] every node still linked together after pruning belongs to the same cluster.

million residents in London in 2017[6], about 5.5 million jobs and short of half a million local establishments or business units in our terminology.

"*Across London, the vast majority (86 per cent) of workplaces are part of very small firms; "micro-enterprises" employing less than 10 employees. [...] The London economy has specialisations in Professional, scientific and technical services; Finance and insurance; and Information and communication. Employment in these three industries is particularly concentrated in inner London, accounting for more than 33 per cent of jobs in Camden, Islington, Southwark and Westminster, almost 50 per cent of jobs in Tower Hamlets and over 70 per cent of jobs in the City of London in 2014. [...] By drawing in workers, tourists, and other visitors, central London areas also support jobs in accommodation, food, arts, entertainment, and retail services in the surrounding areas of inner London. In 2014, the combined Retail, and Accommodation and food services sectors for example accounted for around one in three employee jobs in Kensington and Chelsea, around one in four jobs in Newham and one in five in Haringey, with some evidence of recent growth in the number of jobs around the shopping centre developments in Stratford.* " (Girardi (2017), p.2-35). In order to draw a finer picture of the London economy, at the level of workplaces across the city at different points in time, we turned to a micro dataset recording business organisations in the UK, as well as their different units (or plants) if they operate from multiple sites. The Business Structure Database[7] (BSD) "*is derived primarily from the Inter-Departmental Business Register (IDBR), which is a live register of data collected by HM Revenue and Customs via VAT and Pay As You Earn (PAYE) records. [...] In 2004 it was estimated that the businesses listed on the IDBR accounted for almost 99 per cent of economic activity in the UK*"[8]. This database is provided by the Office for National Statistics (ONS) for free, although under secure access to protect anonymity and non-disclosure. In the context of the ONS non-disclosure rule, it means that no information which can allow the identification of a particular enterprise can be extracted from the secure environment. Most of the time, it means that information needs to be aggregated over at least 10 enterprises or local business units. However, a significant advantage of this database compared to any other free-access source which allows singling out individual enterprises (such as Companies House for example) is that this dataset is longitudinal. Therefore, we are able to follow anonymised local business units, in addition to obtaining the distribution of businesses by industry (SIC code) and by geographical areas through time in a comparable manner.

In this particular paper, we use all the active local units of Greater London in 2007 and in 2014. They represent 550,000 active units in 2014, from slightly over 400,000 in 2007. These local business units were extracted from the 7.5 million units active at one point in time in the UK, using London postcodes as a filter (Table 1). We retrieved from the BSD information about each business units' 5-digit SIC code, which corresponds to their main economic specialisation with a precision of 5 digits, according to the 2007 standard industrial classification (SIC) in the UK. For example, the manufacture of trailers and semi-trailers (29202) and the manufacture of caravans (29203) are distinguished from each other at this last level of the classification.

[6] https://data.london.gov.uk/dataset/projections
[7] Abstract copyright UK Data Service and data collection copyright owner.
[8] Office for National Statistics, 2017, Business Structure Database, 1997-2017: Secure Access, 9th edition, UK Data Service http://doi.org/10.5255/UKDA-SN-6697-10

Table 1 Total number of business units, 2007-2014

Year	United Kingdom	Greater London
2007	2,831,348	414,561
2008	2,833,652	416,450
2009	2,784,761	425,650
2010	2,711,446	417,451
2011	2,662,753	416,075
2012	2,749,395	446,331
2013	2,763,491	458,874
2014	2,874,224	490,011

Source: ONS: BSD 2007–2014

When local units are aggregated into administrative zones such as LSOAs[9], they show a very clear polycentric pattern of density organisation (Fig. 1). Central London (the City, Westminster, Kensington and Chelsea for example) hosts several hundreds to several thousands of business units per km^2, whereas the outer boroughs tend to be more residential and host less than 100 business units per km^2, at the exception of secondary centres, such as Croydon, Barking, Kingston upon Thames, Harrow or Bromley.

Between 2007 and 2014, a period which includes the financial crisis of 2008, the number of business units in London as a whole has grown, declined then grown again to exceed its initial number (Table 1). At the level of LSOAs, we can see that the highest levels of business unit growth have happened in East London, in the formerly industrial boroughs of Hackney, Tower Hamlets, Newham, Redbridge and Barking and Dagenham (Fig. 2). The areas where the number of business units is lower in 2014 than in 2007, are spread out in the outskirts of Greater London (the south of Bromley or Richmond upon Thames, the east of Havering, the west of Hillingdon for example) and in specific areas of central West London (the north riverside of Wandworth, around Elephant and Castle in north Southwark or Regent's park in Westminster).

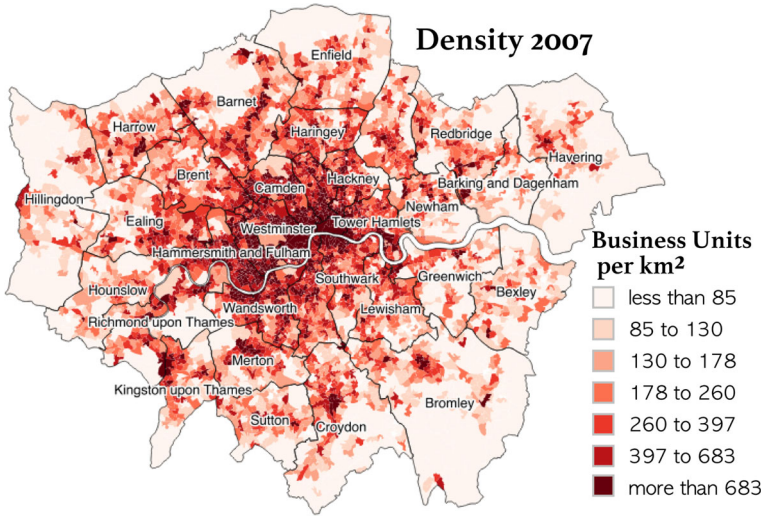

Fig. 1 Density of business units in Greater London LSOAs, 2007. Source: ONS

[9]Lower layer of Super Output Areas, corresponding to a couple thousand residents in the census.

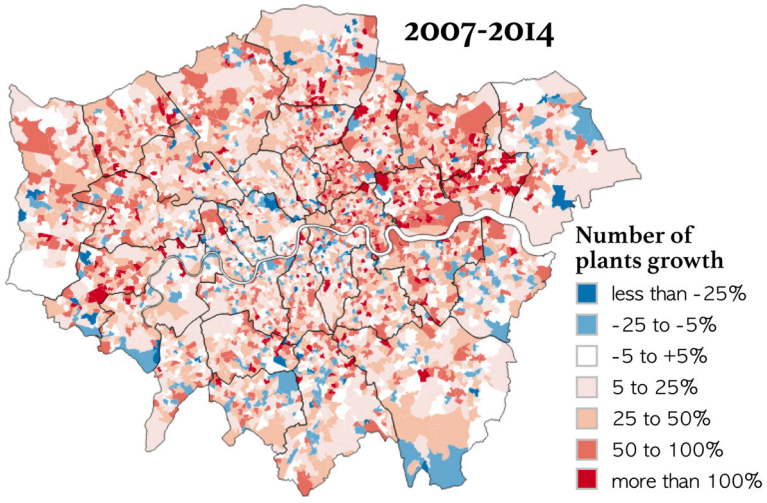

Fig. 2 Variation of the number of business units in Greater London LSOAs, 2007-2014. Source: ONS

If we consider the distribution of business units by industry within LSOAs, we can refine the description of this economic geography with some diversity and specialization measures. For example, for each LSOA j, we compute the Shannon Entropy Index (SEI_j) from the share $b_{i,j}$ of business units in j whose production correspond to the SIC code i among all 5-digit SIC codes ($I = 729$):

$$SEI_j = -\sum_{i=1}^{I} b_{i,j} * \ln(b_{i,j}) \tag{1}$$

The result for all LSOAs in Greater London shows a heterogeneous picture (Fig. 3, left hand side). The most diverse places in terms of industries are Central London (including the areas of Westminster, the City, the South Bank and the West End) and the subcentres around Heathrow airports, Croydon and Wembley. These areas also correspond to the densest in terms of business units (Fig. 1).

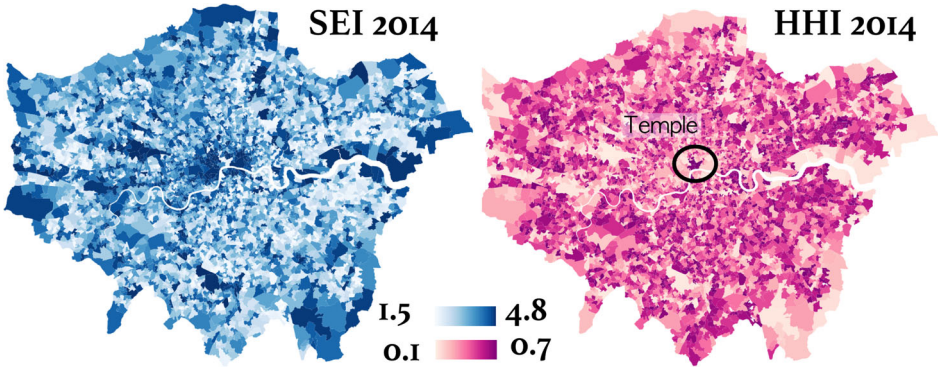

Fig. 3 Entropy index and Herfindhal index of industrial specialisation. Greater London LSOAs, 2014. Source: ONS

In terms of industrial specialisation, the Hirschman Herfindahl Index (HHI_j) highlights the areas j which have an industrial profile strongly different from the overall proportion of industries in Greater London

$$HHI_j = \sqrt{\sum_{i=1}^{I} b_{ij}^2}. \tag{2}$$

It means that areas with high values of HHI have very specific profiles and concentrate an uncommon set of industries in a relatively strong manner. For example, Temple appears as an outlier in the distribution of activities, whereas the other parts of Central London are very representative of the distribution of activities in London overall (Fig. 3, right hand side). The two measures of diversity and specialization are negatively correlated at the level of LSOA ($R = -0.88$ in 2014, and $R = -0.90$ in 2007), as evidenced by Fig. 3.

In order to go beyond the spatial juxtaposition of diversity measures and to account for the similarity between small areas in terms of the mix of business units they host, we built a complete network of LSOAs and consider geoindustrial proximity as the weight of its edges.

Defining geoindustrial proximity

Geoindustrial proximity is here defined as the combination of geographical closeness and industrial similarity between LSOAs. Hence we need to define a measure of geographical distance and a measure of industrial similarity, to apply them to all pairs of LSOAs and then to combine them into a single measure of proximity. For London, the analysis is performed over the 11+ million pairs of Greater London LSOAs.

Geographical distance. There are many different possibilities to account for geographical distance. In the context of a city like London, the connectivity between two different areas is better represented by the availability of public transport between these two zones, instead of the physical distance. We take the transportation networks for the following modes of transport: underground, buses and rail, developed under the project QUANT[10]. The dataset is constructed using the mean travel time according to the timetables as per May 2016 for each mode of transport[11]. The network collapses the three modes on one layer. For each LSOA, the node with higher connectivity, i.e. the node with higher degree is chosen as the *centroid* for the LSOA. Then, the shortest travel time between LSOA centroids is computed, and this is what is used for the weight of the links. The walking time (5 miles/hour) required when changing modes of transport is also taken into account. If for any reason there is no public transport connecting the LSOAs, we use walking time, and the centroid corresponds to the physical centroid.

[10] Spatial interaction model for the whole of the UK considering the above mentioned modes of transport and employment: http://quant.casa.ucl.ac.uk/

[11] Unfortunately, such a network was only constructed for that specific year and month within the QUANT project. Therefore, the dynamics shown in subsequent results depends entirely on the dynamics of business units locations. It would be very interesting to compare these results with the changes induced by accessibility variations over time if the travel time network was available. Overall, the main changes in the public transport system of London are localised to the Overground lines. In particular, these relate to connecting the East of London to peri-central areas in anticipation for the 2012 Olympic Games. The current research might therefore present biases within East London, where further connectivity than the existing one at the time is considered. Extracting the real travel time network of 2007 requires a detective work of recovering those original time tables, since google maps API requests cannot be performed for that date. Such a task is beyond the scope of this research. For future studies however, there will be an opportunity to assess the impact of Crossrail and HS2 on the measured transport times compared to May 2016.

Industrial similarity. In order to compute the industrial similarity, we start by aggregating the business units by 2-digit SIC category (SIC2 level) for each LSOA, and by removing from the analysis all LSOAs containing less than 10 business units (due to the ONS non-disclosure condition). In this sense, we are assuming that *"SIC categories are a reasonable measure of relatedness"* (Bishop and Gripaios (2007), p.1746). Local units are attributed to one of the 88 distinct 2-digit SIC categories. We then use a measure of cosine similarity as in Eq. (3)

$$s_{ij} = \frac{\mathbf{V_i} \bullet \mathbf{V_j}}{\|\mathbf{V_i}\|\|\mathbf{V_j}\|} \tag{3}$$

where $\mathbf{V_i}$ and $\mathbf{V_j}$ are the vectors of LSOAs i and j respectively, defined in the $n = 88$ space of the industrial categories. The cosine similarity has been favoured here over other measures for its simplicity of implementation and computational efficiency, but our approach could very well be implemented with alternative measures of similarity, such as the Jaccard similarity for example.

Geoindustrial proximity. Instead of considering either geographical proximity or industrial similarity, we construct a probability function that takes into account both. Given that we are considering travel time for proximity, we need to transform the time t to $x'_t = 1/t$[12], so that a smaller time reflects a stronger connection. In addition, we normalise x'_t so that both the variable s and the new variable $x_t = x'_t / \max(x'_t)$ lie within the same interval $[0, 1]$. Note that by construction, the similarity s given by Eq. (3) already does.

The joint probability function for s and x_t is constructed using a copula, which is widely used in the field of quantitative finance to model multivariate dependencies (Low et al. 2013). In detail, copulas are functions which encode the joint probability distribution of variables that are not necessarily normally distributed, and whose dependencies are not necessarily linear. Different types of functions are used to encode the different dependencies, and these are called *families*. Each family (for example Gaussian, Student, Clayton, Gumbel, Joe or even a combination of the latter) has a different set of parameters. In more technical detail, a copula of random variables $(X_1, ..., X_d)$ corresponds to the joint cumulative distribution function (CDF) $C : [0, 1]^d \rightarrow [0, 1]$ of the uniformly distributed marginals $(U_1, ..., U_d)$ (Joe 1997). This means, that we first need to find the uniform distribution of (x_t, s) as a bivariate vector (U_1, U_2) (see the Additional file 1 for details). Then we proceed to construct the copula: $C(u_1, u_2) = P(U_1 \leq u_1, U_2 \leq u_2)$ using the *VineCopula* package in *R*[13]. The package provides a function which computes the bivariate copula family and estimates the parameters of the family through maximum likelihood. We find that for both years, the family is the same Archimedian one, and that its parameters are very similar[14], see Fig. 4 for 2007. Ultimately, the copula produces the cumulative joint probability, allowing us to extract specific values p_{ij} which encapsulate simultaneously industrial similarity and geographical proximity between area i and area j. A value of p_{ij} close to 1 means

[12] As shown in Additional file 1, the results are the same if we transform the time distance with $x'_t = e^{-t}$
[13] This package is the continuation of the *CDVine* package (Brechmann and Schepsmeier 2013)
[14] Bivariate copulas used for 2007 and 2014 are from the BB1 family, which correspond to a combination of the Clayton and Gumbel families of generator functions. See Additional file 1: Section A and (Brechmann and Schepsmeier (2013), p.7-8) for more details.

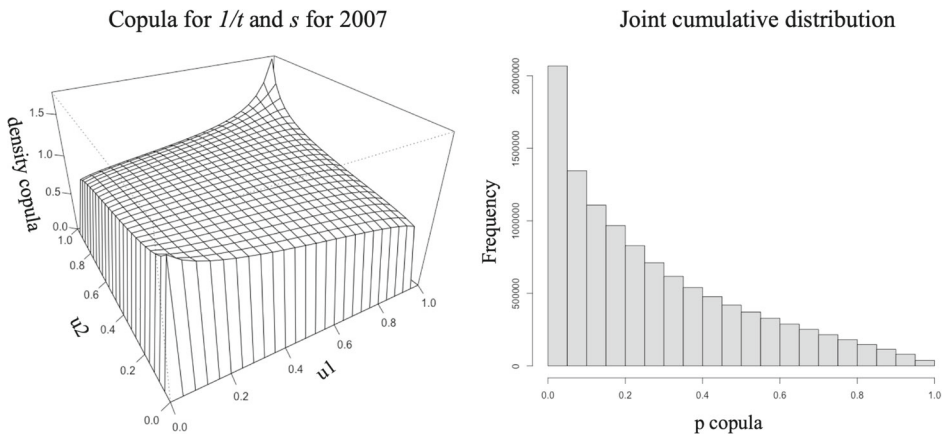

Fig. 4 Copula $C(u1, u2)$ for $1/t$ and s, where t is the time it takes to go from one LSOA to another using public transport, and s is the cosine similarity between their industrial sectors defined by the SIC codes of their firms. London, 2007. Left: density function; right: histogram of the joint cumulative probabilities p_{ij} between LSOAs i and j defined by the copula

that i and j are highly similar and highly accessible to one another, whereas a value of p_{ij} close to 0 means that i and j are highly dissimilar and far from each other.

The main network of geoindustrial proximity between LSOAs is constructed as a fully connected network, weighted by the intensity of their geoindustrial proximity, which is encoded in the copula. This is defined as $G = (V, L)$, where the elements of the set of nodes $V = \{n_1, ..., n_N\}$ correspond to the N LSOAs, and the elements of the set of links $L = \{p_{ij}\}$ for $i, j \in [1, N]$ correspond to the weights of the network, given by the copula probabilities. We illustrate some of the strongest links in the network in Fig. 5. These correspond to the ones whose pairs of LSOAs are the most proximate in terms of temporal distance and industrial similarity. Highly weighted links are dense in Central London,

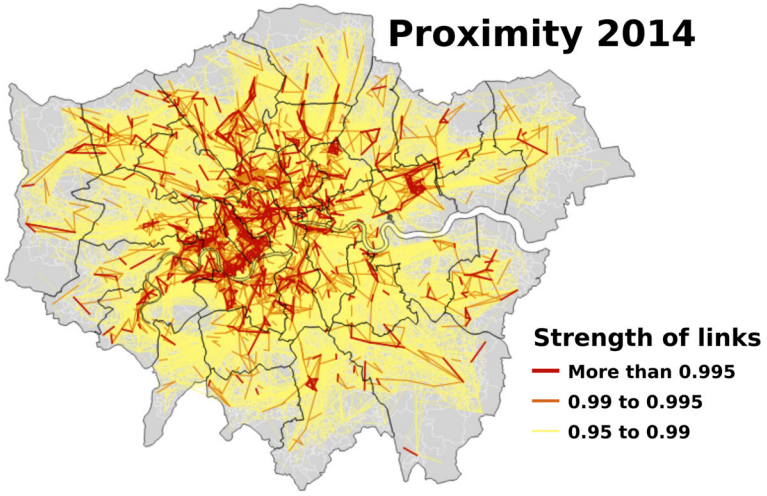

Fig. 5 Network of the most proximate Greater London LSOAs in 2014. This geoindustrial proximity is obtained through the copula for $1/t$ and s, and the weights correspond to the probabilities p_{ij}. Source: ONS

especially in the borough of Kensington and Chelsea, Hammersmith and Fulham, and Westminster, but also around Croydon and Newham.

Clustering method

In this paper, we apply the algorithm derived in Arcaute et al. (2016) to the network G, and obtain the hierarchy of clusters from the multiplicity of transitions. The clusters are the result of a thresholding process such that $p_{ij} > p$, where the value of the threshold probability p, carries no direct interpretation other than the higher p the stronger the geoindustrial proximity. In detail, the percolation process consists in recursively selecting the links from the set L whose weight p_{ij} is above a certain varying threshold p. After each iteration, an induced graph is produced after removing the links that do not meet the constraint. Such a graph might be composed of several disconnected components, which are considered to be the resulting clusters, if these have at least two nodes. In this sense, the process prunes the network of disconnected individual nodes, so that not all original nodes, i.e. LSOAs, are necessarily clustered. Given that the method is iterative, the subsequent clusters obtained with the next threshold p can be represented in relation to clusters of other scales (i.e. values of p) in a hierarchical tree.

One of the main novelties of the way we use percolation theory in the context of geoindustrial clustering, is that instead of looking at a single configuration of clusters in the space, we derive a hierarchical structure by looking at the nested configuration at different scales. This approach has shown promising results in other fields. For example, Gallos et al. (2012) used the rates of obesity to cluster US States to identify spatial clusters of similar health behaviour. Arcaute et al. (2016) used the metric distance of road segments to produce a hierarchical clustering of the UK, showing that different distance thresholds highlight different spatial discontinuities in the road network. Molinero et al. (2017) extended the method using the angular distance, and obtained a classification of the importance of streets in the road network, in addition to deriving the main skeleton of urban systems without further assumptions. All these methods are based on the CCA clustering algorithm developed in Rozenfeld et al. (2008, 2011).

For the following sections, it is important to note that the copula obtained has an extremely small variance, which can be observed in Figs. 4 and 5. This causes the main transitions of interest to occur right at the tip, which correspond to values of $p > 0.9$ for the CDF of the copula. Under this threshold, all LSOAs are close and similar enough to form a single giant cluster for the whole city.

Results

In the following, we look at the structure of business unit clusters and its evolution in London between 2007 and 2014, that is, before and after the financial crisis. We first analyse how clusters are located and structured in London in 2014 and how they nest across scales using different thresholds (section "The nested structure of businesses in London in 2014"). We then present the evolution of clusters between 2007 and 2014 (section "The evolution of clusters through the financial crisis"). Finally, we turn to show how these clusters specialise in two key sectors of the London economy, namely the knowledge intensive sector and the retail and leisure industry (section "Industrial specialisation of clusters in the city").

The nested structure of businesses in London in 2014

In 2014, the tech sector and the post-Olympic industry were flourishing in the eastern part of central London. They are reflected in the clusters formed at the threshold of 0.994, shown in Fig. 6b. Among the largest ones in terms of the number of business units included, in blue, the City of London and technological fringes appear as one cluster. So does Stratford in mint green. We can also spot the finance cluster of Canary wharf and the Docklands in apple green. Other clusters feature in central London: the law cluster of Temple, the media, communication and management area around Hyde Park East or a banana-shaped cluster following the Thames and the train lines in South-West London. With a more restrictive threshold, for example 0.999 (Fig. 6a), LSOAs have to be very close and very similar to be aggregated into such clusters. Therefore, clusters are much smaller. For example, the areas of Shoreditch and of Aldgate appear as two different clusters (although they will belong to the same cluster above the threshold of 0.994). We also identified the cultural area of London Bridge and the research hotspot of Tottenham Court Road as small independent clusters. The exception is the cluster of Temple, which remains pretty much the same size and extent regardless of the threshold chosen, meaning that this cluster is coherent but consistently dissimilar to neighbouring clusters.

The hierarchical tree in Fig. 7 visualises the nested structure of the clusters, by showing how clusters at one level (of threshold value) are merged into bigger clusters at the level above (with lower thresholds). Interestingly, we can thus relate the clustering of London business units across scales and identify the proximity not only between LSOAs but also between clusters by looking at which clusters are fused sooner than others. For example, although Aldgate and Shoreditch are neighbouring clusters, they do not merge until the threshold 0.994. Instead, Aldgate merges with the City of London and Shoreditch merges with Farringdon at the threshold of 0.996 (cf. Fig. 7, left hand side branches). The two merged clusters are fused with the South Bank cluster into the East Central cluster at the threshold of 0.994. Further on, this central cluster fuses with Kensington and Chelsea, the West Thames and the Docklands and other smaller clusters at the threshold of 0.993. The Olympic area of Stratford joins this giant cluster only at the level 0.991. This structure thus highlights proximities between clusters which are usually absent from standard analyses. Let us now look at how this hierarchical organisation changes between 2007 and 2014 in the following section.

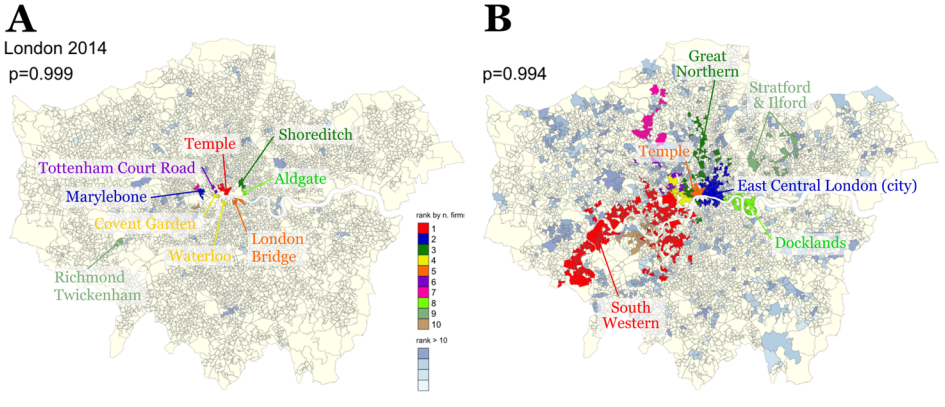

Fig. 6 Geoindustrial clusters in 2014. **a**: percolation threshold of 0.999. **b**: percolation threshold of 0.994

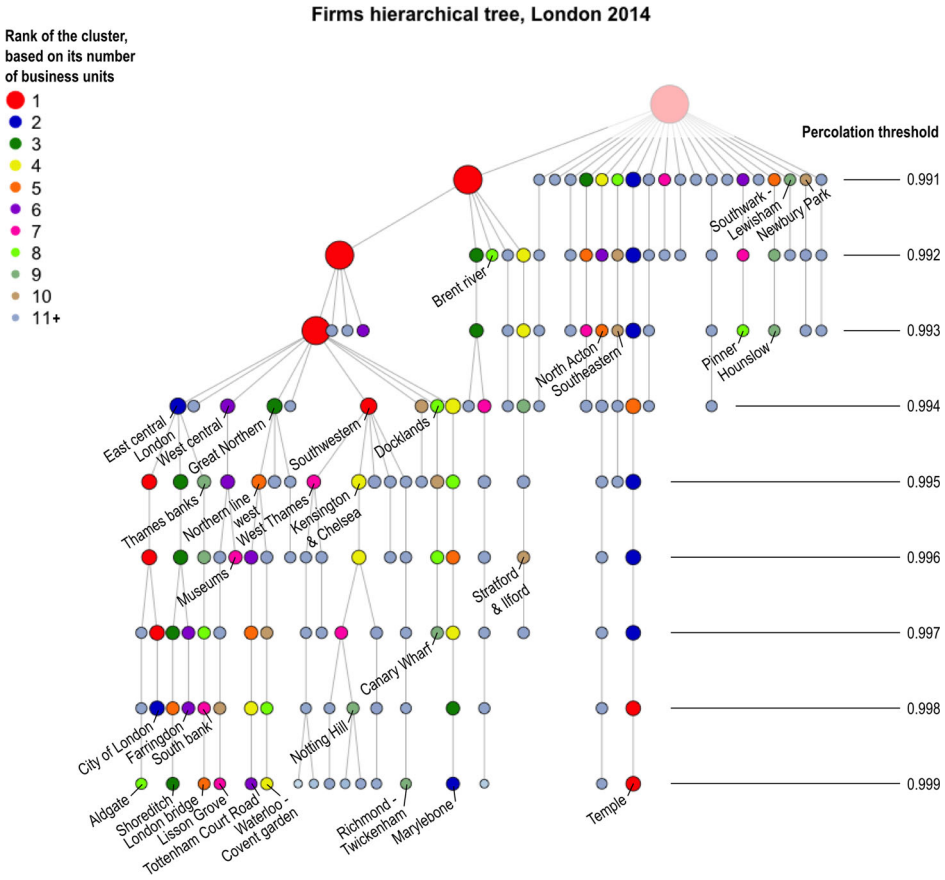

Fig. 7 Hierarchical tree of London geoindustrial clusters in 2014

The evolution of clusters through the financial crisis

First of all, the 2007 tree has overall a different structure. In Fig. 7, we could see on the left hand side a few branches merging separately before being merged into a single giant cluster. In Fig. 8, a reduced version of this phenomenon occurs, but then a giant cluster takes over quicker and small clusters gradually get added to it at the threshold of 0.995 and over. Other differences refer to the "Tech cluster" identified in 2014 on the Eastern fringes of the City. First of all, the City of London remains a single consistent cluster from 0.999 to 0.995, before it takes part into a Central London big cluster. A cluster for Aldgate, equivalent to the one in 2014 is not visible. Similarly, Shoreditch and Farringdon appear as small clusters but do not merge together until reaching the big Central London cluster at the threshold of 0.995. This suggest that between 2007 and 2014, their industrial profile have become more similar, most probably thanks to digital services start-ups in both areas. The only small clusters to agglomerate early on are LSOAs of the West end, with Mayfair and Carnaby Street merging at the level of 0.997 for example, then fusing with Kensington and Belgravia at the level of 0.996. This comparison therefore shows a significant shift of industrial clustering from the West to the East of Central London between 2007 and 2014, where the sectoral and temporal proximity have become stronger over larger extents. This shift might reflect the pressure on rent costs for businesses after the financial crisis of 2008, combined with a shift of the London economy towards knowledge

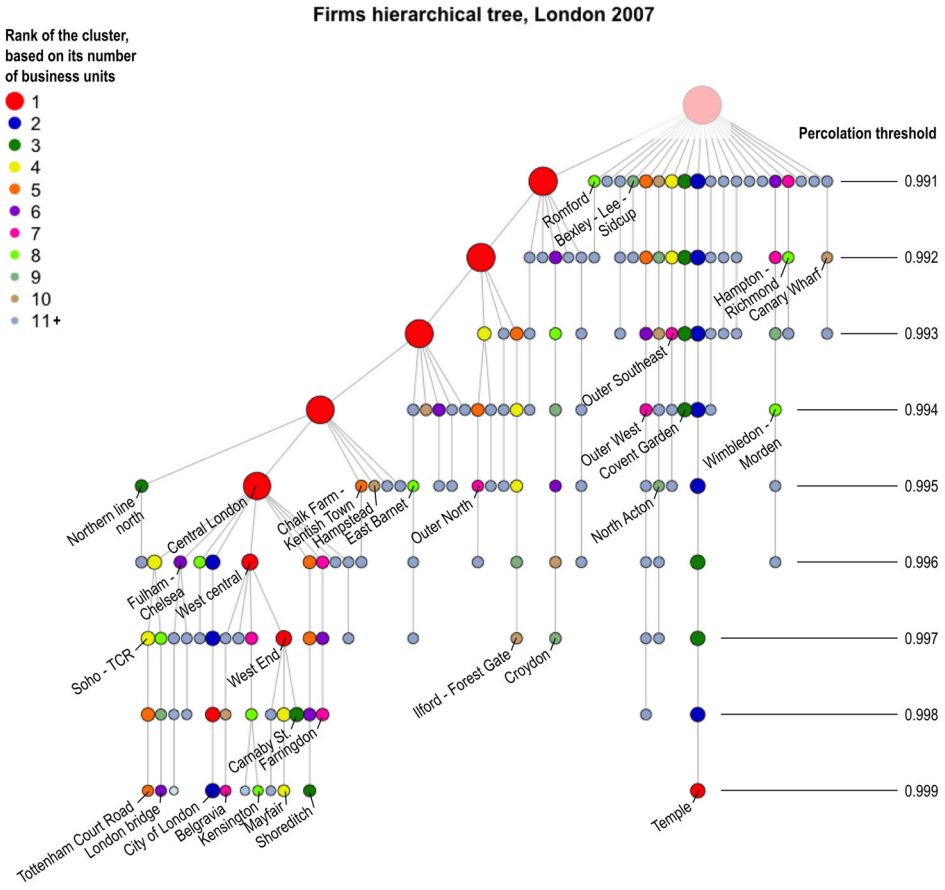

Fig. 8 Hierarchical tree of London geoindustrial clusters in 2007

and technology-based industries powered by small companies and a younger workforce attractive by the work and play lifestyle of East London (McWilliams 2015).

It is also interesting to notice that Temple has remained a coherent and differentiated cluster for all thresholds in both years, whereas Canary Wharf just emerges as a top 10 cluster at the threshold of 0.992 in 2007, while it was already at the top at 0.997 in 2014. The redevelopment of the Docklands dates back from the 1980s, yet they have become a major player in London finance later on. After the financial crisis, many bank units relocated from the City to Canary Wharf taking advantage of the lower rents, which also allowed startups in Fintech to setup. In addition, when banks move, they do so bringing with them all the businesses that provide them with different services, from insurance to catering. Such a move generates the relocation of many business units. This, together with the fact that existing banks in Canary Wharf dissolved to become financial outfits, explain the structural change of the surge of firms in the area in 2014.

Finally, some clusters which look prominent in 2007 have disappeared from the hierarchical tree in 2014. A notable example of such clusters is that of Croydon, which experienced a decrease in job density, partly caused by the crisis in the finance and insurance sector "*including Allianz Global Assistance, RA Insurance Brokers, and AIG Europe*" (Girardi (2017), p.), as well as by urban redevelopments (the Nestlé tower for example). These changes are better interpreted by looking at the specialisation of each cluster throughout the percolation tree. The following section present these results.

Industrial specialisation of clusters in the city

In terms of specialisation, we have looked at two broad sectors of the economy. The first sector aggregates knowledge based industries (KBI), i.e. business units whose dominant industry (in terms of 5-digit SIC classification) relates to digital activities, science, publishing and other scientific services, as defined by the ONS Science and technology classification in 2015[15]. The second sector chosen comprises retail, entertainment and food businesses (RAL), according to Oliver Wymans' classification of the leisure industry in 2012[16]. The KBI and RAL sectors represent respectively 15.6% and 18.4% of all business units in the UK in 2014 and respectively 21.1 and 16.7% in Greater London. However, they are characterised by very different consumers and spatial strategies. Indeed, retail units are generally organised linearly along the high streets and in commercial zones widespread throughout the city, whereas the knowledge sector is thought to be the epitome of industrial clusters. We would then expect to find less homogeneous specialisation in KBI than in RAL, but also a reinforcing trend in specialisation and heterogeneity for KBI between 2007 and 2014, because this (rather) new sector would have clustered even more.

What we find regarding the first hypothesis (Figs. 9 and 10) is that the amplitude of industry share by cluster is much wider for Retail and Leisure than for the Knowledge based industries. Indeed, some clusters like Wimbledon-Morden, Croydon, Ilford and Forest gate, the Outer North in 2007 (Fig. 11c) and Stratford/Ilford in 2014 (Fig. 11d) have more than 40% of RAL businesses, compared to less than 10% in Temple. On the other hand, the maximum shares for KBI (in Farringdon both years, Newbury Park and Hounslow in 2014, cf. Fig. 11b) are under 40%, while the minimum share are also between 5 and 10%. The heterogeneity is therefore more pronounced regarding RAL than KBI for the clusters we have identified using all sectors. As for the evolution between 2007 and 2014, again in contrast with our hypothesis, there seem to be more reinforcing for Retail and Leisure than for the Knowledge Based industries. Indeed, in 2007 there were a few branches of concentrated RAL (Kensington, Romford, Wimbledon, the Outer North, West and SouthEast, Ilford and Forrest Gate, Bexley), whereas there are fewer in 2014 (Lewisham, Stratford and Ilford, Waterloo and Covent Garden for example). For KBI on the other hand, there seems to be a similar number of highly concentrated KBI branches in 2007 (Soho, Shoreditch, Farringdon, Kentish Town, Fulham and Chelsea) and in 2014 (Soho, Shoreditch, Farringdon, Kentish Town/Hampstead, West Thames). There is however a very significant change: in Canary Wharf, the KBI firms went to represent from about 15% to more than 25% of all local businesses, illustrating the spread of FinTech in Canary Wharf finance, in contrast to the more traditional City finance.

The comparison of all four trees shows two interesting areas. The unique cluster of Temple (which remains unchanged between 2007 and 2014 and throughout the thresholds) shows very low shares of both KBI and RAL sectors. Indeed, this cluster is very coherent and very specialised (cf. Additional file 1: Section "Diversity and concentration of firms"), as shown with high values of HHI specialisation index (Fig. 3), but in a different industry to KBI or RAL (i.e. in law and justice related activities). The same is true, to a lesser extent, of other very central areas in West London, around Hyde Park for example).

[15] https://webarchive.nationalarchives.gov.uk/20150905170202/http://www.ons.gov.uk/ons/rel/regional-trends/london-analysis/identifying-science-and-technology-businesses-in-official-statistics/artidentifying-science-and-technology-business.html?format=print

[16] http://www.oliverwyman.com/content/dam/oliver-wyman/global/en/files/archive/2012/20120612_BISL_OW_State_of_the_UK_Leisure_Industry_Final_Report.pdf

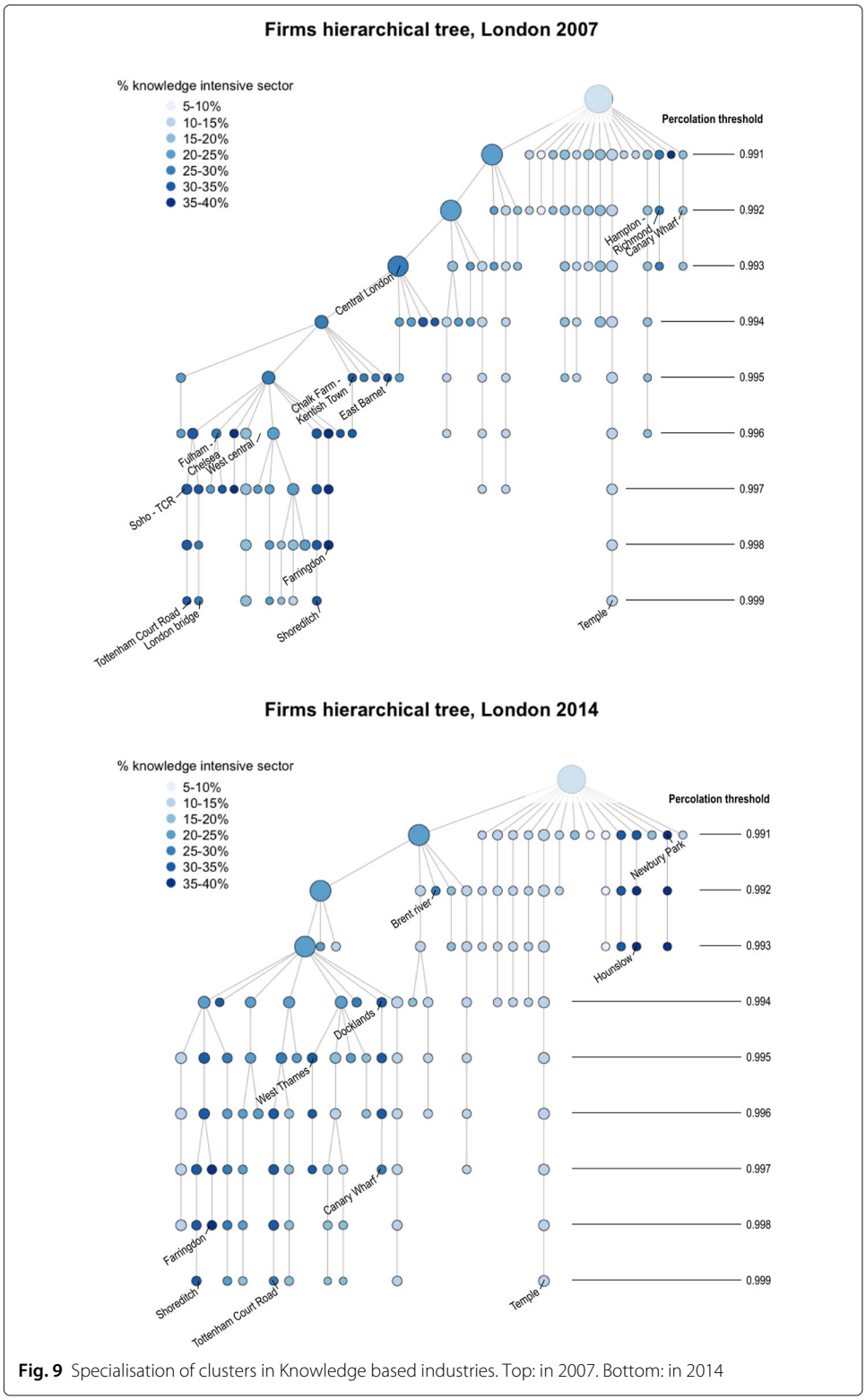

Fig. 9 Specialisation of clusters in Knowledge based industries. Top: in 2007. Bottom: in 2014

On the other hand, a large mix of KBI and RAL characterises the geoindustrial cluster around Tottenham Court Road, where we find around 30% of KBI businesses and around 20% of RAL businesses in 2014, which is an over-representation of both sectors compared

Fig. 10 Specialisation of clusters in retail and leisure industries. Top: in 2007. Bottom: in 2014

to the London average. This area is historically a retail one, but the presence of universities (among which UCL) has attracted publishing and science services companies.

We map in Fig. 11 the KBI and RAL concentration level of clusters at the threshold of 0.997, in 2007 and 2014. The highest percentage of KBI firms in clusters in 2007 seems

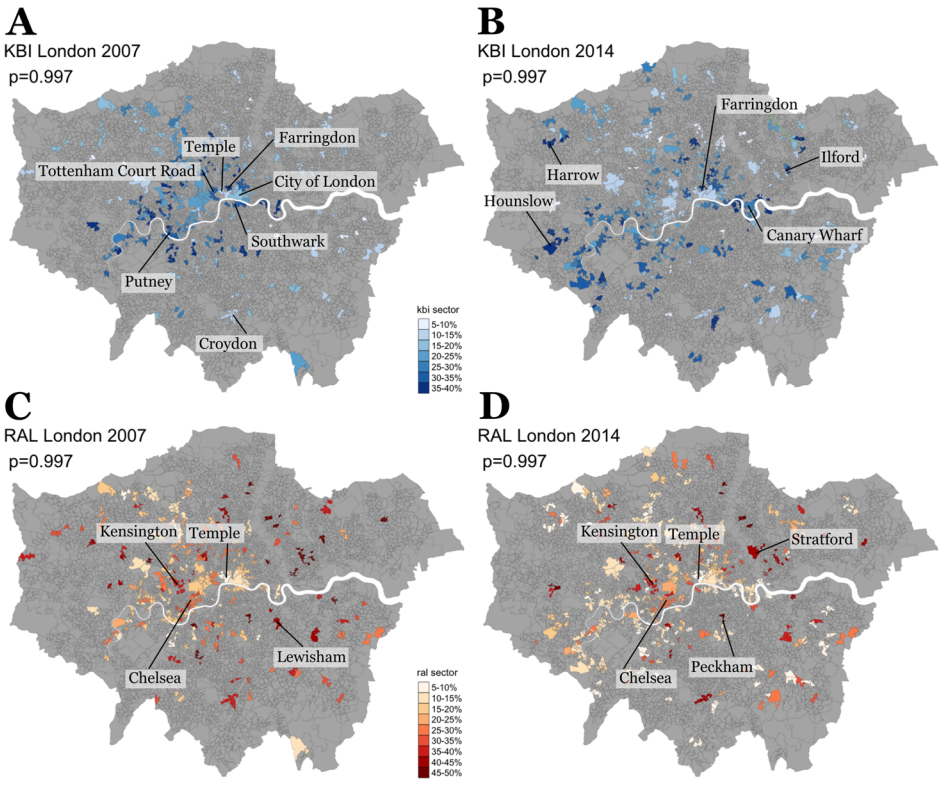

Fig. 11 Specialisation of clusters in knowledge based industries: diffusion of specialised clusters from 2007 (**a**) to 2014 (**b**). Specialisation of clusters in retail and leisure industries: stable structure with the emergence of Stratford between 2007 (**c**) and 2014 (**d**)

to be found in the first ring around central London (including the clusters related to publishing in Southwark and Camden). In 2014, the clusters specialised in Knowledge-based industries have expanded to Outer London boroughs. For example, KBI-dense clusters can be found around Hounslow, Harrow or Richmond. "*Examples of technology companies in these areas include: IBM, Sega Europe, Cisco Systems and SAP offices at Bedfont Lakes Business Park in Hounslow [... They] also show a high level of specialisation in Professional, scientific and technical activities. Within the sector, Richmond upon Thames is particularly specialised in scientific research and development (1,700 jobs, IOS = 5.3). Examples of related employment sites in the area include the scientific parks and research centres associated with Kew Gardens, the National Physical Laboratory and LGC Group30*" (Girardi (2017), p.22-4). This expansion reflects both the increasing share of KBI firms in London, their new location strategies in the outer boroughs where office space is cheaper, but also the fact that Central London is hosting a more diverse set of companies when it is hosting KBI companies.

Regarding the spatial pattern of Retail and Leisure specialisation (RAL), we find two main areas of high concentration across the years: Kensington on the one hand, and the Stratford/Lea Valley on the other hand. *"In Kensington and Chelsea, the main employers are in Retail (23,000 jobs) and Accommodation and Food services (19,000 jobs), likely reflecting the areaś role in attracting visitors to London. [...] Examples of major employers in the sector within the borough include the department stores: Harrods, Peter Jones and*

Harvey Nichols in Knightsbridge" (Girardi (2017), p.12-27). Our method shows that this specialisation holds at the borough level for the lower scale of 0.997 clusters, although with varying intensities between Kensington and Chelsea for example.

Conclusion

With a multidimensional view of proximity which includes time distance and industrial similarity, this paper has offered a renewed take on geoindustrial clusters in London, one that pays particular attention to scales with the use of percolation theory. It has uncovered an evolution of the London economic geography which was not available through other methods, such as the reorganisation of the central London business structure post-crisis, allowing different clusters to co-exist alongside (City-Aldgate, Shoreditch-Farringdon, Notting Hill, Tottenham Court Road for example) rather than a hierarchical central cluster absorbing peripheral extensions as in 2007. We have highlighted changes regarding the structure and the specialisation of clusters in London. It should be noted that this work, through the methodological choices made, is limited firstly to an analysis of aggregated small areas rather than the network of firms through business links or workforce transition. Secondly, the analysis of the present paper does not include economic links to external places, within national boundaries and more generally within the Global Value Chain (Sturgeon et al. 2008): "*The processes of dispersal are not confined to the re-location of economic activity to some newly dynamic center where the agglomeration process can begin anew (Storper and Walker 1989), but also include the unfolding — and perhaps historically novel — dynamics that are presently driving deep functional integration across multiple clusters (Dicken, 2003, 12), a process we refer to as global integration*" (Sturgeon et al. (2008), p.299). Finally, we have restricted our view to the main sectors of KBI and RAL, leaving big parts of the service sector untouched by the analysis. Despite these limitations, our hope is that, by providing a methodology for multiscale cluster analysis, we can emulate comparative works in other regional and national contexts, and unveil different nested structures to inform economic analysis.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1007/s41109-019-0246-9.

Additional file 1: Appendix.

Acknowledgments
We thank the ONS for letting us use the Virtual Microdata Laboratory facility. We are grateful to Richard Milton for providing us with the time distance matrix of London LSOAs, to Max Nathan for insightful discussions and references as well as to the participants of the micro-networks special session at the Regional Studies Association Annual meeting 2017 for their comments and questions.

Authors' contributions
CC and EA contributed equally to this study. Both authors read and approved the final manuscript.

Funding
We acknowledge the funding of the EPSRC grant EP/M023583/1.

Availability of data and materials
The dataset supporting the conclusions of this article (similarity matrix between LSOA, Source: ONS) is available in a FigShare repository, 10.6084/m9.figshare.8035961.
The distance matrix between LSOAs is proprietary data but can be generated again using the Google Maps API. We invited readers interested in this data to contact us or to run the queries on the API.
This work contains statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

Competing interests
The authors declare having no conflict of interest regarding this work.

Received: 6 May 2019 Accepted: 10 December 2019
Published online: 07 January 2020

References

Adam A, Delvenne J-C, Thomas I (2018) Detecting communities with the multi-scale louvain method: robustness test on the metropolitan area of brussels. J Geogr Syst 20(4):363–386
Arcaute E, Molinero C, Hatna E, Murcio R, Vargas-Ruiz C, Masucci AP, Batty M (2016) Cities and regions in britain through hierarchical percolation. Open Sci 3(4):150691
Bergman EM, Feser EJ (1999) Industrial and regional clusters: concepts and comparative applications. University of West Virginia, Morgantown: University of West Virginia Webbook, http://www.rri.wvu.edu/WebBook/Bergman-Feser/contents.htm
Bishop P, Gripaios P (2007) Explaining spatial patterns of industrial diversity: an analysis of sub-regions in great britain. Urban Stud 44(9):1739–1757
Brechmann E, Schepsmeier U (2013) Cdvine: Modeling dependence with c-and d-vine copulas in r. J Stat Softw 52(3):1–27
Catini R, Karamshuk D, Penner O, Riccaboni M (2015) Identifying geographic clusters: A network analytic approach. Res Policy 44(9):1749–1762
Crouch C, Farrell H (2001) Great britain: Falling through the holes in the network concept. In: Le Galès P, Trigilia C, Voelzkow H (eds). Local Production Systems in Europe: Rise or Demise? Oxford University Press, Oxford. pp 154–211
Dicken P (2003) Global shift: Reshaping the global economic map in the 21st century. Sage
Enright M (1996) Regional clusters and economic development: a research agenda. In: Staber U, Schaefer N, Sharma B (eds). Business networks: prospects for regional development. Walter de GRryter, Berlin. pp 190–213
Feser EJ (1998) Old and new theories of industry clusters. Clusters Reg Specialisation:16
Foord J (2013) The new boomtown? creative city to tech city in east london. Cities 33:51–60
Gallos LK, Barttfeld P, Havlin S, Sigman M, Makse HA (2012) Collective behavior in the spatial spreading of obesity. Sci Rep:2. https://doi.org/10.1038/srep00454
Girardi A (2017) A description of london's economy
Iammarino S, McCann P (2016) Network geographies and geographical networks. co-dependence and co-evolution of multinational enterprises and space. The New Oxford Handbook of Economic Geography, Oxford University Press, Oxford
Joe H (1997) Multivariate Models and Multivariate Dependence Concepts. CRC Press
Lambiotte R (2010) Multi-scale modularity in complex networks. In: 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks. pp 546–553
Liu Z, Barahona M (2018) Geometric multiscale community detection: Markov stability and vector partitioning. J Complex Netw 6(2):157–172
Low RKY, Alcock J, Faff R, Brailsford T (2013) Canonical vine copulas in the context of modern portfolio management: Are they worth it? J Bank Finance 37(8):3085–3099
Malmberg A, Maskell P (2002) The elusive concept of localization economies: towards a knowledge-based theory of spatial clustering. Environ Plan A 34(3):429–449
Martin R, Sunley P (2003) Deconstructing clusters: chaotic concept or policy panacea? J Econ Geogr 3(1):5–35
Martins J (2015) The extended workplace in a creative cluster: Exploring space (s) of digital work in silicon roundabout. J Urban Des 20(1):125–145
McWilliams D (2015) The Flat White Economy: How the digital economy is transforming London and other cities of the future. Gerald Duckworth & Co
Molinero C, Murcio R, Arcaute E (2017) The angular nature of road networks. Sci Rep 7(1):4312
Nathan M, Vandore E (2014) Here be startups: Exploring london's 'tech city'digital cluster. Environ Plan A 46(10):2283–2299
Park J, Wood IB, Jing E, Nematzadeh A, Ghosh S, Conover MD, Ahn Y-Y (2019) Global labor flow network reveals the hierarchical organization and dynamics of geo-industrial clusters. Nat Commun 10(1):1–10
Porter ME (1998) Clusters and the new economics of competition, volume 76. Harvard Business Review Boston
Roelandt TJ, Den Hertog P, van Sinderen J, van den Hove N (1999) Cluster analysis and cluster policy in the netherlands. Boosting Innov Clust Approach:315
Rosenfeld SA (1997) Bringing business clusters into the mainstream of economic development. Eur Plan Stud 5(1):3–23
Rozenfeld H, Rybski D, Andrade J, Batty M, Stanley H, Makse H (2008) Laws of population growth. Proc Natl Acad Sci USA 105(48):18702–18707
Rozenfeld H, Rybski D, Gabaix X, Makse H (2011) The area and population of cities: new insights from a different perspective on cities. Am Econ Rev 101:2205–2225
Simmie J, Sennett J (1999) Innovative clusters: global or local linkages? Natl Inst Econ Rev 170(1):87–98
Storper M, Walker R (1989) The capitalist imperative. Blackwell, Oxford
Sturgeon T, Van Biesebroeck J, Gereffi G (2008) Value chains, networks and clusters: reframing the global automotive industry. J Econ Geogr 8(3):297–321
Swann GP, Prevezer M, Stout D (1998) The dynamics of industrial clustering. International comparisons in computing and biotechnology. Oxford university press, Oxford
Van den Berg L, Braun E, Van Winden W (2001) Growth clusters in european cities: An integral approach. Urban Stud 38(1):185–205

Publisher's Note