

# Localisation in UK Manufacturing Industries:

Assessing Non-Randomness Using Micro-Geographic Data \*

Gilles Duranton<sup>§</sup>

*London School of Economics*

Henry G. Overman<sup>‡</sup>

*London School of Economics*

Preliminary, 25 October 2001

ABSTRACT: To study the detailed location patterns of industries, we develop distance-based measures of localisation and apply them to a unique and exhaustive UK data set. In contrast to previous studies, our approach allows us to assess the statistical significance of departures from randomness. In addition, we treat space as continuous instead of using an arbitrary collection of geographical units. This avoids problems relating to scale and borders. For four-digit industries, we find that (i) only 51% of them exhibit excess-localisation at a 5% confidence level, (ii) excess-localisation takes place mostly between 0 and 50 kilometres, (iii) the degree of excess-localisation is very skewed, and (iv) industries follow broad sectoral patterns with respect to localisation. Depending on the industry, smaller establishments can be the main drivers of both excess-localisation and dispersion. Three-digit sectors show similar patterns of excess-localisation at the metropolitan level as well as a tendency to localise at the regional scale (80 – 140 kilometres).

Key words: Localisation, Clusters,  $K$ -density, Spatial Statistics.

JEL classification: C19, R12, L70.

\*Financial support from STICERD is gratefully acknowledged. Discussions with Ian Gordon, Danny Quah, Steve Redding and Tony Venables greatly helped us clarify our minds on crucial steps of the paper. Thanks to seminar participants at CEPR network meeting in Villars, Glasgow University, Sussex University and the London School of Economics. Rachel Griffith and Helen Simpson also helped us getting started with the data. Finally, we are grateful to Nick Gill for his first-class research assistance.

<sup>§</sup>Department of Geography and Environment, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. Also affiliated with the Centre for Economic Policy Research, and the Centre for Economic Performance at the London School of Economics, [g.duranton@lse.ac.uk](mailto:g.duranton@lse.ac.uk), <http://cep.lse.ac.uk/~duranton>.

<sup>‡</sup>Department of Geography and Environment, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. Also affiliated with the Centre for Economic Policy Research, and the Centre for Economic Performance at the London School of Economics, [h.g.overman@lse.ac.uk](mailto:h.g.overman@lse.ac.uk), <http://cep.lse.ac.uk/~overman>.

## 1. Introduction

At least since Alfred Marshall's (1890) *Principles*, the tendency for industries to cluster in some areas has fascinated economists and geographers alike. More recently, some of these clusters have caught the imagination of politicians and policy makers. Following Silicon Valley's success, clusters are seen by many as the magical formula for regional development. In light of this, the tendency for firms to localise (i.e., to concentrate over and above overall economic activity) raises a large number of questions about the forces at work and their welfare implications.<sup>1</sup> Furthermore, what we can learn about spatial clustering is relevant well beyond the realm of economic geography. Many explanations of spatial clustering rely on some form of external increasing returns which also figure prominently in theories dealing with international trade, industrial organisation and economic growth.

In this paper however, we would like to step back from policy and theoretical concerns and think again about the stylised facts to be explained. First and foremost, how general and how strong is the tendency for industries to cluster? We do not question Marshall's historical examples regarding the clustering of cutlery producers in Sheffield or of jewellers in Birmingham. Neither would we deny Silicon Valley's leadership in micro-electronics and software. However, it is worth asking if these are the exception rather than the rule. To inform both theory and policy, it is also crucial to know at which spatial scale this clustering occurs. In the United Kingdom (UK), the localisation of the cutlery industry in one area of Sheffield is different from that of the motor sport industry spreading over more than 100 kilometres along the Thames Valley. It is also important to know what kind of establishments, small or large, are the main driver of localisation. Finally, the scope of localisation (three- versus four- or five-digit industries) must be investigated.

Building on previous research in spatial statistics, we propose a novel way to measure localisation and answer key questions about the extent of localisation, the spatial scale(s) at which it takes place, and its sectoral scope.

In developing our approach, we have been guided by the principle that any measure of localisation should be able to meet five requirements. The first two have been recognised in the literature for a very long time. Obviously, any measure of localisation must be comparable across industries. It must also be able to control for the general tendency of manufacturing to agglomerate. For instance in the United States (US), even in the absence of any tendency towards localisation, we would expect any randomly chosen industry to have more employment in California than in Montana. This is simply because the former has a population more than 30 times as large as the latter. Most traditional measures, like location quotients or Gini indices, when properly employed, are able to satisfy these first two requirements.

Since Ellison and Glaeser (1997), it is also widely recognised that any informative measure of localisation must control for industrial concentration. To understand the distinction between localisation and industrial concentration, note that in an industry with no tendency for clustering, the location patterns of the various plants are determined by idiosyncratic factors with no systematic

---

<sup>1</sup>Following Hoover (1937), the agglomeration of a particular industry after "controlling" for that of general manufacturing is referred to as localisation.

trend. Hence, they are random to the outside observer. A relevant metaphor for the location patterns of such an industry might then be that of darts thrown randomly on a map. Because the number of plants in any industry is never arbitrarily large, such random location processes cannot be expected to generate perfectly regular location patterns. In short, unevenness does not necessarily mean an industry is localised. Unfortunately traditional measures of localisation are not satisfactory in this respect since they only measure unevenness. In the spirit of this dartboard metaphor, Ellison and Glaeser (1997) convincingly make the case that when looking at the location patterns of particular industries, the null hypothesis to be considered is one of spatial randomness conditional on both industrial concentration and the overall concentration of economic activity. The index of localisation they develop satisfies these two requirements together with comparability across industries. Taking a similar dartboard approach, Maurel and Sédillot (1999) and Devereux, Griffith, and Simpson (1999) develop alternative indices of localisation with the same properties.<sup>2</sup>

However, just like the more traditional (or "first generation") indices, these "second generation" measures still ex-ante allocate establishments or plants (i.e., points located on a map), to counties, regions or states (i.e., spatial units at a given level of aggregation). In other words, *they transform dots on a map into units in boxes*. Aggregating data this way has the obvious advantage of making computations simple but it means throwing away a large amount of information and also leads to a range of aggregation problems.

Most obviously, ex-ante aggregation restricts the analysis to only one spatial scale, be it the region, the county or the state. Exploring a different spatial scale requires another aggregation and running the analysis again. This is not very convenient and in most countries the number of available levels of aggregation is commonly limited to two or three. Besides, it is difficult to compare the results across different scales. For instance questions regarding how much industries are localised at the county level after controlling for localisation at the regional level cannot be precisely answered since existing indices are usually not easily additive across different levels of aggregation. Furthermore, most existing spatial units are defined according to administrative needs, which are usually only remotely related to economic relevance. To make matters worse, these units are often very different in population and size so that most existing aggregations tend to mix different spatial scales. For instance, analysing the localisation of industries at the level of US states involves comparisons between Rhode Island and California, which is geographically more than 150 times as big.

Another major issue is that aggregating establishments at any spatial level leads to spurious correlations across aggregated variables. The problem typically worsens as higher levels of aggregations are considered. This problem is well recognised by quantitative geographers and is known in the literature as the Modifiable Areal Unit Problem (MAUP).<sup>3</sup>

Finally, and importantly, after aggregation has taken place, spatial units are treated symmetrically so that two plants in two neighbouring spatial units are treated in exactly the same way as two plants at the two opposite ends of a country. When dealing with localised industries, this creates a downwards bias, which worsens as smaller spatial units are analysed. For instance in Britain,

---

<sup>2</sup>See Devereux *et al.* (1999) for a precise comparison between them.

<sup>3</sup>Its first recognition dates back to Yule and Kendall (1950). Cressie (1993) provides a modern discussion of it.

manufacture of machinery for textile is highly localised but the border between the East Midlands and West Midlands regions cuts the main cluster in the middle. Using UK counties would make matters even worse. We believe any good measure of localisation must avoid these aggregation problems.

The last requirement for any measure of localisation relates to its statistical significance. Given our definitions, in the absence of localisation, the patterns of location of an industry are random conditional on the location of overall manufacturing. Thus any statement about non-randomness can only be probabilistic. In this respect the literature mentioned above only offers localisation indices, i.e., measures with no indication of how statistically significant they are.

The approach we propose here satisfies these five requirements. It relies on the use of distances between observations instead of aggregating them within spatial units. We build on work by quantitative geographers on spatial point patterns (see Cressie, 1993, for a comprehensive review) that we extend to address issues of spatial scale and by constructing proper confidence intervals and confidence bounds. The basic idea in our geo-computations is to consider the distribution of distances between pairs of establishments in an industry and to compare it with that of hypothetical industries with the same number of establishments which are randomly distributed conditional on the distribution of aggregate manufacturing.

We apply this approach to a unique and exhaustive UK manufacturing data set. Four main conclusions emerge with respect to four-digit industries: (i) only 51% of them exhibit excess-localisation at a 5% confidence level, (ii) excess-localisation takes place mostly between 0 and 50 kilometres, (iii) the degree of excess-localisation is very skewed across industries, and (iv) they reveal broader patterns as industries that belong to the same industrial branches tend to have similar localisation patterns.

When looking at the location patterns of smaller establishments, a wide variety of behaviours can be found. In some industries (often related to publishing, chemicals or instruments/electric appliances), smaller establishments tend to be more localised than larger establishments. In many other industries, the opposite holds: Smaller establishments are located away from the main clusters. The analysis of the location patterns of large establishments yields findings consistent with this. In some traditional industries like wood, pulp and paper or petroleum and other non-metallic mineral products, large establishments are very localised whereas many textile, apparel and publishing industries no longer appear to be excessively-localised. In these sectors a few large establishments tend to locate away from each other and existing clusters. Regarding the sectoral scope of localisation, a range of interesting facts emerge. There are no marked differences between four- and five-digit industries, whereas three-digit sectors tend to exhibit different patterns. In particular, with three-digit sectors, excess-localisation is equally important at low geographical scale (0 – 50 kilometres) and at a more regional level (80 – 140 kilometres).

The rest of the paper is organised as follows. The next section describes our data. Section 3 outlines our methodology. Our baseline results for the localisation of four-digit UK industries are given in Section 4. These results are complemented further in Section 5 where we take into account the size of establishments. Section 6 presents further results about the scope of localisation. The last Section contains some concluding thoughts.

## 2. Data

Our empirical analysis uses exhaustive establishment level data from the 1997 Annual Respondent Database (ARD) which is the data underlying the Annual Census of Production in the UK. Collected by the Office for National Statistics (ONS), the ARD is an extremely rich data set which contains information about all UK establishments (see Griffith, 1999, for a detailed description of this data). For our purpose, we restricted ourselves to production establishments in manufacturing industries using the Standard Industrial Classification (SIC)92 industrial classification (SIC15000 to 36639) for the whole country except Northern Ireland. For every establishment, we know its complete postcode, its five-digit industrial classification, and its number of employees. Note that, when referring to SIC two-digit, three-digit, four-digit and five-digit categories, we will speak of industrial branches, sectors, industries and sub-industries respectively.

The availability of a complete postcode is particularly useful for locating plants. In the UK, postcodes units typically refer to one property, one construction, or a very small group of dwellings. Large buildings may even comprise more than one postcode.<sup>4</sup> The CODE-POINT data set from the Ordnance Survey (OS) gives spatial coordinates for all UK postcodes. This data is the most precise postcode georeferencing data available for the UK. Each Code-Point record contains information about its location, and about the number and type of postal delivery points. By merging this data together with the ARD, very detailed information about the geographical location of all UK manufacturing establishments can be generated. In so doing, we could establish the eastings and northings for around 90% of our population of establishments. These give the grid reference for any location taking as the origin a point located South West of the UK. The main problem for the remaining 10%, for which the postcode could not be matched with spatial co-ordinates, relates to postcode updates. These take place when new postcodes need to be created in a particular postcode area. Unfortunately, this could be a source of systematic rather than random errors as wrong postcodes will be reported more frequently in areas where an update recently took place. To reduce this source of systematic errors to a minimum, we checked our data against all postcode updates since 1992. This left us with 5% of establishments that could not be given a grid reference. We believe that the missing 5% of the ARD we could not match with CODE-POINT truly reflect random errors due to bad reporting and typing mistakes.

This left us with a population of 176,106 establishments. For 99.99% of them, the OS acknowledges a potential location error below 100 metres. For the remaining 26 observations, the maximum error is a few kilometres. Figures 1(a-d) map this location information for four industries. ONS disclosure rules prohibit us from naming these industries – so instead, we refer to them as industries *A* – *D* in what follows. As can be seen from the maps, industry *D* looks very localised whereas *C* is very dispersed. These two industries are extreme cases. The other two, *A* and *B*, are more representative of the typical patterns. Whether or not these last two industries are localised is far from obvious. Throughout the rest of this section, we keep using these four industries for illustrative purpose.

---

<sup>4</sup>See Raper, Rhind, and Shepherd (1992) for a complete description of the UK postcode system.



(a) Industry A



(b) Industry B



(c) Industry C



(d) Industry D

**Figure 1.** Four illustrative industries

Our analysis is conceptually simple but its implementation involves resolving some tricky technical issues. For any particular branch/sector/industry (and more generally for any partition of our population of establishments), we must first select the relevant observations. The main issue to consider here is the large number of very small establishments with one or two workers. These establishments may have different location patterns. For instance in naval constructions, there are very many small establishments of one to ten employees located inland whereas all the large establishments are located on a coast. It seems very likely that these establishments, although classified in the same industry, do not do the same thing. The first solution to this problem is to keep the data as it is on the grounds that it is best to consider the whole population of production establishments in Britain. A second possibility would be to consider an absolute size threshold and retain only establishments with employment above this threshold. However, getting rid off all establishments with less than 10 workers may be reasonable for naval construction but less so for publishing. A third possibility is then to consider a relative threshold and select establishments by decreasing size so that say 90% of employment in the sector or industry is considered. A last possibility is to weight establishments by their employment.

We implement all three approaches. In our baseline cross-industry analysis (Section 4), we consider all establishments independently of their size. In Section 5, we then consider only the largest establishments of any industry comprising at least 90% of employment. In the same section, we also weight establishments by their employment. Note that this captures a slightly different concept: weighting by employment will give a measure of the localisation of employment and no longer that of establishments.<sup>5</sup>

### 3. Methodology

#### *Constructing and smoothing K-densities*

Once we have decided what sample of plants to use, we begin by calculating the Euclidian distance between every pair of establishments in the sample. An industry  $A$  with  $n$  establishments will thus generate  $\frac{n(n-1)}{2}$  unique bilateral distances. We can then calculate the frequency for each distance level and plot the corresponding density function. Denote  $D(i,j)$ , the Euclidian distance between establishments  $i$  and  $j$  rounded to the closest hundred metres. Then define  $\delta(i,j,d)$  such that  $\delta(i,j,d) = 1$  when  $D(i,j) = d$  and  $\delta(i,j,d) = 0$  otherwise. Our un-smoothed distance density function or  $K$ -density function is then<sup>6</sup>

$$K_A(d) \equiv \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{\delta(i,j,d)}{n(n-1)}. \quad (1)$$

---

<sup>5</sup>Both types of analysis are a-priori legitimate. In his analysis of economies of scale at the metropolitan level, Henderson (1999) finds significant effects of both industry employment and of the number of establishments in the industry.

<sup>6</sup>Traditionally, spatial statisticians have used the cumulative of the  $K$ -density, the  $K$ -function. See Cressie (1993) for details. Given that a major objective of our analysis is to distinguish the spatial scale(s) at which excess-localisation takes place, the focus on the density distribution rather than its cumulative is warranted.

When doing the analysis with a relative employment cut-off point, the distance density function is defined in the same way (albeit with a smaller population for the industry). When we turn to weighting plants by their employment we analyse the distance between pairs of workers employed by different plants. Denoting the employment of firm  $i$  by  $e(i)$ , the definition of the  $K$ -density function is thus slightly different:

$$K_A^{emp}(d) \equiv \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{\delta(i,j,d)e(i)e(j)}{\sum_{i=1}^{n-1} \sum_{j=i+1}^n e(i)e(j)}. \quad (2)$$

Although the location of nearly all establishments in our data is known with a very high degree of precision, any Euclidian distance is only a proxy for the real economic distance between two establishments. The curvature of the earth is an obvious first source of systematic error. However it is easy to verify that in the UK the maximum possible error caused by the curvature of the earth is below one kilometre. The second source of systematic error is that journey times for any given distance might differ between low and high-density areas. However, there are opposing effects at work. In low-density areas, roads are fewer (so actual journey distance are much longer than Euclidean journey distance) whereas in high-density areas they are more numerous (so Euclidean distance is a good approximation to actual) but also more congested. It is unclear which effect dominates so that no specific correction was imposed. We are still left with random errors. For example, the real distance between two points along a straight road is equal to its Euclidian distance whereas that between two points on opposite sides of a river is usually well above its Euclidian counterpart. Given this noise in the measurement of distances, we decided to kernel-smooth our  $K$ -density function.<sup>7</sup> The solid lines in Figures 2(a-d) plot these density distributions for the same four industries as previously ( $A - D$ ).

### *Constructing counterfactuals*

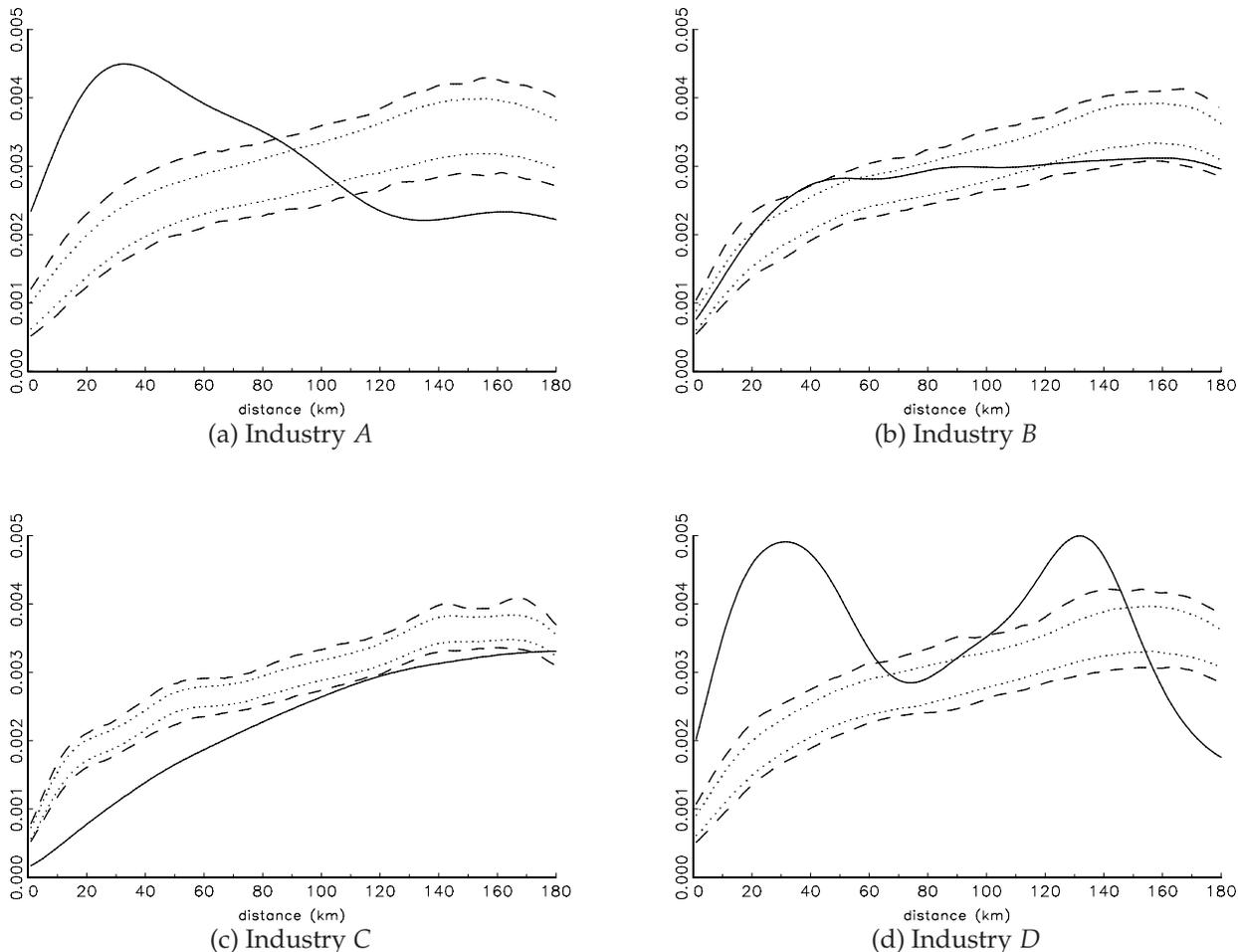
To go further, we need to decide on the relevant counterfactual that our  $K$ -density functions should be compared to. In this respect, note that the analysis of localisation is informative only to the extent that it captures interactions between establishments or between establishments and their environment. Consequently the number of firms in each industry and the size-distribution therein are taken as given. That is we take the increasing returns within the firms as given.<sup>8</sup>

Since our objective is to explore the tendencies for firms to cluster over and above general manufacturing, we need to control for the overall tendency of manufacturing to agglomerate. Furthermore, we need to allow for the fact that in Great Britain zoning and planning restrictions are ubiquitous. These restrictions imply that manufacturing cannot locate in many areas of the country (e.g., parts of Dorset and Devon, the Lake district, London's green belt, etc). To control

---

<sup>7</sup>The density plots can be interpreted as the continuous equivalent of a histogram, in which the number of intervals has been let tend to infinity and then to the continuum. All densities are calculated nonparametrically using a Gaussian Kernel with the bandwidth set as per Section 3.4.2 of Silverman (1986).

<sup>8</sup>Ultimately however, any fully satisfactory approach to these issues must treat the size of firms as endogenous. Thus, a joint-analysis of the spatial distribution of firms together with that of employment within firms is in order. Our analysis is able to deal with the spatial distribution of any subset of establishments as well as with the spatial distribution of employment but cannot directly say anything about the boundaries of the firm. This is left for future work.



**Figure 2.** K-density, local confidence intervals and global confidence bands for four illustrative industries

for overall agglomeration and the stringent regulatory framework, we consider that the set of all existing "sites",  $S$ , currently used by a manufacturing establishment constitutes the set of all possible locations for any plant.<sup>9</sup>

Thus, after calculating the  $K$ -densities and smoothing them, the third step is to generate counterfactuals by sampling from the overall population of sites in our data. For each industry we run a number of simulations. For each simulation we sample as many sites as there are establishments in the industry under scrutiny. Since establishments are created over time and since any existing site hosts only one establishment, sampling is done without replacement. Thus for any industry  $A$  with  $n$  establishments, we generate our counterfactuals  $\tilde{A}_m$  for  $m = 1, 2, \dots, M$ , where  $M$  is the number of simulations, by sampling  $n$  elements without replacement from  $S$  so that each simulation is in fact equivalent to a random reshuffling of establishments across sites. This controls for both the concentration in the industry and the overall agglomeration of manufacturing.

Each simulated industry is then processed in exactly the same way as the real industry. We

<sup>9</sup>A site is where one establishment is located – when two establishments share the same postcode, two different sites are distinguished.

first construct a raw distance density function as in equation (1). We then optimally smooth this counterfactual  $K$ -density function. In order to be able to construct proper confidence intervals, we repeat our sampling 1,000 times for each industry.<sup>10</sup>

Note that we sample by first drawing points and then calculating the set of bilateral distances associated with these points. This is rather computer-intensive. Sampling distances directly from the density of distances for the whole of manufacturing would be faster and would make it possible to calculate exact confidence intervals.<sup>11</sup> However, this short-cut amounts to treating the bilateral distances between points as independent, which they are not. To see this, consider the simplest case of three plants and imagine drawing bilateral distances between these plants from the overall distribution of distances. It is possible with a small but non-negligible probability to obtain three bilateral distances above 700 kilometres. However, it is impossible to have three plants a distance of 700 kilometres from each other in Great Britain as such an equilateral triangle simply cannot fit on its territory. It turns out that this lack of independence between distances is important beyond pathological cases and very small samples. By running a very large number of simulations (10,000) for a few medium-sized industries, we found that the differences between point-generated  $K$ -density functions and distance density functions generated directly from distances are too large for us to be able to sample distances directly.<sup>12</sup>

### *Local confidence intervals*

The fourth step is to calculate local confidence intervals. We consider all distances between 0 and 180. This threshold is the median distance in the set of distances between all pairs of manufacturing establishments. Any 'abnormally' high values for the distance density function,  $K(d)$ , for  $d > 180$  could in principle be interpreted as excess-dispersion but this information is redundant if we consider both *lower* and upper confidence bounds for  $d < 180$ . Hence we restrict our analysis to the interval  $[0,180]$ . For each kilometre in this interval, we rank our simulations in ascending order and select the 5<sup>th</sup> and 95<sup>th</sup> percentile to obtain a lower 5% and an upper 5% confidence interval that we denote  $K_5(d)$  and  $K_{95}(d)$  respectively. When for industry  $A$ ,  $K_A(d) > K_{A,95}(d)$ , this industry is said to exhibit *excess-localisation at distance  $d$* . Symmetrically, when  $K_A(d) < K_{A,5}(d)$ , this industry is said to exhibit *excess-dispersion at distance  $d$* .<sup>13</sup> We can also define an index of excess-localisation

$$\gamma_A(d) \equiv \max(K_A(d) - K_{A,95}(d), 0) , \quad (3)$$

---

<sup>10</sup>We also repeated our simulations 2,000 and 10,000 times for a few industries. We ended up with very similar confidence bands.

<sup>11</sup>The density at each level of distance could then be treated as the result of repeated binomial draws for which exact confidence intervals are readily available.

<sup>12</sup>For an industry with around 200 establishments, we found that the confidence intervals were about twice as large when drawing distances directly.

<sup>13</sup>Excess-dispersion here is precisely defined as having fewer establishments at distance  $d$  than randomness would predict. In other words the distribution of an industry with excess-dispersion is "too regular". A direct analogy can be made with random draws of zeros and ones under equal probability. A string of 10 zeros out of ten draws is rather unlikely and akin to our concept of excess-localisation. Alternatively, five zeros alternating with five ones is also fairly unlikely and this excessive regularity is interpreted in a geographical context as excess-dispersion.

as well as an index of dispersion

$$\psi_A(d) \equiv \max(K_{A,5}(d) - K_A(d), 0). \quad (4)$$

To reject the hypothesis of randomness at distance  $d$  because of excess-localisation (resp. dispersion), we only need  $\gamma_A(d) > 0$  (resp.  $\psi_A(d) > 0$ ). The exact value of these two indices does not matter. However, the indices do indicate how much excess-localisation and excess-dispersion there is at any level of distance.

Graphically, excess-localisation (resp. dispersion) is detected when the  $K$ -function of one particular industry lies above (resp. below) its local upper (resp. lower) confidence interval. The two dotted lines in Figures 2(a-d) plot these local confidence intervals for our four illustrative industries. For instance, industry  $D$  exhibits excess-localisation for every kilometre from 0 to 60 whereas  $C$  exhibits excess-dispersion over the same range of distances.

### *Global confidence bands*

Note that the calculation of  $\gamma_A(d)$  and  $\psi_A(d)$  only allows us to make local statements (i.e. at a given distance) about departures from randomness. These local statements however do not correspond to statements about the global location patterns of an industry as captured by its  $K$ -density function. Even a completely randomly distributed industry will exhibit excess-dispersion or excess-localisation for some level of distance with quite a high probability. To see this, recall that there is 5% probability an industry shows excess-localisation for each kilometre, so that the probability of this happening for at least one kilometre among 180 is quite close to 1 even when we account for the fact that smoothing induces some autocorrelation in the  $K$ -density estimates across distances.

Our last step is thus to construct global confidence bands so that statements can also be made about the overall location patterns of an industry. There are infinitely many ways to draw bands such that no less than 95% of a series of randomly generated  $K$ -density functions lie completely within the bands. The restriction we impose here is standard: we choose identical local confidence bounds at all levels of distance such that the global confidence level is 5%. That is, deviations by randomly generated  $K$ -densities are equally likely across all levels of distances to make the confidence bands neutral with respect to distances.

Here, we cannot use the standard Bonferroni method which considers the local confidence interval  $y$  such that in our case  $(1 - y)^{181} = 5\%$  since it ignores the positive autocorrelation across distances and would thus give us confidence bands that are too wide. Instead, the solution is to go back to our simulated industries and look for the upper and lower local confidence intervals such that, when we consider them across all distances between 0 and 180, only 5% of our randomly generated  $K$ -density functions hit them. The local confidence levels associated with these global confidence bands will of course be below 5%.

Even with 1000 simulations however, there may not be any local confidence level such that we can capture exactly 95% of our randomly generated  $K$ -density functions. This problem is solved easily by interpolating. The second worry is that we may need to consider the local 99.9<sup>th</sup> or even the 100.0<sup>th</sup> percentile to get a global 5% confidence band. The variance for these

randomly generated extreme bounds (i.e., the extreme or the second extreme value in the simulations) is potentially quite high which means a low degree of precision for the corresponding bands. However, excess-localisation and excess-dispersion are correlated across distances (be it only because of optimal smoothing). For most of our industries, this implies that the local confidence level such that 5% of our randomly generated industries deviate is typically around 99%, i.e., the 10<sup>th</sup> extreme value, for which the variance is much lower.

Denote  $\bar{K}_A(d)$  the upper global confidence band of industry  $A$ . This band is hit by 5% of our simulations between 0 and 180 kilometres. When  $K_A(d) > \bar{K}_A(d)$  for at least one  $d \in [0,180]$  this industry is said to exhibit *global excess-localisation*. Turning to global excess-dispersion, note first that by construction, our distance densities must sum to one. Thus an industry which is very excessively localised at short distances (say the city level) can show excess-dispersion at larger distances. Similarly, an industry which is excessively localised at medium distances (say the regional level) can show excess-dispersion at shorter distances. In other words, for strongly localised industries, excess-dispersion is just an implication of excess-localisation. This discussion suggests the following definition: The lower global confidence band of industry  $A$ ,  $\underline{K}_A(d)$ , is such that it is hit by 5% of the randomly generated  $K$ -density functions that are *not* excessively localised. An industry is then said to exhibit *global excess-dispersion* when  $K_A(d) < \underline{K}_A(d)$  for at least one  $d \in [0,180]$  and the industry does not exhibit excess-localisation. As before, we can define:

$$\Gamma_A(d) \equiv \max(K_A(d) - \bar{K}_A(d), 0), \quad (5)$$

an index of global excess-localisation and

$$\Psi_A(d) \equiv \begin{cases} \max(\underline{K}_A(d) - K_A(d), 0) & \text{if } \sum_{d=0}^{d=180} \Gamma_A(d) = 0, \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

an index of global excess-dispersion.

Graphically, global excess-localisation is detected when the  $K$ -function of one particular industry lies above its upper confidence band. Global excess-localisation is detected when the  $K$ -density of one particular industry lies below the global lower confidence band and never lies above the upper confidence band. For our four illustrative industries, the global confidence bands are represented by the two dashed lines in Figures 2(a-d). For instance, industry  $D$  exhibits global excess-localisation whereas  $C$  exhibits excess-dispersion. Industry  $B$  shows neither global excess-localisation nor dispersion while  $A$  shows excess-localisation and thus by definition no excess-dispersion even though its  $K$ -density does go beneath the global lower confidence band.

In summary, we use the density of distances between pairs of establishments in an industry from 0 to 180 kilometres, the latter being the median distance between pairs of establishments across UK manufacturing. We then compare this distribution to counterfactuals generated by a random re-shuffle of establishments across existing manufacturing sites.

### *Examples and interpretation*

To understand better what these tests capture, let us consider a few examples. Take first an industry with a cluster of plants around London like industry  $A$ . This cluster in London implies a high

density for distances between 0 and 60 kilometres and this industry thus shows excess-localisation between these distances. Consider now an industry like industry  $D$  with a cluster of plants around Manchester and another around Birmingham. The large number of establishments located close to each other in both Manchester and Birmingham still implies excess-localisation between 0 and 40 – 50 kilometres. Furthermore, Manchester and Birmingham are quite close to each other so that there is also excess-localisation for distances between 100 and 140 kilometres. Had the second cluster been in London instead of Birmingham, this second peak of distance would not show-up in our analysis as Manchester and London are distant of more than 180 kilometres from each other. A multiplicity of peaks in our distance density distribution thus indicates a multiplicity of clusters close to each other.

Consider now the more contrived case of an industry located mostly in one region, say the South East but with regularly dispersed plant (in order to serve local markets for instance). Such an industry would be both locally dispersed at short distances, but also localised at higher levels of distances (capturing the fact that it is present in only one region). Conversely an industry with nearly all its plants in one city (like industry  $A$ ) is localised for low distances and locally dispersed for distances above 100 kilometres since the  $K$ -density function is a probability distribution function and thus must sum up to one across distances.

These examples should help clarify our terminology outlined above. To reiterate: An industry is said to be locally excessively localised (resp. dispersed) at distance  $d$  if it goes outside the upper (resp. lower) local confidence band at distance  $d$ . We will call an industry globally excessively localised if it goes outside the global upper confidence band somewhere in the range 0 – 180km. We call an industry globally excessively dispersed only if it goes below its global lower confidence band at some distance *and* it does not show excess-localisation. When this second condition is not satisfied, excess-dispersion is a consequence of excess-localisation at some distance between 0 and 180km.<sup>14</sup>

Finally, note that a cluster of establishments is more likely to be found in the Midlands, which has a lot of manufacturing than, say, in Northern Scotland which has very little. Our analysis does not directly deal with this, since as a first step we want to be able to make statements about patterns in particular industries in relation to general manufacturing and not about the patterns of specialisation of some particular local economies, here or there. The analysis of specialisation is conceptually distinct from that of localisation and as it requires different tools it is beyond the scope of this paper.

Before presenting our results, it worth considering what we can and cannot learn from this type of analysis. Excess-localisation or dispersion is compatible with any explanation of clustering that relies on some form of external effect but also with any explanation based on fixed natural endowments. Like Ellison and Glaeser (1997), we think that it is helpful to be able to make statements about the location pattern of an industry without knowing the right mix of external economies and natural endowments that led to its pattern.

---

<sup>14</sup>Notice that with this convention, it is possible that we may incorrectly classify a couple of pathological industries like our example above. In fact, this never happens for the industries in our sample. In addition, the classification has the distinct advantage of partitioning the group of industries in a way that corresponds to most peoples conception of excessive localisation and dispersion.

This said, the scales at which deviations from randomness take place may provide highly suggestive evidence of some underlying causes. Ellison and Glaeser (1999) make a very broad list of the kind of natural endowments that may matter for manufacturing location. Their list includes the prices of various forms of energy, some agricultural variables, population density and some location variables such as the proximity to a coast, etc.<sup>15</sup> In the UK, none of these "natural" endowments imposes tight constraints on the location of production establishments at small spatial scales (i.e., within a few kilometres).<sup>16</sup> Moving from first to second nature explanations, localisation economies can be divided into those accruing from economising on the transport of physical goods (input-output linkages), on the transport and commuting of workers (labour market pooling) and on the transport of ideas (technological externalities). Here the literature suggests that the range for technological externalities is largely limited to a few kilometres, the distance for frequent word-of-mouth interactions. The range for the labour market pooling externalities is surely the same as that of commuting by workers dispersed around a plant (i.e., maybe 50 kilometres – the urban and metropolitan scale), whereas the range for input-output linkages is probably much larger. Hence we can use these extra identifying assumptions to draw further conclusions about our results.

#### 4. Baseline results for four-digit industries

In this section we describe the patterns of localisation of UK four-digit industries using the complete population of plants.

##### *How many industries deviate and where*

Out of a total of 234 industries, 204 deviate locally at some distance over the range.<sup>17</sup> Correcting for global confidence bands as outlined in Section 3 leads us to conclude that 180 industries differ significantly from randomness at the 5% level of significance. The detailed breakdown is as follows. We find that 120 industries, that is 51% of them, exhibit excess-localisation whereas 60 industries, that is 26% of them, exhibit excess-dispersion, and 54, that is 23% of them do not deviate significantly from randomness. From our results, excess-localisation is not as widespread as earlier studies led us to believe whereas dispersion seems much more prevalent (all the more so given our restrictive definition of dispersion). Devereux *et al.* (1999) on comparable UK data, Ellison and Glaeser (1997) on US data, or Maurel and Sédillot (1999) on French data find that between 75 and 95% of industries are localised according to the Ellison-Glaeser (EG) index and less than 15% are dispersed.<sup>18</sup>

---

<sup>15</sup>See Rosenthal and Strange (2001) for a further discussion of this and a more restrictive list of natural endowments variables.

<sup>16</sup>Especially since extractive activities are not considered here.

<sup>17</sup>Five industries (15430, 15920, 15950, 17140, and 2652) with 10 plants or less are not considered here because of their very small size.

<sup>18</sup>Note however that these analyses deal with the localisation of employment and not that of plants. See below for a comparison between our approach when weighting by employment and the EG index using the same data.

Fraction of four-digit industries localised at:				
5km	5km only	5 and 30km only	5 and 150km only	5, 30 and 150km
37.2	6.4	21.4	0.4	9.0
30km	30km only	30 and 150km only		
36.8	5.1	1.3		
150km	150km only			
16.7	5.9			

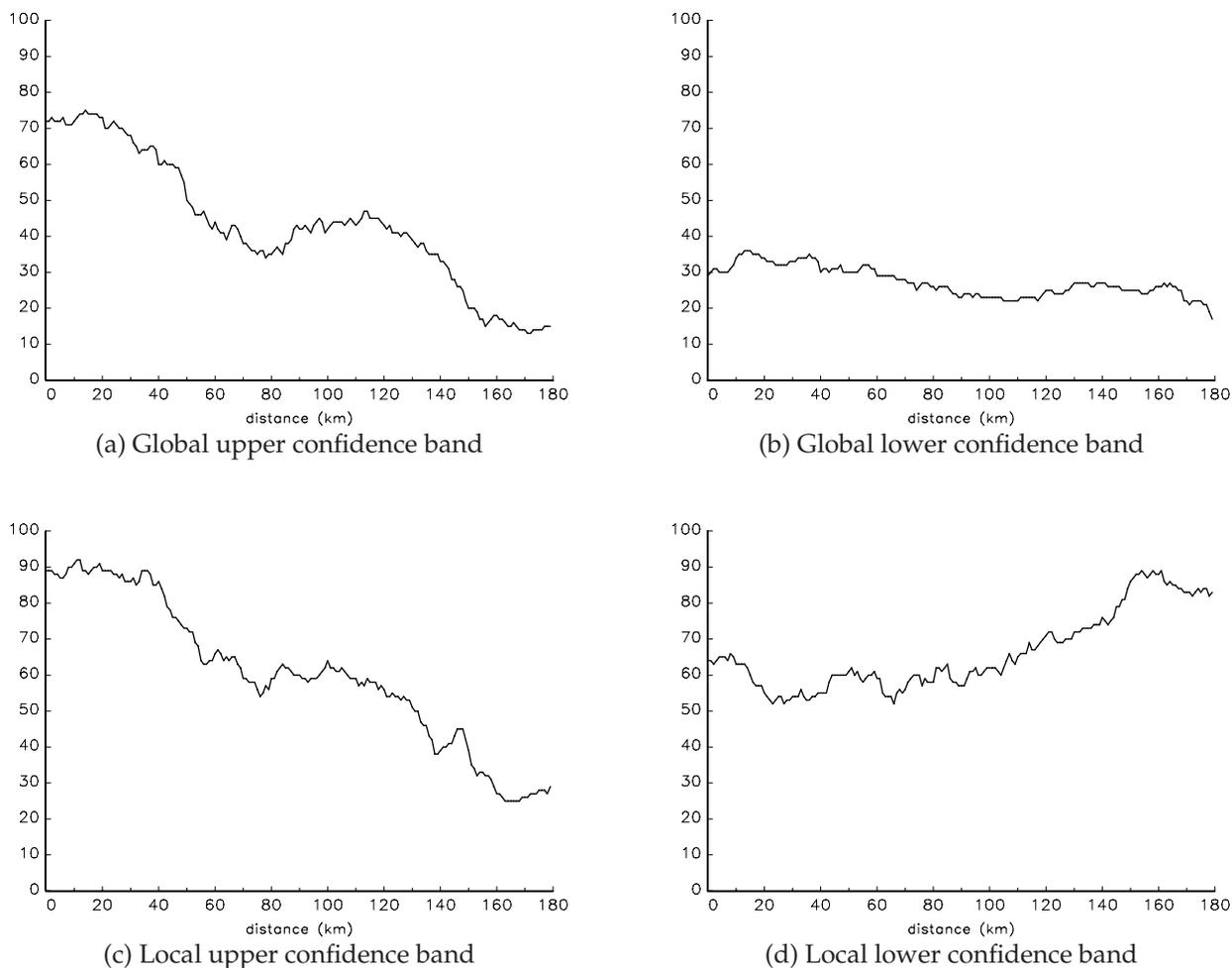
**Table 1.** Excess-localisation at three key thresholds for four-digit industries

To go further and to look at scale issues, Table 1 considers the fraction of industries which show local excess-localisation at three critical thresholds (5, 30 and 150 kilometres). Note that a majority of industries who deviate for any of these three threshold tend to do it for both 5 and 30 kilometres. These results are confirmed when looking more broadly at cross-industry patterns. Figure 3 shows the number of industries with excess-localisation and excess-dispersion (local and global) for each level of distance. Both local and global confidence bands show roughly similar patterns for excess-localisation. At low distances, a significant proportion of industries are excessively localised. The number of excessively localised industries is on a high plateau between 0 and 40 kilometres, then falls sharply with distance up to around 80 kilometres and then begins to rise again with a second and lower peak between 100 and 120 kilometres. These findings regarding the scale(s) at which excess-localisation takes place are markedly different from those of previous studies based on the EG index. They find that in the US industrial localisation is persistently stronger for states than counties and stronger for counties than ZIP-codes (Ellison and Glaeser, 1997, Rosenthal and Strange, 2001).<sup>19</sup> Excess-dispersion shows very different patterns. Global excess-dispersion tends to occur equally across all distances. In contrast, local excess-dispersion (which includes industries that are localised) tends to rise slowly with distance reflecting the ‘reflection’ problem that we discussed earlier.

Although these figures tell us how many industries deviate from randomness at any given distance, they are not informative about the extent of the deviations from randomness. We can base a measure of excess-localisation at any given distance on the index of excess-localisation  $\Gamma_A(d)$  defined in equation (6). We proceed as follows. Denote by  $I_4$  the set of industries at the four-digit level. Then construct  $\Gamma(d) \equiv \sum_{A \in I_4} \Gamma_A(d)$ , by summing the index of excess-localisation across industries for each level of distance. Similarly, we can construct a measure of the extent of cross-industry excess-dispersion,  $\Psi(d) \equiv \sum_{A \in I_4} \Psi_A(d)$  using the index of excess-dispersion  $\Psi_A(d)$ . We should be clear - these measures have no formal interpretation, but are useful to summarise our results. Figure 4 reports both measures for the 234 industries. Note that the measures are directly comparable across distances, but not across the two figures.<sup>20</sup> It is immediately apparent that the extent of localisation is much greater at small distances than large distances. In contrast, as before, dispersion does not show any marked pattern.

<sup>19</sup>See below for a detailed comparison on UK data.

<sup>20</sup>This is because excess-localisation is unbounded from above whereas excess-dispersion is bounded from below by zero.



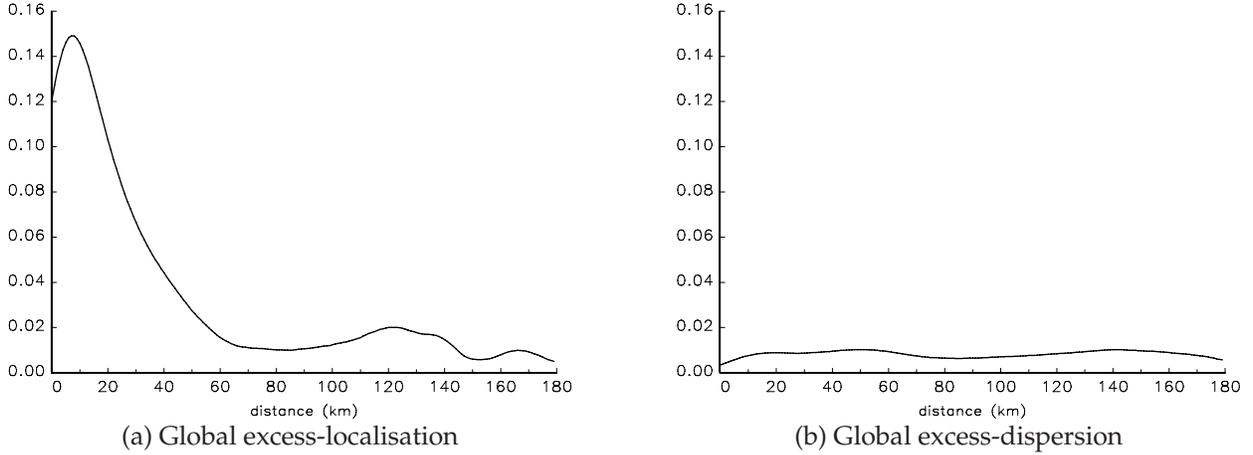
**Figure 3.** Number of four-digit industries with excess-local/global localisation and dispersion

The important conclusion we draw here is that excess-localisation tends to take place mostly at the urban/metropolitan scale. The high levels of excess-localisation between 0 and 40 kilometres point at an explanation based on both labour market and word-of-mouth externalities. The second peak of excess-localisation after 100 kilometres is more difficult to interpret with respect to input-output linkages as it can be the result of single clusters arising at a larger geographical scale or that of small clusters that happen to be close to each other – and such a pattern may or may not be random. Cursory inspection of industries with excess-localisation between 100 and 150 kilometres suggests that both patterns are indeed present.<sup>21</sup>

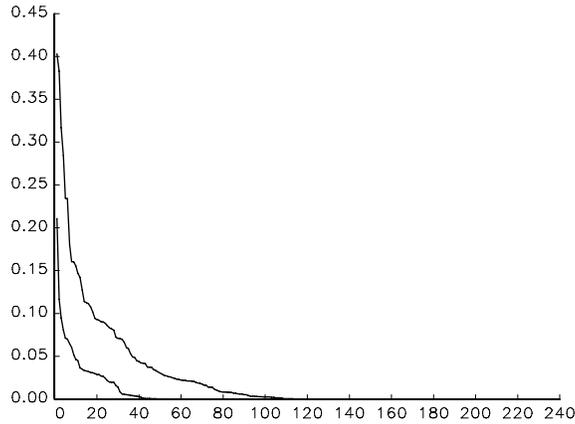
### *Differences between industries*

We now turn to the examination of differences between industries. We start by constructing a measure of the extent to which different industries deviate from randomness. Proceeding as before,

<sup>21</sup>For instance, the double peaked  $K$ -density function for industry  $D$  is caused by three small-scale clusters in Manchester, Birmingham and London with two of the bilateral distances between these cities being around 120 kilometres.



**Figure 4.** Amount of global excess-localisation and dispersion in four-digit industries



**Figure 5.** Distribution of local/global excess-localisation/dispersion across four-digit industries

for each industry  $A$  we can define the following cross-distance indices:  $\Gamma_A \equiv \sum_{d=0}^{180} \Gamma_A(d)$ , and  $\Psi_A \equiv \sum_{d=0}^{180} \Psi_A(d)$ . Respectively, these measures are the sum for each industry of the index of global excess-localisation and dispersion across all levels of distance. To illustrate the variations in industry outcomes, we rank industries by decreasing order along these quantities and plot them in Figure 5. The upper line is the measure of the extent of excess-localisation, the lower that of excess-dispersion.<sup>22</sup> As is immediately clear, there are a few industries that show very high excess-localisation or excess-dispersion, but the majority of industries do not see such extreme outcomes. This highly skewed distribution of excess-localisation confirms previous findings (Ellison and Glaeser, 1997, Devereux *et al.*, 1999, Maurel and Sédillot, 1999).

To give some idea of the reality underlying Figure 5, Table 2 lists the 10 industries with most

<sup>22</sup>Recall that we do not include industries that are dispersed and localised when calculating the extent of excess-dispersion

sic92	Industry	$\Gamma$ or $\Psi$
Most excess-localised		
2214	Publishing of sound recordings	0.403
1711	Preparation and spinning of cotton-type fibres	0.383
2231	Reproduction of sound recordings	0.317
1760	Manufacture of knitted and crocheted fabrics	0.284
1771	Manufacture of knitted and crocheted hosiery	0.234
1713	Preparation and spinning of worsted-type fibres	0.233
2861	Manufacture of cutlery	0.181
1822	Manufacture of other outerwear	0.160
2211	Publishing of books	0.160
1824	Manufacture of other wearing apparel and accessories n.e.c.	0.156
Most excess-dispersed		
1520	Processing and preserving of fish and fish products	0.211
3511	Building and repairing of ships	0.117
1581	Manufacture of bread, fresh pastry goods and cakes	0.095
2010	Saw milling and planing of wood, impregnation of wood	0.081
1752	Manufacture of cordage, rope, twine and netting	0.071
1551	Operation of dairies and cheese making	0.061
3615	Manufacture of mattresses	0.052
2030	Manufacture of builders' carpentry and joinery	0.046
1571	Manufacture of prepared feeds for farm animals	0.045
1596	Manufacture of beer	0.037

**Table 2.** Most localised and most dispersed four-digit industries

excess-localisation and the 10 with most excess-dispersion.<sup>23</sup> Interestingly, more than a century after Marshall (1890), Cutlery (SIC2861) is still amongst the most localised industries. Furthermore, six textile or textile-related industries are also in the same list together with three media based industries. These highly localised industries are fairly exceptional. The mean industry (after ranking industries by their degree of localisation) by contrast is barely more localised than if randomly distributed. On the other hand, it is mostly food-related industries together with industries with high transport costs or high dependence on natural resources that show excess-dispersion.

Finally, it is also interesting to notice that for many (two-digit) industrial branches, related industries tend to follow similar patterns. Table 3 breaks down excess-localisation of industries by branches. Out of 22 branches, only six show heterogenous patterns for their industries. For instance nearly all Food and Drink industries (SIC15) or Wood, Petroleum, and Mineral industries (SIC20, 23, and 26) are not localised. By contrast, most Textile, Publishing, Instrument and Appliances industries (SIC17 – 19, 22, and 30 – 33) are localised. The two main exceptions are Chemicals (SIC24) and Machinery (SIC29). In these two branches, however, the more detailed patterns are telling. Chemical industries related to dispersed industries such as Fertilisers (SIC2415)

<sup>23</sup>We exclude two industries 17250 Manufacture of other textiles, and 29320 Manufacture of other agricultural and forestry machinery, from the list of most dispersed industries. It is not clear whether these are actually dispersed, or whether these industries are aggregates of other well defined clustered sub-industries.

Two-digit branch	Number of four-digit industries	no. global excess-localisation $\leq 60$ km	no. global excess-localisation $> 60$ km
Food products and beverages	30	1	0
Tobacco products	1	1	0
Textiles	20	15	8
Wearing apparels, dressing, etc	6	6	3
Tanning and dressing of leather, footwear	3	3	3
Wood and products of wood, etc	6	0	0
Pulp, paper and paper products	7	2	1
Publishing, printing and recorded media	13	13	8
Coke, refined petroleum products	3	0	0
Chemical and chemical products	20	9	9
Rubber and plastic products	7	1	3
Other non-metallic mineral products	24	4	2
Basic metals	17	12	10
Fabricated metal products	16	10	12
Other machinery and equipment	20	6	10
Office machinery and computers	2	2	2
Electrical machinery	7	2	5
Radio, Televisions and other appliances	3	3	3
Instruments	5	3	4
Motor vehicles, trailers, etc	3	1	3
Other transport equipment	8	1	1
Furniture and other products	13	4	5
Aggregate	234	99	92

**Table 3.** Excess-localisation by two-digit branch

or Organic chemicals (SIC2420) are dispersed whereas those related to localised industries like Basic pharmaceuticals (SIC2441) or Preparation of recorded media (SIC2465) are themselves very localised. The same holds for machinery: Agricultural machinery (SIC2932) is very dispersed like most agriculture related industries, whereas Textile machinery (SIC2954) is very localised like most textile industries.

## 5. Establishment size and UK localisation

Four main conclusions emerge so far: (i) only 51% of industries show excess-localisation, (ii) excess-localisation mostly takes place at the urban and metropolitan scale, (iii) deviations from randomness are very skewed across industries and (iv) they seem to reveal broader patterns where industries that belong to the same branch tend to have similar localisation patterns. These findings may be driven by particular types of establishments or particular sectoral definitions. To evaluate how robust they are and to gain insights about the size of localised establishments and the scope

of localisation, we need to replicate our analysis with alternative samples of plants and alternative sectoral definitions. This section deals with size issues; questions relating to scope are examined in the next section.

Note that the issue of size is particularly crucial as firm-size distributions are very skewed in most industries. In our population of plants, 36% of establishments employ two persons or less and represent only a very small fraction (2.4%) of total manufacturing employment. The issue of firm-size is also important from a policy perspective. Policies encouraging dispersion are not likely to be very successful if it is only small establishments that can be dispersed, whereas clustering policies might be more difficult to implement if it is only large establishments that cluster. Finally, the type of establishments, big or small, that cluster or disperse is potentially very informative about the relevance of particular theories.

#### *Four-digit industries when censoring the smallest plants*

In this section we repeat our baseline analysis after censoring for the smallest plants in each industry. There are two reasons for doing this. First it checks the robustness of our results to aggregation errors introduced by the classification system. In certain industries (say shipbuilding to take our earlier example) small plants might not do the same thing as large plants. Second, in industries where aggregation errors do not occur, it is still possible that the location behaviour of small plants differs from the location behaviour of large plants.<sup>24</sup> As discussed in Section 2 imposing the same absolute threshold across industries is problematic when average plant sizes differ substantially across industries. Instead, we use a relative threshold obtained by ranking plants by decreasing size and then selecting a cut-off size such that those plants account for 90% of employment in the industry. Once we have the cut-off plant size, we redo our analysis of Section 4 using the sample of plants that meet this cut-off criteria for the same 234 four-digit industries.<sup>25</sup>

The first key finding is that after censoring the smallest plants only 41% of industries (against 51% in the baseline simulations) show any amount of global excess-localisation. At the same time however, the amount of excess-localisation,  $\Gamma(d)$ , is above that in the baseline simulations despite there being fewer establishments and thus larger confidence bands. Hence, in some industries, excess-localisation gets stronger when smaller establishments are ignored whereas in others small plants are the main drivers of localisation.

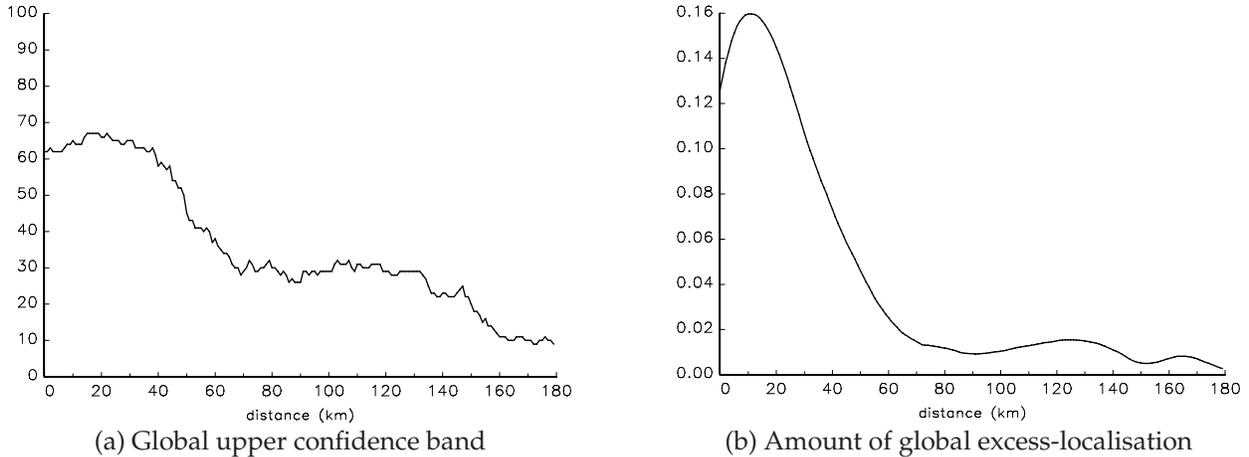
Turning to the spatial scales at which the deviations take place, note that they are the same as before. As can be seen from Figure 6, the number of localised industries is large between 0 and 40 kilometres and then decreases to reach a low plateau after 60 kilometres. From the same figure, the amount of excess-localisation,  $\Gamma(d)$ , also follows patterns similar to those observed when all plants are considered.

Industries continue to show very different location patterns and a very skewed distribution of both excess-localisation and dispersion. When comparing these results with our baseline across

---

<sup>24</sup>Note that the ability to focus on and separately analyse any subset of establishments in a consistent way is one of the strengths of our approach.

<sup>25</sup>Of course, the sample of plants will usually account for more than 90% of employment once we include all plants that are at least the cut-off size.



**Figure 6.** Global excess-localisation when censoring for the smallest establishments

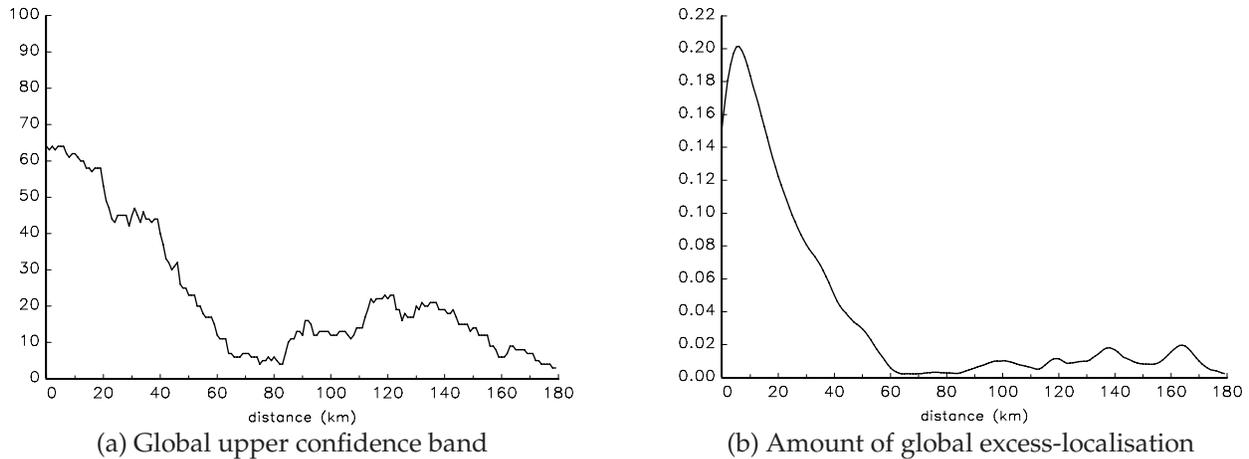
industrial branches, we note that the declines in excess-specialisation are concentrated in Publishing (SIC22), Chemicals (SIC24), Computers (SIC30), and Radios and TVs (SIC32). This is evidence that in these branches, excess-localisation is driven by small establishments who tend to cluster.

On the other hand, in a few textile industries, that are among the most localised industries when treating all plants symmetrically, the coefficient of excess-localisation,  $\Gamma$ , increases by more than 50% when the smallest plants are ignored. This is observed not only in Textile and related industries (SIC17 – 19) but also in Petroleum and Other non-metallic mineral products (SIC23 and 26). In these industries, smaller establishments are more dispersed. This finding is confirmed when looking at the patterns of excess-dispersion. Only 18% of industries (against 26% in the baseline) exhibit global excess-dispersion when censoring for the smallest establishments. Overall, the location patterns of small establishments seem to vary a lot across industries but tend in general to be more extreme with stronger tendencies towards either localisation or dispersion.

#### *Four-digit industries when weighting for employment*

We can complete our analysis of location patterns by establishment size by weighting them by their employment as in equation (2). As made clear above, censoring for the smallest establishments can shed light on their location patterns when comparing the results with those obtained for the whole population of plants. By contrast, weighting establishments by their employment is mostly helpful for studying the location patterns of the larger establishments. The reason is that according to equation (2), the distance between two establishments with 100 employees each is given 100 times the weight of the distance between two establishments with 10 employees each.

Turning to the results, note first that only 43% of four-digit industries (against 51% in the baseline) exhibit some global excess-localisation. The scales at which these deviations occur are very similar to what has been observed before. According to Figure 7, when weighting for employment excess-localisation is even more strongly biased in favour of short distances (below 50 kilometres). At the same time, as in the previous analysis, the total amount of excess-localisation



**Figure 7.** Global excess-localisation when weighting establishments by their employment

is higher than in the baseline simulations for distances below 50 kilometres (see again Figure 7).<sup>26</sup> Hence when weighting for employment, fewer industries show excess-localisation but those who do deviate more strongly from randomness. Results with respect to excess-dispersion are as in the baseline simulations – 26% of all industries are excessively dispersed.

Industries are still highly heterogeneous with respect to their localisation/dispersion behaviour. Further interesting patterns emerge when we compare these results with the baseline results across industrial branches. First, in Apparels (SIC18), Tanning (SIC19), Publishing (SIC22), Chemicals (SIC24), Radio and TVs (SIC32), and Instruments (SIC33) excess-localisation is far less prevalent than in the baseline. For instance, all six industries in Apparels show excess-localisation in the initial results whereas only one still shows excess-localisation when weighting establishments by size. Similarly only seven in 13 publishing industries remain excessively localised when weighting by employment instead of all of them when treating all plants symmetrically.

In other branches like Wood products (SIC20), Pulp and paper (SIC21), Petroleum (SIC23), and Other non-metallic mineral products (SIC26), the exact opposite happens. For instance out of the seven sectors that make up the Pulp and paper industry, none shows excess-localisation in the baseline while six in seven do when weighting by employment. When weighting by establishment size, the three most localised industries are Builders' carpentry and joinery (SIC2030), Tobacco (SIC1600), and Corrugated paper and paper boards (SIC2121), i.e., industries dominated by a few large establishments located close to each other.<sup>27</sup>

These findings are fully consistent with those obtained when censoring for smaller firms. However they stand in contrast with those obtained in the literature: Holmes and Stevens (2000) examine differences between small and large plants in US manufacturing using the Ellison and Glaeser (1997) approach. They find that large plants are very significantly more localised than

<sup>26</sup>Although, we cannot strictly speaking compare the  $\Gamma(d)$ s in this analysis with those of the initial analysis, note that in the  $K$ -density in (2) both the numerator and the denominator are weighted in the same fashion.

<sup>27</sup>Interestingly Builders' carpentry and joinery (SIC2030) is also one of the most dispersed industry in our baseline.

small plants. In the UK, we observe this pattern only for a few industrial branches. This could stem from differences between the UK and the US in their location structure or from the methodology they use.

Before turning to a detailed comparison between our approach and that of Ellison and Glaeser (1997), it must be emphasised that taking plant size explicitly into account in our approach reinforces the four main conclusions obtained so far. Excess-localisation is detected in at most half of the industries. Deviations still occur at a scale of 0 to 50 kilometres. There is still a lot of cross-industry heterogeneity with respect to localisation and dispersion. This is compounded by cross-industry differences in location patterns between small and large establishments. Finally we still observe broad patterns by industrial branches which also materialise with respect to the clustering of small vs large establishments.

### *Comparison with other approaches*

The index developed by Ellison and Glaeser (1997) is equal to

$$G_A \equiv \frac{g_A - H_A}{1 - H_A}, \quad (7)$$

where  $H_A \equiv \sum_j x_A(j)^2$  is the Herfindhal index of industrial concentration for industry  $A$ ,  $x_A(j)$  is the share of employment of establishment  $j$  in industry  $A$ ,  $g_A$  is a raw localisation index equal to

$$g_A \equiv \frac{\sum_i (s_A(i) - s(i))^2}{1 - \sum_i s(i)^2}, \quad (8)$$

$s_A(i)$  is the share of area  $i$  in industry  $A$ , and  $s(i)$  the areas share in total manufacturing. Any positive value for this index is interpreted as excess-localisation. Ellison and Glaeser (1997) also argue that a value between 0 and 0.02 signals weak excess-localisation and anything above 0.05 is interpreted as a strong tendency to localise. To compare with our methodology, we apply this index to the 120 postcode areas of Great Britain using the total population of plants. Note that postcode areas are on average smaller and less populated than US states but larger than US counties.

As this index not only controls for the lumpiness of plants but also for their size distribution, it is best compared to our results when plants are weighted by employment. The first crucial difference is that only 6% of industries have a negative EG index whereas 18% of industries were found to be significantly dispersed with our approach when weighting for employment. Conversely, the EG index implies that 94% of industries are localised whereas the equivalent figure with our approach is only 43%.<sup>28</sup> When ranking industries by decreasing EG index, we need to choose a cut-off value of 0.015 to ensure that 43% of industries are defined as localised. Thus, for UK manufacturing plants Ellison and Glaeser (1997)'s definition of weakly localised (EG index less than 0.02) industries mostly defines industries whose location patterns are not *significantly* different from randomness. Furthermore, in addition to the differences in terms of number of localised industries, individual industries show very different outcomes between the two measures. For

---

<sup>28</sup>Note that the mean value of the EG index across 234 industries is 0.034 and the median is at 0.011. These figures are above their corresponding values for US counties but below those of US states according to Ellison and Glaeser (1997)'s calculations.

example, there is no overlap between the 10 most localised industries with our approach and the 10 industries with the highest EG index. Worse, our 10 most localised industries have an average rank above 100 with the EG index.<sup>29</sup> These differences hold more generally – the Spearman rank-correlation between the overall rankings delivered by the two methods is insignificant and very low at 0.04.

The correspondence is much better when we take our 10 most localised industries after censoring for the smallest plants. The two methods agree on 4 out of the 10 most localised industries and the average EG ranking of our 10 most localised industries is below 30. Ignoring three publishing industries which we find to be very localised, the EG index average ranking of our 10 most localised industries would even fall below 10. These three media/publishing industries are quite interesting. Studying the maps for these sectors, we see that they are indeed very localised around London. The EG index cannot capture this as the London region is divided into many postcode areas which are then treated as completely unrelated entities in the calculation of the index.<sup>30</sup> More generally, the Spearman rank-correlation between the EG index and our ranking when censoring for the smallest plant is very significant and much higher at 0.40.

We believe these comparisons highlight a number of advantages of our approach. First, allocating dots on a map to units in a box introduces border effects that bias downwards existing measures of localisation. On its own, this would tend to increase the number of localised industries identified by our methodology which avoids this border effect. However, offsetting this is the fact that ignoring the significance of departures from randomness bias existing measures of localisation. Our methodology removes border effects and allows for significance, and our results show that the latter effect dominates. Second, the relevant geographical scales for localisation emerge naturally from our analysis because we do not need to arbitrarily define the size of units *ex-ante*. Existing indices such as the EG are calculated over only one partition of space, whereas we have shown that different industries localise at different spatial scales. This problem is compounded by the fact that the urban and metropolitan scale turn out to be particularly important and this level of aggregation is not very well captured when using spatial units such as US states, European regions, UK or US counties or even UK postcode areas (for which the correspondence with the urban scale is good only for medium-sized cities). Finally we are able to deal flexibly with the crucial issue of the size distribution of establishments. To understand why flexibility is important, note that our three approaches outlined above yield similar aggregate results with respect to the extent and scales of localisation, but that there are large differences at the level of particular industries. This reflects the fact that there are marked differences in location patterns between small and large establishments and that the nature of those differences also vary across industries. Existing indices are narrowly constrained in the way they deal with the distribution of establishment size (e.g. through a Herfindhal index in the the EG case) whereas our approach is flexible and could easily be extended to other weighting methods for establishment size.

---

<sup>29</sup>If we had the same top 10 localised industries, but a different ordering, we would expect the average rank to be 5.5.

<sup>30</sup>We believe this downward border bias is why the EG index is consistently found to increase with the size of spatial units.

## 6. The scope of localisation in UK manufacturing

We now consider three extensions of our methodology related to the scope of the localisation that we observe. First, we evaluate the sectoral scope of localisation by replicating our baseline analysis for alternative three and five-digit sectoral classifications. Second, we consider whether we can identify localisation effects for four-digit industries *within* three-digit sectors. Third, we examine the tendency for different four-digit industries within the same sectors to co-locate.

### *Localisation of five-digit sub-industries*

In the UK, only 33 four-digit industries (out of 239) are sub-divided into 76 more finely defined five-digit sub-industries. We consider only the 58 of them that have more than 10 establishments.<sup>31</sup> Correcting for global confidence bands, we find that 43 of these sub-industries (or 74%) deviate significantly from randomness. More precisely, 26 sub-industries, that is 45%, show excess-localisation, 29% are excessively dispersed, while we cannot reject randomness for the remaining 26%.

These figures are not very different from those for four-digit industries. However, it is more meaningful to compare them with the patterns observed in the corresponding industries instead of the whole sample since sub-industries are more prevalent in some branches than in others. The branch with most sub-industries, 15 out of 58, is Food and beverage (SIC15). There, four-digit industries generally show dispersion and so do sub-industries. 14 out of 58 sub-industries are in Textiles (SIC17). These are mostly localised, sometimes highly so, just like their four-digit counterparts. The third large group, with 13 sub-industries, is in Chemicals (SIC24) and Machinery (SIC29). They show mixed patterns just like their corresponding industries.

When looking more closely at the differences between industries and their related sub-industries, three types of findings emerge. First, when the patterns of localisation are strong for industries, they are often even stronger for their sub-industries.<sup>32</sup> Second, we find (in both Food and Machinery) that industries showing either a dispersed or a seemingly random pattern are sometimes composed of one sub-industry that is localised and one that is only slightly dispersed. This implies that some of the lack of localisation that we detect for industries reflects a classification problem – five-digit sub-industries can show different non-random behaviours which look random when these are lumped together. Hence, using more finely defined industrial categories

---

<sup>31</sup>The following sub-industries were excluded: 15139, 15209, 15519, 15899, 15949, 17519, 17549, 18249, 21219, 23209, 25239, 26829, 29401, 29409, 36509 and 36639.

<sup>32</sup>However, some interesting details emerge. For example, for clothing industries, plants that produce women's clothing are always more localised than plants producing men's clothing.

Fraction of three-digit sectors localised at:					
5km	5km only	5 and 30km only	5 and 150km only	5, 30 and 150km	
35.9	5.8	19.4	1.0	9.7	
30km	30km only	30 and 150km only			
38.8	6.8	2.9			
150km	150km only				
20.4	6.8				

**Table 4.** Excess-localisation at three key thresholds for three-digit sectors

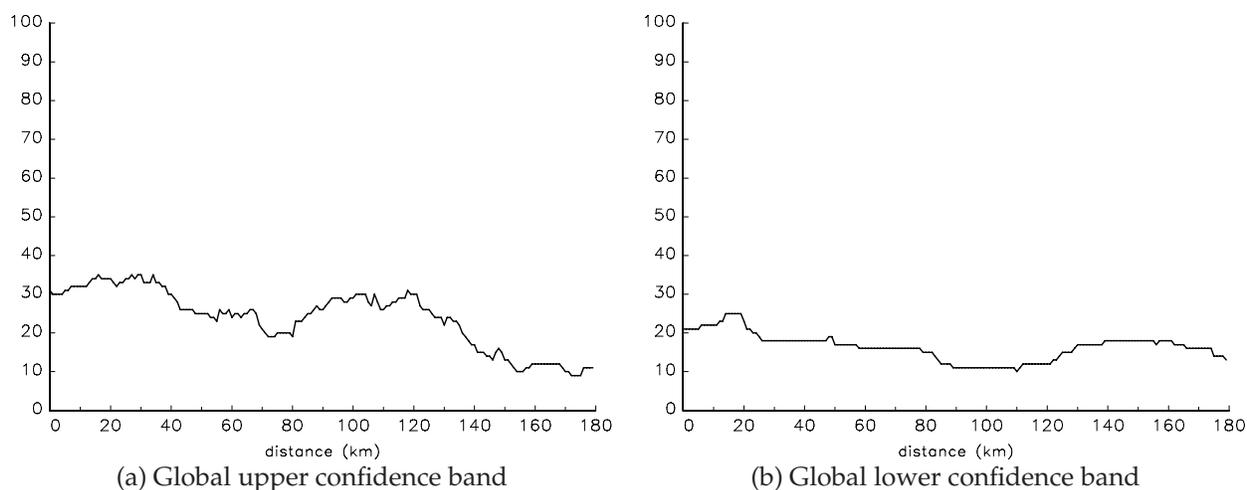
allows us to uncover some patterns that were so far hidden.<sup>33</sup> Third, in some instances when an industry shows minor excess-dispersion or localisation (in terms of  $\Psi_A$  or  $\Gamma_A$ ) we often cannot reject randomness for related five-digit components because these are smaller and thus have larger confidence bands. Thus, moving to a five-digit classification sometimes allows us to pick up more detail in the location patterns, but at the cost of greater imprecision reflected in the width of the confidence bands. In total, the increase in sectoral detail appears to offset the imprecision, so that we reject randomness in approximately the same proportion of industries.

#### *Localisation of three-digit sectors*

We now turn to the comparison between three-digit sectors and four-digit industries. Of 103 sectors, 87% of them deviate globally at some level of distance. The proportion with excess-localisation and excess-dispersion is higher than with four-digit industries: 58% and 29% respectively against 51% and 26% for industries.

To gain further insights, it is useful to consider the same three critical threshold as we did previously (5, 30 and 150 kilometres). The results are reported in Table 4. Comparing with Table 1, note that for the combinations involving the 5 and 30 but not the 150 kilometres thresholds, the figures are relatively similar. In contrast, for figures relating to the 150 kilometres threshold, excess-localisation seems more prevalent for sectors than industries. This finding is confirmed when looking at how many sectors deviate for each level of distance – plotted in Figure 8. The figure shows that there is first a high plateau of excess-localisation between 0 and 40 kilometres, then a decline followed by rise and a second plateau between 80 and 140 kilometres. The shape of this curve shares some similarities with its counterpart for four-digit industries represented in Figure 3. However, the relative number of sectors with excess-localisation at distances above 80 kilometres is much higher. Excess-dispersion shows a pattern similar to four-digit industries.

<sup>33</sup>However, note that production establishments must report only one SIC even though they may be engaged in different industries. Since multi-activity is more likely in closely related industries, classifications errors become more important as industries are more finely defined. Note also that five-digit product groups are only marginally more finely defined than four-digit industries. For instance SIC1751, Carpet and rugs, distinguishes between SIC17511, Woven carpet and rugs, and SIC17512, Tufted carpet and rugs. Such a fine distinction may not capture very many differences across establishments possibly using the same type of workers, and sharing the same customers and suppliers. In contrast, the difference between three-digit sectors and four-digit industries is markedly stronger. For instance SIC17500, Manufacture of other textile, is sub-divided into four very different industries: Carpets and rugs (17510), Cordage, rope and netting (17520), Non-wovens (17530) and Other textiles (17540).



**Figure 8.** Number of three-digit sectors with global excess-localisation and dispersion

In summary, for short distances (metropolitan level and below) excess-localisation is as frequent in sectors as in industries but localisation at the regional level is much more prevalent for sectors. How can we interpret these findings? The importance of excess-localisation above 80 kilometres is a new feature arising at this level of sectoral aggregation. This suggests an important role for vertical linkages in sectors, but is consistent with different processes governing firm location decisions. A first explanation for this large amount of excess-localisation above 80 kilometres is that firms in three-digit sectors may have a tendency to localise at fairly large spatial scales. Alternatively, this could be the result of four-digit industries being clustered at the metropolitan level (as seen before) and having these clusters located next to each other.

Problems also arise when trying to interpret the amount of excess-localisation observed for distances below 40 kilometres. It could be the case that what we detect here as excess-localisation for three-digit sectors is in fact excess-localisation in the four-digit industries within these sectors. For instance, excess-localisation in Pharmaceuticals (SIC244), might be driven by the strong tendency of Basic pharmaceuticals (SIC2441) to cluster. Alternatively, this finding could be driven by a tendency for firms across different industries part of the same sector to co-localise at this spatial scale. For instance in Pharmaceuticals (SIC244), firms in Pharmaceutical preparations (SIC2442) may try to locate close to firms in Basic pharmaceuticals (SIC2441) just like producers of car parts may seek to locate close to car assemblers. To resolve these issues, it is necessary first to measure the tendency for four-digit industries to localise after controlling for the localisation of the sector they belong to. The main reason for doing so is that if four-digit industries are not excessively localised taking their respective three-digit sectors as benchmarks, then the relevant level of aggregation will be the three-digit sector. By contrast, if the baseline results are unchanged despite the change of benchmark, it will imply that four-digit industries are the relevant level of aggregation to study excess-localisation.

### *Localisation within three-digit sectors*

Whether four-digit industries still show excess-localisation after controlling for the overall localisation of their three-digit sectors can be answered with a simple modification of the main approach described in Section 3. Instead of sampling the counterfactuals from the overall set of manufacturing sites,  $S$ , it is possible to sample only from the sites occupied by a firm in the same sector. Thus the approach is the same throughout but for any four-digit industry  $A$  which comprises  $n$  establishments and is part of sector  $B$ , just sample without replacement  $n$  sites from  $S_B$  the set of sites occupied by an establishment part of  $B$  instead of sampling from  $S$ .<sup>34</sup>

There are 184 industries that are part of a three-digit sector containing at least two industries. Of these, we find that 50% show excess-localisation. This is only marginally less than with our baseline. However, for each level of distance, fewer industries deviate and the amount of excess-localisation captured by  $\Gamma$  is much lower than in the baseline. With respect to excess-dispersion, the figure is at 18%, much lower than for the baseline (26%). Furthermore, there are no clear distance patterns with respect to these deviations from randomness.

From this we conclude that when controlling for the location of the parent sector, industries tend to exhibit less excess-localisation and dispersion. This implies that industries tend to follow broadly the locations patterns of their own sectors and are closer to them than to the patterns of general manufacturing. Interestingly, we note that for Textile (SIC17), Publishing (SIC22) and Basic metals (SIC27) industries, far fewer industries are localised with respect to their parent sector than general manufacturing. For instance, in Publishing only six industries in 13 show excess-concentration in their parent sector against all of them in the baseline analysis.<sup>35</sup> Still, five of the 10 most localised industries in Table 2 are still among the 10 most localised industries when controlling for the location of their parent sectors. However, their  $\Gamma$  has on average a value less than half of that in the baseline. This tendency for industries to follow broadly the patterns of their parent sectors can shed lights on our results regarding three-digit sectors. First, this can explain why for sectors, there is still a tendency to localise at small geographical scales. Second and since the overlap between industry clusters is not perfect, this can also explain why sectors also localise at larger scales (100 kilometres and above).

### *Localisation across three-digit sectors*

Establishing that four-digit industries still have a tendency, albeit weaker, to localise after controlling for the location of their parent sector is not enough. It is also worth exploring whether industries that are part of the same sector tend to co-localise. Despite some attention in the recent literature (Ellison and Glaeser, 1997, Devereux *et al.*, 1999) and its crucial importance with respect to many theoretical and policy concerns, very little is known about co-localisation.<sup>36</sup> Although it

---

<sup>34</sup>The concept of "co-agglomeration" used by Ellison and Glaeser (1997) is also in this spirit since it is based on the difference between the localisation index of the sector and the weighted average of the localisation indices of each individual industry.

<sup>35</sup>The reason is that most industries in Publishing are based around London. The same hold for Birmingham and Basic Metals, Manchester and textile, etc.

<sup>36</sup>The term co-agglomeration is also used in the literature. For consistency with respect to the terminology used so far, we speak of co-localisation.

is easy to define co-localisation as the tendency of different industries (or more generally different partitions in a population of firms) to cluster together, measuring co-localisation is much more complex than localisation. To see why, consider the following generalised version of equation (1):

$$K_{(A,B)}(d) \equiv \sum_{i=1}^{n_A} \sum_{\substack{j=1 \\ j \neq i}}^{n_B} \frac{\delta(i,j,d)}{P(n_A, n_B)}. \quad (9)$$

where  $A$  and  $B$  are two (possibly overlapping) subsets from the population of establishments,  $S$ , and  $P(n_A, n_B)$  is the total number of unique bilateral distances between pairs of establishments with one establishment from each subset.<sup>37</sup> The density  $K_{(A,B)}(\cdot)$  is a straightforward generalisation of  $K_A(\cdot)$  which allows us to calculate the density of bilateral distances between establishments in any two subsets from a population.

Then, it remains to define the counterfactuals which this distribution should be compared to. In this respect, note that tests over  $K_{(A,B)}(\cdot)$  may involve counterfactuals  $\tilde{A}$  and  $\tilde{B}$  drawn from any subset of  $S$ . A priori this implies a considerable number of possible tests. Many of these tests are not very informative with respect to co-localisation. For instance calculating the distance density function using equation (9) for four-digit industries  $A$  and  $B$  that belong to the same sector and comparing it to counterfactuals generated by sampling from the overall population of firms leads to results that are highly problematic to interpret. The reason is that localisation in two industries can also generate some form of co-localisation across the same two industries. Fundamentally two industries may cluster in the same area because they are attracted to each other or because they are attracted by whatever this area has to offer (and it may not even be the same elements that attract the two industries). A measure of gross co-localisation based on sampling from the overall population cannot distinguish between these explanations and thus confuses co-localisation with localisation.

Among all the possible tests one could construct using (9) and given the observational equivalence between localisation and co-localisation, we believe one exercise is of particular interest here. The most informative test is to see whether there is some tendency for co-localisation across four-digit industries after controlling for the overall tendency of the parent three-digit industry to cluster. To investigate this, we can apply equation (9) to any two four-digit industries,  $A$  and  $B$  part of the same three-digit industries, and sample our counterfactuals from the set of all sites occupied by an establishment in any of these industries,  $A \cup B$ . For instance after controlling for the tendency of the partition composed of Basic pharmaceuticals and Pharmaceutical (SIC2441 and 2442) to localise, it allow us to test whether plants in SIC2442 tend to locate closer to plants in SIC2441 than randomness would suggest. Note that this is a strong test since the rejection of randomness means that establishments in these industries are attracted to each other even after controlling for whatever tendency they have to cluster. At the same, being unable to reject randomness is no rejection of industries being co-localised in some sense.

[Results to be reported]

---

<sup>37</sup>If  $A$  and  $B$  are the same set ( $A \sim B$ ) then  $P(n_A, n_B) = \frac{n_A(n_A-1)}{2}$ . If  $A$  and  $B$  are disjoint sets ( $A \cap B = \emptyset$ ) then  $P(n_A, n_B) = n_A n_B$ .

## 7. Conclusion

To study the detailed location patterns of industries, we developed distance-based measures of localisation. We were guided by the principle that any such measure must satisfy five requirements: (i) comparability across industries, (ii) control for the uneven distribution of overall manufacturing or population, (iii) control for industrial concentration, (iv) no aggregation bias, and (v) statistical significance. Our measures can satisfy all five requirements whereas previous studies could satisfy at most three.

We applied these measures to a unique and exhaustive UK manufacturing data set. Our main findings are the following:

- 51% of four-digit industries exhibit excess-localisation at a 5% confidence level. Many of them still show excess-localisation even within their three-digit industries.
- 26% of four-digit industries exhibit excess-dispersion at a 5% confidence level.
- Excess-localisation in four-digit industries takes place mostly between 0 and 50 kilometres.
- Four- and five-digit industries follow broad sector and branch patterns with respect to localisation.
- In some industrial branches, excess-localisation at the industry level is driven by larger establishments, whereas in others it is smaller establishments which have a tendency to cluster.
- The extent of excess-localisation and dispersion are very skewed across industries.
- Excess-localisation and dispersion are as frequent in three-digit sectors as in four-digit industries for distances below 80 kilometres.
- Three-digit sectors also show a lot of excess-localisation at the regional scale (80 – 140 kilometres) due in part to the tendency of four-digit industries to co-agglomerate at this spatial scale.
- Co-agglomeration [To be reported]

Some of these findings confirm previous findings in the literature. For instance, the high levels of heterogeneity in location patterns across industries have been observed by most previous studies. Other results are in stark contrast to previous studies. For instance, we do not find as many industries to be localised as previously claimed. Along the same lines, we found the propensity for excess-dispersion to be stronger than one may have believed. Our results about broad sectoral effects are also stronger than previously obtained. Our results on plant size suggest that differences in location behaviour between large and small plants are much more nuanced than earlier studies have led us to believe. Finally our results about the scale and the scope of localisation are to a large extent completely new as the tools previously available were not suited to an exploration of these issues.

Many detailed issues remain to be investigated as regards the issues of localisation, dispersion, and co-localisation. For instance, one may wish to compare the behaviour of independent plants with that of plants that are part of multi-unit firms or that of foreign plants versus domestic ones. Also, much remain to be learnt about co-localisation in vertically linked industries, etc. We hope to be able to shed light on these questions in future research. It must be noted that distance-based analyses can be applied beyond industrial geography. Any data with detailed geographical information readily lends itself to this type of analysis. In the past, studies involving distance based measures could be performed only on very small populations (Cressie, 1993) for lack of computing power and precise enough data. These two obstacles are gradually being removed and we hope to see more of this type of study in the future.

Furthermore, and as shown in part by our study, distance-based analysis not only allows us to answer long standing empirical questions in a more precise and accurate way but it also allows us to obtain answers that were previously unavailable.

## References

- Cressie, Noel A. C. 1993. *Statistics for Spatial Data*. New York: John Wiley.
- Devereux, Michael P., Rachel Griffith, and Helen Simpson. 1999. The geographic distribution of production activity in the uk. Institute for Fiscal Studies Working Paper 26/99.
- Ellison, Glenn and Edward L. Glaeser. 1997. Geographic concentration in US manufacturing industries: A dartboard approach. *Journal of Political Economy* 105(5):889–927.
- Ellison, Glenn and Edward L. Glaeser. 1999. The geographic concentration of industry: Does natural advantage explain agglomeration? *American Economic Review Papers and Proceedings* 89(2):311–316.
- Griffith, Rachel. 1999. Using the ard establishment level data: an application to estimating production functions. *Economic Journal* 109(456):F416–F442.
- Henderson, J. Vernon. 1999. Marshall's economies. Working Paper 7358, National Bureau of Economic Research. URL <http://www.nber.org/>.
- Holmes, Thomas J. and John J. Stevens. 2000. Geographic concentration and establishment scale. Processed, University of Minnesota.
- Hoover, Edgar M. 1937. *Location Theory and the Shoe and Leather Industries*. Cambridge, Mass.: Harvard University Press.
- Marshall, Alfred. 1890. *Principles of Economics*. London: Macmillan.
- Maurel, Françoise and Béatrice Sédillot. 1999. A measure of the geographic concentration of french manufacturing industries. *Regional Science and Urban Economics* 29(5):575–604.
- Raper, Jonathan F., David W. Rhind, and John W. Shepherd. 1992. *Postcodes : The New Geography*. Harlow, Essex: Longman Scientific and Technical.
- Rosenthal, Stuart A. and William C. Strange. 2001. The determinants of agglomeration. *Journal of Urban Economics* (forthcoming).

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Yule, George U. and Maurice G. Kendall. 1950. *An Introduction to the Theory of Statistics*. London: Griffin.