Review of Economic Studies (2005) **72**, 1077–1106 0034-6527/05/00421077$02.00

Testing for Localization Using Micro-Geographic Data

GILLES DURANTON and HENRY G. OVERMAN
London School of Economics

First version received December 2002; *final version accepted December* 2004 (*Eds.*)

To study the detailed location patterns of industries, and particularly the tendency for industries to cluster relative to overall manufacturing, we develop distance-based tests of localization. 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. We apply these tests to an exhaustive U.K. data-set. For four-digit industries, we find that (i) 52% of them are localized at a 5% confidence level, (ii) localization mostly takes place at small scales below 50 km, (iii) the degree of localization is very skewed, and (iv) industries follow broad sectoral patterns with respect to localization. Depending on the industry, smaller establishments can be the main drivers of both localization and dispersion. Three-digit sectors show similar patterns of localization at small scales as well as a tendency to localize at medium scales.

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 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 localize (*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.[1] Furthermore, what we may 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 of international trade, industrial organization and economic growth.

In this paper however, we step back from policy and theoretical concerns and think again about the stylized 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 of the clustering of cutlery producers in Sheffield or jewellers in Birmingham. Neither do we deny Silicon Valley's importance in microelectronics and software. However, it is worth asking whether these examples 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 (U.K.), the localization of the cutlery industry in one area of Sheffield is different from that of the motor sport industry spreading over more than 100 km along the Thames Valley. Finally, it is important to know whether small or large establishments are the main driver of localization and investigate its sectoral scope.

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

Building on previous research in spatial statistics, we develop a novel way to test for localization and answer key questions about the extent of localization, the spatial scales at which it takes place, and its sectoral scope. Our test is based on a measure of localization, which is comparable across industries, and controls for both the overall tendency of manufacturing to agglomerate and for the degree of industrial concentration. These three requirements have already been recognized by the literature. However, in this paper we argue that any measure of localization should also be unbiased with respect to scale and spatial aggregation and that any test of localization should report the significance of the result. Our approach satisfies these two additional properties. Let us consider each of these five requirements in turn.

Obviously, any test of localization must be based on a measure that is comparable across industries. This measure must also control for the general tendency of manufacturing to agglomerate. For instance in the United States (U.S.), even in the absence of any tendency towards localization, we would expect any typical 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. These first two requirements have been recognized in the literature for a long time. Most traditional measures, like Gini indices, are able to satisfy them when properly employed.

Since Ellison and Glaeser (1997), it is also widely recognized that any informative measure of localization must control for industrial concentration. To understand the distinction between localization and industrial concentration, note that in an industry with no tendency for clustering, the location patterns of the plants are determined by purely idiosyncratic factors. 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 at 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. For instance, according to Ellison and Glaeser (1997), 75% of the employees in the U.S. vacuum cleaner industry work in one of four main plants. Even if these plants locate separately, four locations must account for at least 75% of the employment in this industry without it being localized in any meaningful way. In short, unevenness does not necessarily mean an industry is localized. Unfortunately traditional measures of localization 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 should be one of spatial randomness conditional on both industrial concentration and the overall agglomeration of manufacturing. The index they develop satisfies these two requirements and is comparable across industries. Taking a similar dartboard approach, Maurel and Sédillot (1999) and Devereux, Griffith and Simpson (2004) develop alternative indices of localization with the same properties.

However, just like the more traditional indices, these "second generation" measures still *ex ante* allocate establishments (*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 in this way has the obvious advantage of making computations simple but it means throwing away a large amount of information and leads to a range of aggregation problems.

Most obviously, aggregation restricts the analysis to only one spatial scale, be it the county, region or state. Exploring a different spatial scale requires another aggregation and running the analysis again. This is limiting because in most countries the number of levels of aggregation is commonly limited to two or three. More importantly, it is difficult to compare the results across different scales. For instance, questions regarding how much industries are localized at the county level after controlling for localization 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 not 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 localization of industries at the level of U.S. states involves comparisons between Rhode Island and California, which is geographically more than 150 times as large.

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 recognized by quantitative geographers (Yule and Kendall (1950), Cressie (1993)) and is known as the Modifiable Areal Unit Problem (MAUP).

Finally, and importantly, after aggregation has taken place, spatial units are treated symmetrically so that plants in neighbouring spatial units are treated in exactly the same way as plants at opposite ends of a country. This creates a downwards bias when dealing with localized industries that cross an administrative boundary. This problem worsens as smaller spatial units are analysed. For instance in the U.K., manufacture of machinery for textile is highly localized but the border between the East and West Midlands regions cuts the main cluster in half. Using U.K. counties would make matters even worse. Any good measure of localization must avoid these aggregation problems. Ours does, by directly using the distances between observations and thus working in continuous space rather than aggregating observations within administrative units.

The last requirement for any test of localization relates to its statistical significance. Given our definitions, in the absence of localization, the location of an industry is random conditional on industrial concentration and the location of overall manufacturing. Thus any statement about non-randomness can only be probabilistic. The literature mentioned above only offers localization indices with no indication of statistical significance. In contrast, we analyse the statistical significance of departures from randomness using a Monte Carlo approach.

In summary, any test of localization should rely on a measure which (i) is comparable across industries; (ii) controls for the overall agglomeration of manufacturing; (iii) controls for industrial concentration; (iv) is unbiased with respect to scale and aggregation. The test should also (v) give an indication of the significance of the results. The approach we propose here satisfies these five requirements. 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 significance. 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.[2]

We apply our approach to an exhaustive U.K. manufacturing data-set. Four main conclusions emerge with respect to four-digit industries: (i) 52% of them are localized at a 5% confidence level, (ii) localization takes place mostly between 0 and 50 km, (iii) the degree of localization is very skewed across industries, and (iv) industries that belong to the same industrial branch tend to have similar localization patterns. In part, our results are entirely new as we know of no previous systematic attempt to measure the scale of localization. Where our results can be compared with previous work, there are marked differences. For instance, 94% of U.K. four-digit

2. Our philosophy when developing this methodology has been to impose a set of statistical requirements on our test. An alternative would be to develop tests based on an underlying economic model. As will become clear below, our test is also consistent with such a model-based approach. In this sense, our approach develops a test of the simplest possible location model (*i.e.* pure randomness conditional on the distribution of overall manufacturing). In future work, we expect to develop our methodology to test more sophisticated theories of industrial location. The key barrier to achieving this is the difficulty of generating counterfactual location patterns from such theories. An alternative would be to use the indices of localization we derive below as endogenous variables and regress them on a set of industry characteristics as in Rosenthal and Strange (2001).

industries are localized according to the Ellison–Glaeser (EG) index compared to 52% using our approach.

When looking at the location patterns of establishments with few employees we find a wide variety of behaviours. In some industries (often related to publishing, chemicals or electric machinery), smaller establishments tend to be more localized than larger establishments. In other industries, like food and beverages or non-metallic mineral products, the opposite holds: Smaller establishments are located away from the main clusters. Regarding the sectoral scope of localization, 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, localization is equally important at small scales (0–50 km) and at a more regional level (80–140 km). We find that these regional effects are caused by the tendency of four-digit industries that are part of the same sectors to co-localize at this spatial scale.

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

2. DATA

Our empirical analysis uses exhaustive establishment level data from the 1996 Annual Respondent Database ARD which are the data underlying the Annual Census of Production in the U.K. Collected by the Office for National Statistics (ONS), the ARD is an extremely rich data-set which contains information about all U.K. establishments (see Griffith, 1999, for a detailed description of these data). We restrict ourselves to production establishments in manufacturing industries using the Standard Industrial Classification (SIC) 92 (SIC15000 to 36639) for the whole country except Northern Ireland.[3] For every establishment, we know its postcode, five-digit industrial classification and number of employees. Note that, when referring to SIC two-, three-, four- and five-digit categories, we will speak of industrial branches, sectors, industries and sub-industries, respectively.

The postcode is particularly useful for locating plants. In the U.K., postcodes typically refer to one property or a very small group of dwellings. Large buildings may even comprise more than one postcode. See Raper, Rhind and Shepherd (1992) for a complete description of the U.K. postcode system. The CODE-POINT data-set from the Ordnance Survey (OS) gives spatial coordinates for all U.K. postcodes. These data are the most precise postcode geo-referencing data available for the U.K. Each Code-Point record contains information about its location, and about the number and type of postal delivery points. By merging these data together with the ARD we can generate very detailed information about the geographical location of all U.K. manufacturing establishments. In so doing, we could directly establish the Eastings and Northings for around 90% of establishments. These give the grid reference for any location taking as the origin a point located south-west of the U.K.

The main problem for the remaining 10%, for which the postcode could not be matched with spatial coordinates, relates to postcode updates. These take place when new postcodes are 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 error to a minimum, we checked our data against all postcode updates since 1992. This left us with 5% of establishments that could not

3. We use the terms establishment and plant interchangeably.

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 reporting 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 m. For the remaining 26 observations, the maximum error is a few kilometres.

Figures 1(a)–(d) map this location information for four industries: Basic Pharmaceuticals (SIC2441), Pharmaceutical Preparations (SIC2442), Other Agricultural and Forestry Machinery (SIC2932), and Machinery for Textile, Apparel and Leather Production (SIC2954). Each dot represents a production establishment. As can be seen from the maps, Machinery for Textile, Apparel and Leather Production (SIC2954) looks very localized whereas Other Agricultural and Forestry Machinery (SIC2932) is very dispersed. These are extreme cases. Basic Pharmaceuticals (SIC2441) and Pharmaceutical Preparations (SIC2442) are more representative of the typical pattern. Whether or not these last two industries are localized is far from obvious. In the exposition of the methodology below, we keep using these four industries for illustrative purposes. However, the main results will consider all industries.

3. METHODOLOGY

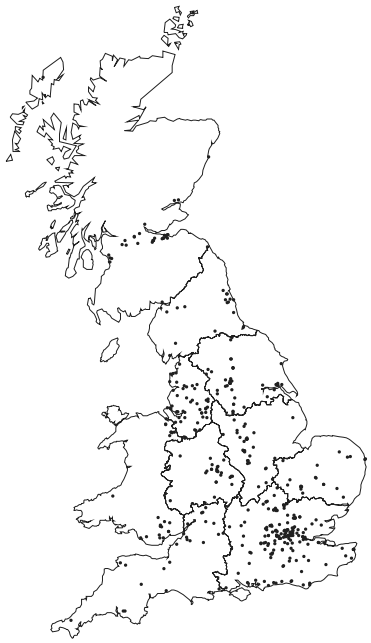
Our analysis is conceptually simple. We first select the relevant establishments. The second step is to compute the density of bilateral distances between all pairs of establishments in an industry. This measure is unbiased with respect to spatial scale and aggregation and thus satisfies our fourth requirement. The third step is to construct counterfactuals. To satisfy our first and third requirements about comparability across industry and the need to control for industrial concentration, we consider hypothetical industries with the same number of establishments. Any existing establishment, regardless of its industry, is assumed to occupy one site. Establishments in our hypothetical industries are randomly allocated across these existing sites. This controls for the overall distribution of manufacturing and thus satisfies our second requirement. Finally we construct local confidence intervals and global confidence bands to take care of our fifth requirement. This allows us to compare the actual distribution of distances to randomly generated counterfactuals and to assess the significance of departures from randomness. We now describe these steps in greater detail.

Selection of observations

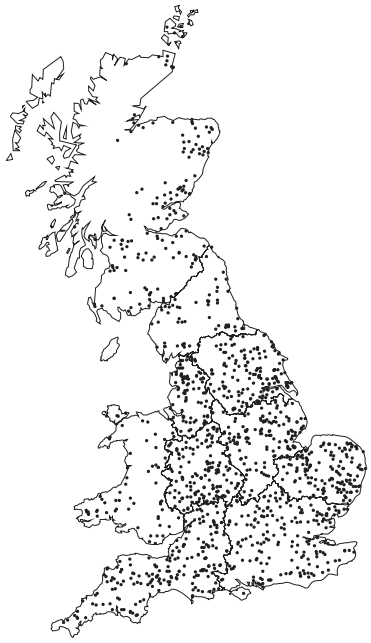
For any particular industry (and more generally for any partition of our population of establishments), we first select the relevant observations. The main issue to consider here is the large number of small establishments, which may have different location patterns. For instance in naval constructions, there are many very small establishments of 1–10 employees located inland whereas all the large establishments are located on a coast. It seems likely that these establishments, although classified in the same industry, do not do the same thing. One alternative is to ignore this problem and consider the whole population of U.K. production establishments. A second possibility would be to consider a size threshold and retain only establishments with employment above this threshold. We then need to choose between an absolute and a relative threshold. Dropping all establishments with less than 10 workers may be reasonable for naval construction but less so for publishing. Such reasoning suggests the use of a relative threshold where we select establishments by decreasing size so that say 90% of employment is considered. A last possibility is to weight establishments by their employment.

We implement all three approaches. In our baseline analysis (Section 4), we consider all establishments independent of 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

(a) Basic Pharmaceuticals (SIC2441)



(b) Pharmaceutical Preparations (SIC2442)



(c) Other Agricultural and Forestry Machinery (SIC2932)

(d) Machinery for Textile, Apparel and Leather Production (SIC2954)

FIGURE 1
Maps of four illustrative industries

establishments by employment. Note that this captures a slightly different concept: weighting by employment gives a measure of the localization of employment and no longer that of establishments.

Kernel estimates of K-densities

Next, for industry A with n establishments, we calculate the Euclidean distance between every pair of establishments. This generates $\frac{n(n-1)}{2}$ unique bilateral distances. Although the location of nearly all establishments in our data is known with a high degree of precision, any Euclidean distance is only a proxy for the true physical distance between establishments. The curvature of the earth is a first source of systematic error. However it is easy to verify that in the U.K. 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 distances are much longer than Euclidean distances) whereas in high-density areas they are more numerous (so Euclidean distances are a good approximation to actual) but also more congested. It is unclear which effect dominates so we impose no specific correction.[4] We are still left with random errors. For example, the real distance between two points along a straight road is equal to its Euclidean distance whereas that between two points on opposite sides of a river is usually well above its Euclidean counterpart. Given this noise in the measurement of distances, we decided to kernel-smooth when estimating the distribution of bilateral distances.

Denote by $d_{i,j}$, the Euclidean distance between establishments i and j. Given n establishments, the estimator of the density of bilateral distances (henceforth K-density) at any point d is

$$\widehat{K}(d) = \frac{1}{n(n-1)h} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} f\left(\frac{d - d_{i,j}}{h}\right) \tag{1}$$

where h is the bandwidth and f is the kernel function. All densities are calculated using a Gaussian kernel with the bandwidth set as per Section 3.4.2 of Silverman (1986). The solid lines in Figures 2(a)–(d) plot these densities for the same four industries as previously. The dashed and dotted lines in the figures, which refer to local and global confidence bands, will be explained later.

Estimation issues

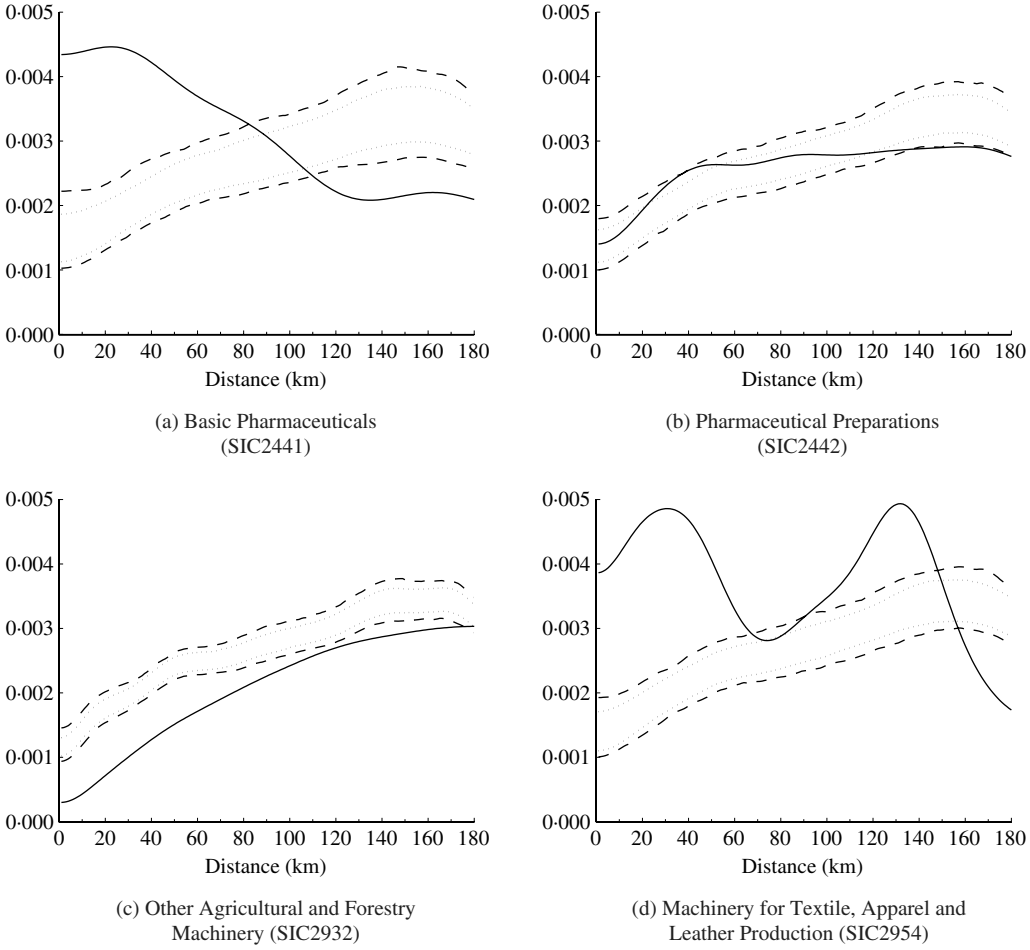
Before going further, we briefly discuss several estimation issues.

Because distances cannot be negative, we need to constrain our density estimates to be zero for negative distances. One possible solution is to estimate the kernel density ignoring this boundary restriction, then to set the estimate to zero for negative distances and re-scale at all other distances to ensure that the density still sums to one. Unfortunately, this uniform re-scaling means that distances near to the boundary will contribute less to our density estimate and thus that the weight of the distribution near zero will be underestimated. Instead we deal with this boundary problem by adopting the reflection method proposed in Section 2.10 of Silverman (1986).[5]

The nature of our data implies two important differences with respect to standard kernel density estimation. Note first that each actual K-density is calculated on the basis of a *census* of

4. According to Combes and Lafourcade (2005), the correlation between Euclidean distances and generalized transport costs (computed from real transport data) for France is extremely high at 0·97.

5. If the original data-set for industry A is $X_1, X_2, \ldots$ the reflected data-set is $X_1, -X_1, X_2, -X_2 \ldots$. We then estimate $\widehat{K}^*_A(d)$ using this augmented data-set and define $\widehat{K}_A(d) = 2\widehat{K}^*_A(d)$ if $d > 0$ and $\widehat{K}_A(d) = 0$ if $d \leq 0$.



(a) Basic Pharmaceuticals (SIC2441)

(b) Pharmaceutical Preparations (SIC2442)

(c) Other Agricultural and Forestry Machinery (SIC2932)

(d) Machinery for Textile, Apparel and Leather Production (SIC2954)

FIGURE 2

K-density, local confidence intervals and global confidence bands for four illustrative industries

the entire industry population. If, instead of a census, we had a random sample of firms from each industry we would need to worry about the statistical variation due to the estimation of the actual K-density. Applications of the techniques developed below to samples of firms from particular industries could allow for this statistical variation to be taken into account but the exhaustive nature of our data means that we are able to ignore it in what follows (see Efron and Tibshirani, 1993, and Quah, 1997, for further discussion of these issues as well as Davison and Hinkley, 1997, for a discussion more focused on point patterns).

The second difference stems from the fact that the spatial nature of our data implies strong dependence between the bilateral distances that are used to calculate the density. This strong dependence arises because the observations of interest are actually the points that generate these bilateral distances. Even if the underlying points are independently located, the bilateral distances between these points will not be independent.[6] This has implications for the sampling theory of our estimator, $\widehat{K}_A(d)$. In situations where the observations are independent (or only

6. See below for more on this issue.

weakly dependent) then the limiting distribution of this estimator can be derived using a central limit theorem for U-statistics. Section 3.7 of Silverman (1986) provides more details of these asymptotic properties. Unfortunately, such results are not available for the spatial point pattern data that we consider here. See Cressie (1993) and Diggle (2003) for further discussion. This means that we will have to rely on Monte Carlo results in order to test for departures from randomness. This is the issue to which we now turn.

Constructing counterfactuals

At this stage, we need to decide on the relevant counterfactuals to which our K-densities should be compared. In this respect, note that the analysis of localization is informative only to the extent that it captures interactions across establishments or between establishments and their environment. Consequently the number of firms in each industry and the size distribution therein are taken as given.[7]

To satisfy our second condition, we need to control for the overall tendency of manufacturing to agglomerate. Furthermore, we need to allow for the fact that in the U.K. zoning and planning restrictions are ubiquitous. Manufacturing cannot locate in many areas of the country (*e.g.* the Lake District, London's green belt, etc.). Hence, to control for overall agglomeration and the 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.[8]

How should we test whether the sites occupied by a particular industry can be considered a random draw from all existing sites? One possibility would be to calculate the distribution of bilateral distances for all establishments and then to compare the estimated distribution for a particular industry to this distribution. Sampling distances directly from the density of distances for the whole of manufacturing would make it possible to calculate exact confidence intervals because the density at each level of distance could then be treated as the result of repeated binomial draws. However, this short-cut amounts to treating the bilateral distances between points as independent, which they are not as already argued above.[9] To see this, consider directly drawing bilateral distances for the simplest case of three plants. It is possible with a small but non-negligible probability to obtain three bilateral distances above 700 km. However, it is impossible to have three plants a distance of 700 km from each other in the U.K. as such an equilateral triangle simply cannot fit on its territory.

To avoid this problem, we need to construct counterfactuals by first drawing locations from the overall population of sites and then calculating the set of bilateral distances. For each industry we run 1000 simulations.[10] 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 $m = 1, 2, \ldots, 1000$, by sampling n elements without replacement from S so that each simulation is equivalent to a random reshuffling of establishments across sites. This controls for both industrial

7. That is we take increasing returns within the firm as given. 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.

8. A site is where one establishment is located—when two establishments share the same postcode, two different sites are distinguished.

9. It is this dependence that rules out the use of the asymptotic results reported in Section 3.7 of Silverman (1986).

10. We also repeated our simulations 2000 and 10,000 times for a few industries with very similar results.

concentration and the overall agglomeration of manufacturing. Other alternatives are possible. For instance one could draw the counterfactuals from the set of all U.K. postcodes. Given that the number of residential addresses far outweigh that of manufacturing addresses, this would control for the overall distribution of population instead. However we think it is more informative to control for the overall distribution of manufacturing given the constraints on manufacturing location in the U.K.[11] Once we have the counterfactual, we calculate the smoothed density for each simulation exactly as we did for the real industry.

Before moving on, we note that detailed investigation suggests that the lack of independence between distances is important beyond 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 and distance-generated K-densities are too large for us to be able to sample distances directly. For instance, for an industry with 200 establishments (which corresponds roughly to the median number of establishments for four-digit industries), we found that the confidence intervals were about twice as large when drawing distances directly.

Local confidence intervals

The next step is to calculate local confidence intervals. We consider all distances between 0 and 180 km. This threshold is the median distance between all pairs of manufacturing establishments. Any "abnormally" high values for the distance density, $\widehat{K}_A(d)$, for $d > 180$ could in principle be interpreted as dispersion but this information is redundant if we consider both *lower* and upper confidence intervals for $d < 180$. This reflects the fact that the densities must sum to one over the entire range of distances (0–1000 km). Hence we restrict our analysis to the interval [0, 180]. For each industry, for each kilometre in this interval we rank our simulations in ascending order and select the 5-th and 95-th percentile to obtain a lower 5% and an upper 5% confidence interval that we denote $\overline{K}_A(d)$ and $\underline{K}_A(d)$, respectively. When for industry A, $\widehat{K}_A(d) > \overline{K}_A(d)$, this industry is said to exhibit *localization at distance d (at a 5% confidence level)*. Symmetrically, when $\widehat{K}_A(d) < \underline{K}_A(d)$, this industry is said to exhibit *dispersion at distance d (at a 5% confidence level)*.[12] We can also define an index of localization

$$\gamma_A(d) \equiv \max\left(\widehat{K}_A(d) - \overline{K}_A(d), 0\right), \tag{2}$$

as well as an index of dispersion

$$\psi_A(d) \equiv \max\left(\underline{K}_A(d) - \widehat{K}_A(d), 0\right). \tag{3}$$

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

Graphically, localization (dispersion) is detected when the K-density of one particular industry lies above (below) its local upper (lower) confidence interval. The two dotted lines in Figures 2(a)–(d) plot these local confidence intervals for our four illustrative industries. For instance, Machinery for Textile, Apparel and Leather Production (SIC2954) exhibits localization for every kilometre from 0 to 60 whereas Other Agricultural and Forestry Machinery (SIC2932)

11. An important avenue for future research is to develop counterfactuals from more sophisticated theories of industrial location and test them in the same fashion as below. We view randomness conditional on overall agglomeration as the first and most obvious null hypothesis to be tested but this is by no means the only one.

12. Dispersion here is precisely defined as having fewer establishments at distance d than randomness would predict. In other words the distribution of a dispersed industry 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 10 draws is rather unlikely and akin to our concept of localization. Alternatively, five zeros alternating with five ones is as unlikely and this extreme regularity is interpreted in a geographical context as dispersion.

exhibits dispersion over the same range of distances. Note that the shape of these confidence intervals reflects the distribution of overall manufacturing.

Global confidence bands

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. Even a randomly distributed industry will exhibit dispersion or localization for some level of distance with quite a high probability. To see this, recall that there is a 5% probability an industry shows localization for each kilometre, so that the probability of this happening for at least one kilometre among 180 is quite high 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 a band such that no less than 95% of a series of randomly generated K-densities lie above or below that band. The restriction we impose here is standard: we choose identical local confidence intervals 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-densities hit them. The local confidence levels associated with these 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-densities. This problem is solved easily by interpolating. The second worry is that we may need to consider the local 99·9-th or even the 100·0-th percentile to get a 5% confidence band. The variance of 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, because localization and dispersion are correlated across distances (be it only because of optimal smoothing), the local confidence level such that 5% of our randomly generated industries deviate is typically around 99%, *i.e.* the 10-th extreme value, for which the variance is much lower.

Denote $\overline{\overline{K}}_A(d)$ the upper confidence band of industry A. This band is hit by 5% of our simulations between 0 and 180 km. When $\widehat{K}_A(d) > \overline{\overline{K}}_A(d)$ for at least one $d \in [0, 180]$ this industry is said to exhibit *global localization (at a 5% confidence level)*. Turning to global dispersion, recall that by construction, our distance densities must sum to one. Thus an industry which is very localized at short distances can show dispersion at larger distances. In other words, for strongly localized industries, dispersion is just an implication of localization. This discussion suggests the following definition: The lower confidence band of industry A, $\underline{\underline{K}}_A(d)$, is such that it is hit by 5% of the randomly generated K-densities that are *not* localized. An industry is then said to exhibit *global dispersion (at a 5% confidence level)* when $\widehat{K}_A(d) < \underline{\underline{K}}_A(d)$ for at least one $d \in [0, 180]$ *and* the industry does not exhibit localization. As before, we can define

$$\Gamma_A(d) \equiv \max\left(\widehat{K}_A(d) - \overline{\overline{K}}_A(d), 0\right), \tag{4}$$

an index of global localization and

$$\Psi_A(d) \equiv \begin{cases} \max\left(\underline{\underline{K}}_A(d) - \widehat{K}_A(d), 0\right) & \text{if } \sum_{d=0}^{d=180} \Gamma_A(d) = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

an index of global dispersion.

Graphically, global localization is detected when the K-density of one particular industry lies above its upper confidence band. Global dispersion is detected when the K-density lies below the 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, Machinery for Textile, Apparel and Leather Production (SIC2954) exhibits global localization whereas Other Agricultural and Forestry Machinery (SIC2932) exhibits dispersion. Pharmaceutical Preparations (SIC2442) shows neither global localization nor dispersion while Basic Pharmaceuticals (SIC2441) shows global localization and thus by definition no dispersion even though its K-density does go beneath the lower confidence band.

Interpretation and examples

To understand better what these tests capture, let us consider a few examples. Take first an industry like Basic Pharmaceuticals (SIC2441) with a cluster of plants around London. This cluster implies a high density for distances between 0 and 60 km and this industry thus shows localization between these distances. Consider now an industry like Machinery for Textile, Apparel and Leather Production (SIC2954) 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 localization between 0 and 50 km. Furthermore, Manchester and Birmingham are quite close to each other so that there is also localization for distances between 100 and 140 km. 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 more than 180 km from each other. A multiplicity of peaks in our distance density thus indicates a multiplicity of clusters close to each other. Consider now the more contrived case of an industry located mostly in one region but with regularly dispersed plants (in order to serve local markets for instance). Such an industry would be locally dispersed at short distances, but also localized at higher levels of distances (capturing the fact that it is present in only one region).

A limit common to all approaches, including ours, is that we cannot detect non-random patterns if they do not involve localization or dispersion. Thus we may accept as random some industries whose location is clearly non-random, for example industries located along a coast or a railway line. However, the likelihood of such a type-II error decreases with the number of establishments. For instance, an industry with plants along a straight line is localized at short distances (the number of neighbours increases linearly with distance whereas if location is random, the increase is quadratic) and this should be detected provided there are enough establishments.

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 specialization of particular local economies. The analysis of specialization is conceptually distinct from that of localization and as it requires different tools it is beyond the scope of this paper.

Before presenting our results, it is worth considering what we can and cannot learn from this type of analysis. Localization 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 this pattern. In the Appendix, we develop a simple economic model which shows that our test is indeed one of randomness vs. localization resulting from either external economies or natural endowments.

4. BASELINE RESULTS FOR FOUR-DIGIT INDUSTRIES

In this section we describe our results for U.K. four-digit industries using the complete population of plants.

How many industries deviate and where

We consider 234 industries (out of 239) that have more than 10 establishments. Two hundred and five deviate locally at some distance over the range. Correcting for global confidence bands as outlined in Section 3 leads us to conclude that 177 industries differ significantly from randomness at the 5% level of significance. The detailed breakdown is as follows. We find that 122 industries, that is 52% of them, are localized whereas 55 industries (24%) are dispersed, and 57 (24%) do not deviate significantly from randomness. From our results, localization is not as widespread as earlier studies led us to believe whereas dispersion seems much more prevalent. Devereux *et al*. (2004) on comparable U.K. data, Ellison and Glaeser (1997) on U.S. data, or Maurel and Sédillot (1999) on French data find that between 75 and 95% of industries are localized according to the EG index and less than 15% are dispersed. Note however that these papers deal with the localization 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.

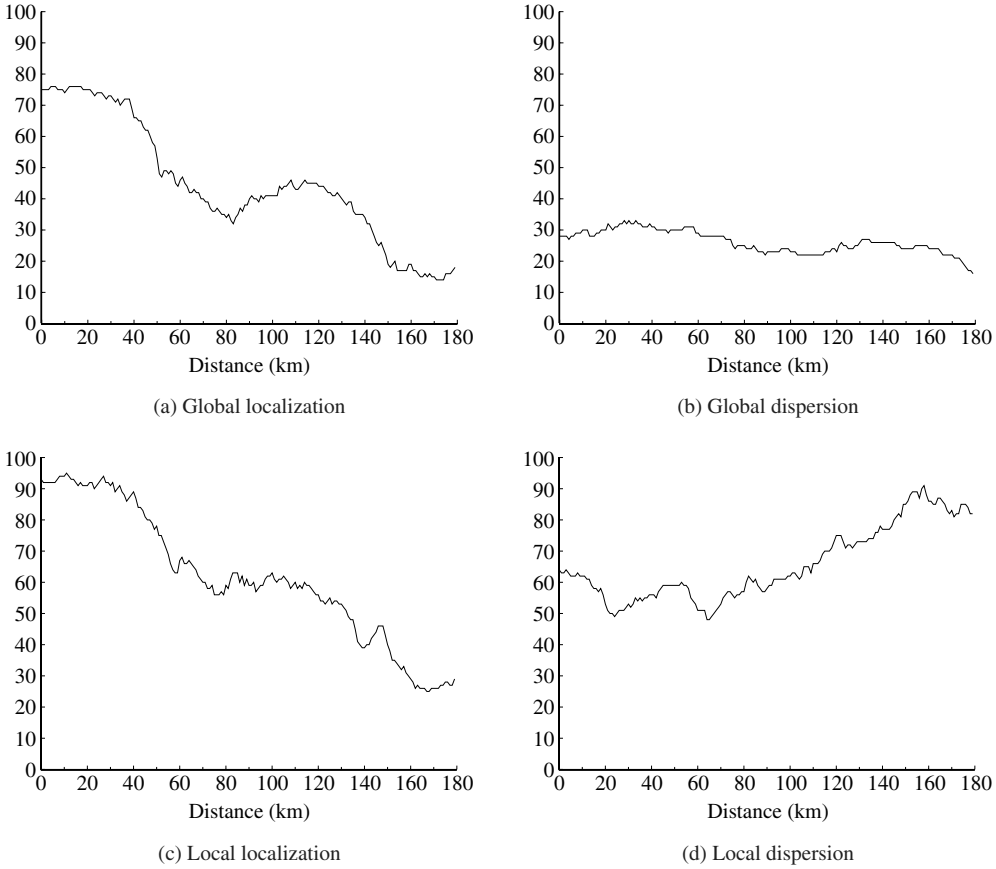
To go further and to look at scale issues, Table 1 considers the fraction of industries which show localization at three thresholds (5, 30 and 150 km). A majority of industries that deviate for any of these three threshold tend to do it for both 5 and 30 km. These results are confirmed when looking more broadly at cross-industry patterns. Figure 3 shows the number of localized and dispersed industries for each level of distance. Both local and global localization results show roughly similar patterns for localization. At low distances, a significant proportion of industries are localized. The number of localized industries is on a high plateau between 0 and 40 km, then falls sharply with distance up to around 80 km and then begins to rise again with a second and lower peak between 100 and 120 km. These findings regarding the scales at which localization takes place are markedly different from those of previous studies based on the EG index. They find that in the U.S. industrial localization is persistently stronger for states than counties and stronger for counties than ZIP-codes (Ellison and Glaeser (1997), Rosenthal and Strange (2001)). Dispersion shows very different patterns. Global dispersion tends to occur equally across all distances. In contrast, local dispersion (which includes industries that are localized) tends to rise slowly with distance as a result of 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. We can base a measure of localization at any given distance on the index of localization $\Gamma_A(d)$ defined in equation (4). We construct the measure $\Gamma(d) \equiv \sum_A \Gamma_A(d)$, by summing the index of localization across all industries for each level of distance. Similarly, we can construct a measure of the extent of cross-industry dispersion, $\Psi(d) \equiv \sum_A \Psi_A(d)$ using the index of dispersion $\Psi_A(d)$. Figure 4 reports both measures for the 234 industries. Note that the measures are directly comparable

TABLE 1

Localization at three thresholds for four-digit industries

Percentage of four-digit industries localized at:				
5 km	5 km only	5 and 30 km only	5 and 150 km only	5, 30 and 150 km
39·3	6·4	22·6	0·9	9·4
30 km	30 km only	30 and 150 km only		
38·9	6·0	0·9		
150 km	150 km only			
17·1	6·0			



(a) Global localization (b) Global dispersion

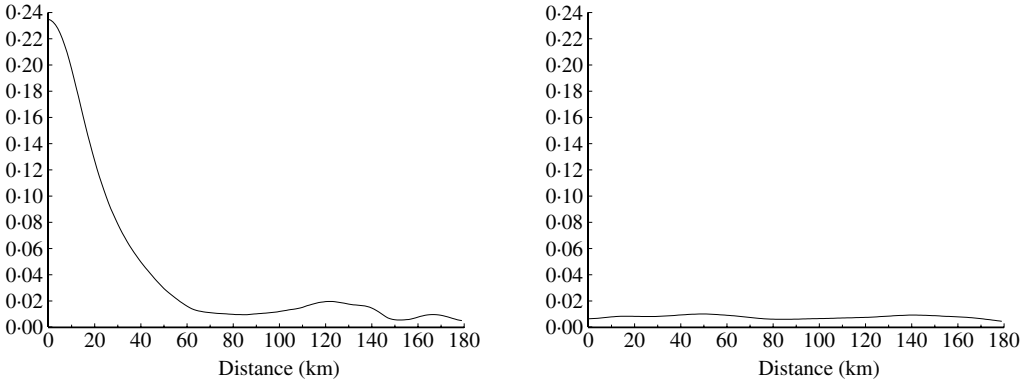
(c) Local localization (d) Local dispersion

FIGURE 3

Number of four-digit industries with local/global localization and dispersion

across distances, but not across the two figures.[13] It is immediately apparent that the extent of localization is much greater at small distances than large distances. As before, dispersion does not show any marked pattern. The important conclusion we draw here is that localization tends to take place mostly at fairly small scales.

13. This is because for an industry that exhibits localization the density is unbounded from above whereas the density of an industry that exhibits dispersion is bounded from below by zero.



(a) Global localization (b) Global dispersion

FIGURE 4
Index of global localization and dispersion by distance

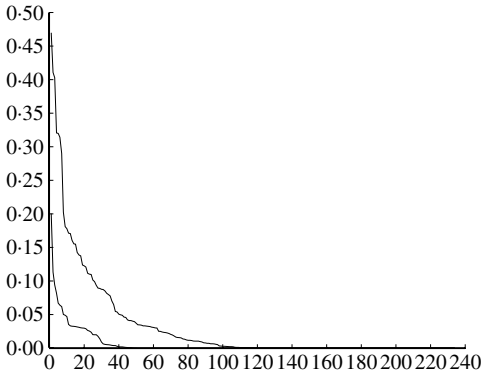


FIGURE 5
Distribution of global localization and dispersion by four-digit industries

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, 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 localization and dispersion across all levels of distance. To illustrate the variations in industry outcomes, we rank industries by decreasing order of these indices and plot them in Figure 5. The upper line is the measure of localization, the lower that of dispersion. As is immediately clear, there are a few industries that show very high localization or dispersion, but the majority of industries do not see such extreme outcomes. This highly skewed distribution of localization confirms previous findings (Ellison and Glaeser (1997), Maurel and Sédillot (1999), Devereux *et al*. (2004)).

To give some idea of the reality underlying Figure 5, Table 2 lists the 10 most localized industries and the 10 most dispersed. Interestingly, more than a century after Marshall (1890), Cutlery (SIC2861) is still amongst the most localized industries. Six textile or textile-related

TABLE 2

Most localized and most dispersed four-digit industries

SIC92	Industry	Γ or Ψ
Most localized		
2214	Publishing of sound recordings	0·470
1711	Preparation and spinning of cotton-type fibres	0·411
2231	Reproduction of sound recordings	0·403
1760	Manufacture of knitted and crocheted fabrics	0·321
1713	Preparation and spinning of worsted-type fibres	0·319
2861	Manufacture of cutlery	0·314
1771	Manufacture of knitted and crocheted hosiery	0·290
1810	Manufacture of leather clothes	0·203
1822	Manufacture of other outerwear	0·181
2211	Publishing of books	0·178
Most dispersed		
1520	Processing and preserving of fish and fish products	0·200
3511	Building and repairing of ships	0·113
1581	Manufacture of bread, fresh pastry goods and cakes	0·094
2010	Saw milling and planing of wood, impregnation of wood	0·082
2932	Other agricultural and forestry machinery	0·067
1551	Operation of dairies and cheese making	0·064
1752	Manufacture of cordage, rope, twine and netting	0·062
3615	Manufacture of mattresses	0·050
1571	Manufacture of prepared feeds for farm animals	0·049
2030	Manufacture of builders' carpentry and joinery	0·047

industries are also in the same list together with three media-based industries. These highly localized industries are fairly exceptional. In contrast, the mean industry (after ranking industries by their degree of localization) is barely more localized than if randomly distributed. It is mostly food-related industries together with industries with high transport costs or high dependence on natural resources that show dispersion.

Our main focus in this paper is on the proportion of manufacturing sectors that are localized. However, it is interesting to notice that a number of industries that appear in Table 2 are fairly small in terms of overall employment. This raises the question as to whether the percentage of manufacturing workers employed in localized industries is above or below the percentage of sectors that are localized. Weighting sectors by their share in manufacturing employment, we find that 67% of U.K. manufacturing employers work in sectors that are localized. This shows that localized sectors tend to have a larger share of manufacturing employment. Offsetting this, however, is the fact that the employment share weighted mean of the index of globalization, Γ_A, is 30% lower than the unweighted mean of the index. That is, larger sectors tend to be less strongly localized.

Finally, it is also interesting to notice that for many (two-digit) branches, related industries within the same branch tend to follow similar patterns. Table 3 breaks down localization of industries by branches. For instance nearly all Food and Drink industries (SIC15) or Wood, Petroleum, and Mineral industries (SIC20, 23 and 26) are not localized. By contrast, most Textile, Publishing, Instrument and Appliances industries (SIC17–19, 22 and 30–33) are localized. The two main exceptions are Chemicals (SIC24) and Machinery (SIC29). In these two branches, however, the more detailed patterns are telling. Chemical industries such as Fertilisers (SIC2415) vertically linked to dispersed industries are also dispersed whereas those like Basic Pharmaceuticals (SIC2441) or Preparation of Recorded Media (SIC2465) vertically linked to localized industries are themselves very localized. The same holds for machinery: Other

TABLE 3

Localization by two-digit branch

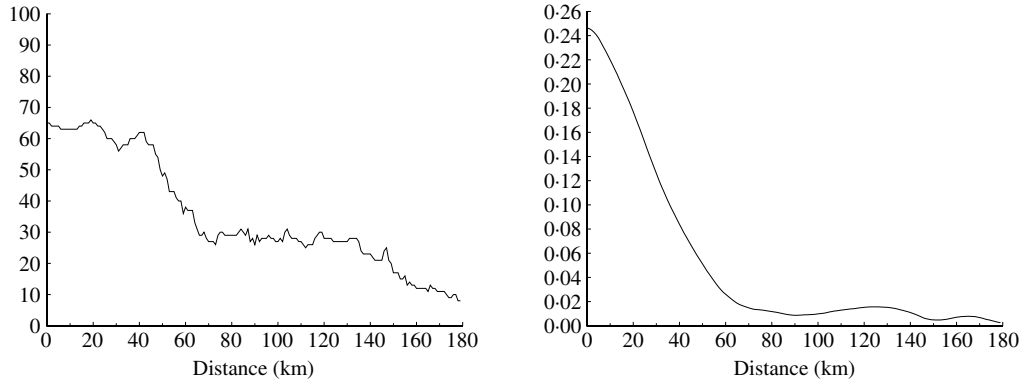
	Two-digit branch	Number of four-digit industries	No. global localization ≤60 km	No. global localization >60 km
15	Food products and beverages	30	1	0
16	Tobacco products	1	1	0
17	Textiles	20	16	9
18	Wearing apparel, dressing, etc.	6	6	3
19	Tanning and dressing of leather, footwear	3	3	3
20	Wood and products of wood, etc.	6	0	0
21	Pulp, paper and paper products	7	2	1
22	Publishing, printing and recorded media	13	13	8
23	Coke, refined petroleum products	3	0	0
24	Chemical and chemical products	20	8	8
25	Rubber and plastic products	7	1	3
26	Other non-metallic mineral products	24	4	2
27	Basic metals	17	11	10
28	Fabricated metal products	16	9	12
29	Other machinery and equipment	20	6	9
30	Office machinery and computers	2	2	2
31	Electrical machinery	7	2	5
32	Radio, televisions and other appliances	3	3	3
33	Instruments	5	3	4
34	Motor vehicles, trailers, etc.	3	1	3
35	Other transport equipment	8	2	2
36	Furniture and other products	13	4	5
	Aggregate	234	98	92

Agricultural and Forestry Machinery (SIC2932) is very dispersed like most agriculture-related industries, whereas Machinery for Textile, Apparel and Leather Production (SIC2954) is very localized like most textile industries.

5. ESTABLISHMENT SIZE AND LOCALIZATION

Four main conclusions emerge so far: (i) 52% of industries are localized, (ii) localization mostly takes place at small scales, (iii) deviations from randomness are very skewed across industries and (iv) industries that belong to the same branch tend to have similar localization patterns. These findings may be driven by particular types of establishments or particular sectoral definitions. To gain insights about the size of localized establishments and the scope of localization, we 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 may be 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.



(a) Number of localized industries

(b) Index of global localization

FIGURE 6
Global localization when censoring for the smallest establishments

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 and large plants differ. 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. As discussed in Section 2 imposing the same absolute threshold across industries is problematic given that 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, 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.[14]

The first key finding is that after censoring the smallest plants only 43% of industries (against 52% in the baseline simulations) show any amount of global localization. At the same time however, the index of localization, $\Gamma(d)$, is slightly above that in the baseline simulations at short distances despite there being fewer establishments and thus larger confidence bands. Hence, in some industries, localization is stronger when smaller establishments are ignored whereas in others, small plants are the main drivers of localization.

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 localized industries is large between 0 and 40 km and then decreases to reach a low plateau after 60 km. From the same figure, the index of localization, $\Gamma(d)$, also follows a pattern similar to that when all plants are considered.

Industries continue to show very different location patterns and a very skewed distribution of both localization and dispersion. When comparing these results with our baseline across industrial branches, we note that the declines in localization are concentrated in Publishing (SIC22), Chemicals (SIC24), Computers (SIC30), and Radios and TVs (SIC32). This is evidence

14. 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.

that in these branches, localization is driven by small establishments. On the other hand, in a few textile industries, that are among the most localized industries when treating all plants symmetrically, the index of localization, Γ, increases by more than 50% when the smallest plants are ignored. This is observed not only in Textile industries (SIC17) 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 dispersion. Only 19% of industries (against 24% in the baseline) exhibit global dispersion when censoring for the smallest establishments. Overall, the location patterns of small establishments vary a lot across industries but in general are more extreme with stronger tendencies towards either localization or dispersion.

Four-digit industries when weighting for employment

As we made clear above, censoring for the smallest establishments sheds light on their location patterns when comparing the results with those obtained for the whole population of plants. However this approach does not allow a detailed analysis of location patterns for larger establishments. It also maintains the establishment as the basic unit of observation. In some instances it is interesting to take instead the worker as the unit of observation. In this case we need to consider the bilateral distances between all pairs of workers who belong to different establishments.

Denoting the employment of firm i by $e(i)$, the estimator of the K-density becomes

$$\widehat{K}_A^{\text{emp}}(d) = \frac{1}{h \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} e(i)e(j)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} e(i)e(j) f\left(\frac{d - d_{i,j}}{h}\right) \qquad (6)$$

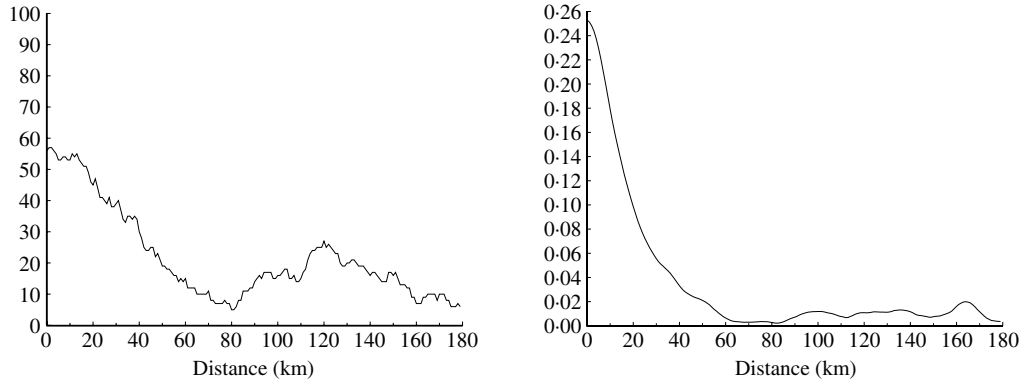
where bandwidth, h, and kernel function, f, are chosen as for equation (1). Note that we do not consider the zero distances between employees in the same plant so localization cannot be driven by the concentration of employment within a particular plant but could be driven by a few large plants located close to each other.[15]

Turning to the results, only 43% of four-digit industries (against 52% in the baseline) exhibit some global localization. The scales at which these deviations occur are very similar to those observed before. According to Figure 7 when weighting for employment, localization is even more strongly biased in favour of short distances (below 50 km). At the same time, as in the previous analysis, the total amount of localization is higher than in the baseline simulations for distances below 50 km.[16] Hence when weighting for employment, fewer industries are localized but those that are deviate more strongly from randomness. Results with respect to dispersion are very close to those in the baseline simulations: 22% of all industries are dispersed (against 24% in the baseline).

Industries are still highly heterogeneous with respect to their localization/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), Electrical machinery (SIC31), Radio and TVs (SIC32), and Instruments (SIC33) localization is less prevalent than in the baseline. For instance, five in six

15. According to equation (6), the distance between two establishments with 10 employees each is given 100 times the weight of the distance between two establishments with a single employee each. This multiplicative weighting in equation (6) gives a lot of weight to large establishments close to one another. Milder forms of weighting are possible. For instance it could be possible to use an additive specification, $e(i)+e(j)$, rather than a multiplicative one. The thought experiment corresponding to this specification would, however, be less clear since it implicitly assumes a "link" between each worker in an establishment and each neighbouring establishment.

16. 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 (6) both the numerator and the denominator are weighted in the same fashion.



(a) Number of localized industries

(b) Index of global localization

FIGURE 7

Global localization when weighting establishments by their employment

industries in Electrical machinery show localization in the initial results whereas only one still shows localization when weighting establishments by employment. In two other branches, Food and beverages (SIC15) and Other non-metallic mineral products (SIC26), the exact opposite happens. For instance, only one food industry in the baseline shows localization while five do when weighting by employment.

These findings are fully consistent with those obtained when censoring for smaller firms. Furthermore findings reported in Holmes and Stevens (2002) allow a comparison with U.S. manufacturing although we note that his comparison should be interpreted with caution given differences in the methodologies employed. Holmes and Stevens (2002) examine the location patterns of large plants in U.S. manufacturing using the EG index. They find that large plants tend to be more localized than their whole industry. In broad agreement with this tendency, we observe an increase in our index of localization, $\Gamma(d)$, for the U.K. when censoring for small plants. However, in contrast with U.S. findings, localization in the U.K. is driven by large firms in only some industrial branches.

More generally, it must be emphasized that taking plant size into account reinforces the four main conclusions obtained so far.[17] Localization is detected in at most half of the industries. Deviations still occur at a scale of 0–50 km. There is still a lot of cross-industry heterogeneity with respect to localization and dispersion. This is compounded by cross-industry differences in location patterns between small and large establishments. Finally we observe broad patterns of clustering of small vs. large establishments by industrial branch.

Before turning to a detailed comparison between our approach and the EG index, it must be noted that the two approaches developed here to examine patterns of localization by establishment size could be further refined. Instead of considering only one threshold, we could consider finer classes of establishment sizes in each industry. The counterfactuals could also be modified. For instance, imagine that establishment size constrains location choices to a set of "appropriate" sites. Then, we could construct counterfactuals that only allow large firms to locate on large sites and small firms to locate only on sites currently occupied by a small firm. These questions as well as broader issues about which type of establishment localize (*e.g.* independent

17. Interestingly when industries are ranked by decreasing localization the Spearman rank correlation when weighting by employment with the baseline is 0·77 whereas that with the ranking when censoring for establishment size is 0·74.

vs. affiliated, indigenous vs. foreign owned, etc.) are explored in Duranton and Overman (2005). The important thing to note here is that our technique is flexible enough to accommodate for these variants and this makes it possible to explore many other questions.

Comparison with the EG index

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

$$EG_A \equiv \frac{g_A - H_A}{1 - H_A}, \tag{7}$$

where $H_A \equiv \sum_j x_A(j)^2$ is the Herfindahl 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 localization index equal to

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

$s_A(i)$ 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 localization. Ellison and Glaeser (1997) also argue that a value between 0 and 0·02 signals weak localization and anything above 0·05 is interpreted as a strong tendency to localize. To compare with our methodology, we apply this index to the 120 postcode areas of the U.K. (without Northern Ireland) using the total population of plants. Note that postcode areas are on average less populated than U.S. states but larger than U.S. counties.

The mean value of the EG index across 234 U.K. industries is 0·034 and the median is at 0·011. These figures are above their corresponding values for U.S. counties but below those of U.S. states according to Ellison and Glaeser's (1997) calculations. We also find that 94% of U.K. industries have a positive EG index and thus exhibit some localization. This proportion is very close to that obtained by Ellison and Glaeser (1997) for the U.S. (97%).

As the EG index not only controls for the lumpiness of plants but also for their size distribution, it is *a priori* best compared to our results when plants are weighted by employment. The contrast is strong since we find 43% of industries to be localized (against 94%) and 22% to be dispersed (against 6%). 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 localized, suggesting that Ellison and Glaeser's (1997) definition of weak localization (EG index above 0 but less than 0·02) is probably not appropriate. For U.K. manufacturing plants this definition of weak localization mostly defines industries whose location patterns are not *significantly* different from randomness.

Furthermore, in addition to the substantial differences in terms of number of localized industries, individual industries show different outcomes between the two measures. Across all industries, the Spearman rank correlation between the EG index and our ranking is statistically significant at 0·41. Focusing on the most localized industries we find that the two methods agree on only 5 out of 10 industries. Interestingly two publishing industries and Jewelry, which we find to be very localized, are ranked above 30 according to the EG index. Looking at the detailed location patterns of Jewelry and these two publishing industries is informative. Studying the maps for these sectors, we see that they are indeed very localized around London (with a second cluster in Birmingham for Jewelry). The EG index does not capture this localization as the Greater London region is divided into different postcode areas, which are then treated as completely unrelated entities in the calculation of the index.

We believe this comparison highlights 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 localization. We believe this downward border bias is why the EG index

is consistently found to increase with the size of spatial units. On its own, this downward bias would tend to increase the number of localized industries identified by our methodology which avoids this border effect. However, offsetting this is the fact that ignoring the significance of departures from randomness biases existing measures of localization upwards. Our methodology removes border effects and allows for significance, and our results show that the latter effect dominates. Second, the relevant geographical scales for localization emerge naturally from our analysis because we do not need to arbitrarily define the size of units *ex ante*. Existing indices are calculated over only one partition of space, whereas we have shown that different industries localize at different spatial scales. This problem is compounded by the fact that small scales (urban and metropolitan) turn out to be particularly important and this level of aggregation is not very well captured when using spatial units such as U.S. states, European regions, U.K. or U.S. counties or U.K. 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 localization, but that there are sizable differences for 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 varies across industries. Existing indices are narrowly constrained in the way they deal with the distribution of establishment size (*e.g.* through a Herfindahl index in the EG case) whereas our approach is flexible and could easily be extended to other weighting methods.

6. THE SCOPE OF LOCALIZATION

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

Localization of five-digit sub-industries

In the U.K., 33 four-digit industries (out of 239) are subdivided into more finely defined five-digit sub-industries. We consider only the 58 (out of 76) five-digit sub-industries that have more than 10 establishments. Correcting for global confidence bands, we find that 44 of these sub-industries (76%) deviate significantly from randomness. More precisely, 45% are localized, 31% are dispersed, while we cannot reject randomness for the remaining 24%.

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 industrial branches 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. Thirteen out of 58 sub-industries are in Textiles (SIC17) and Apparel (SIC18). These are mostly localized, sometimes highly so, just like their four-digit counterparts. The third large group, accounting for 12 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 findings emerge. First, when the patterns of localization are strong for industries,

they are often even stronger for their sub-industries.[18] 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 localized and one that is only slightly dispersed. This implies that some of the lack of localization 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 allows us to uncover some patterns that were so far hidden.[19] Third, in some instances when an industry shows minor dispersion or localization (in terms of Ψ_A or Γ_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 for approximately the same proportion of industries.

Localization 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 that are localized or dispersed is higher than in the baseline: 58% and 29% respectively against 52% and 24% for four-digit industries.

To gain further insight, it is useful to consider the same three thresholds as we did previously (5, 30 and 150 km). The results are reported in Table 4. Comparing with Table 1, note that for localization at the 5 and 30 km thresholds, the figures are relatively similar. In contrast, for figures relating to the 150 km threshold, localization is 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 a high plateau of localization between 0 and 40 km, then a decline followed by a rise and a second plateau between 80 and 140 km. Although the shape of this curve shares some similarities with its counterpart for four-digit industries represented in Figure 3, the relative number of sectors localized at distances above 80 km is much higher. Dispersion shows a pattern similar to four-digit industries.

In summary, for short distances localization is as frequent in sectors as in industries but localization at medium distances is much more prevalent for sectors. The importance of localization above 80 km is a new feature arising at this level of sectoral aggregation. This finding is consistent with different processes governing firm location decisions. A first explanation is that firms in industries that are part of the same three-digit sectors may opt for each others' company and co-localize at fairly large spatial scales. Alternatively, four-digit industries may cluster at smaller scales (as seen before) but these clusters then locate next to each other.

Problems also arise when trying to interpret the amount of localization observed for distances below 40 km. It could be the case that what we detect here as localization for three-digit sectors is in fact localization in the four-digit industries within these sectors. For instance,

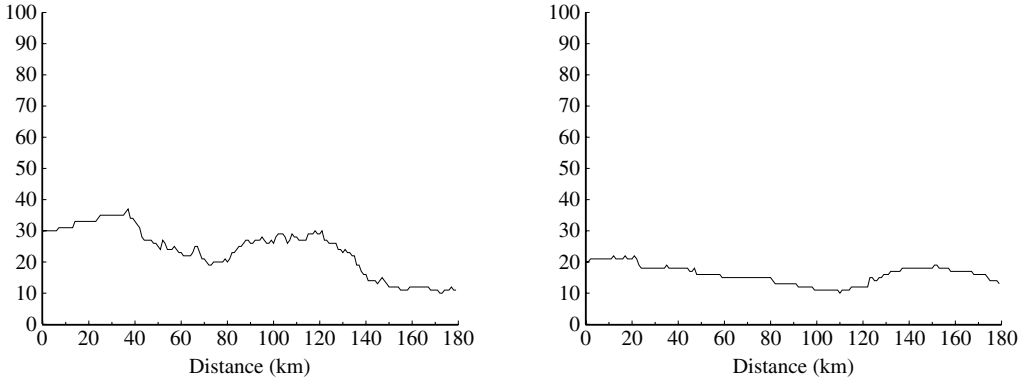
18. However, some interesting details emerge. For example, for clothing industries, plants that produce women's clothing are always more localized than plants producing men's clothing.

19. 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, classification errors become more important as industries are more finely defined. Note also that five-digit sub-industries 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 SIC175, Manufacture of other textiles, is subdivided into four very different industries: Carpets and rugs (1751), Cordage, rope and netting (1752), Non-wovens (1753) and Other textiles (1754).

TABLE 4

Localization at three thresholds for three-digit sectors

Percentage of three-digit sectors localized at:				
5 km	5 km only	5 and 30 km only	5 and 150 km only	5, 30 and 150 km
35·9	5·8	19·4	1·0	9·7
30 km	30 km only	30 and 150 km only		
38·8	7·8	1·9		
150 km	150 km only			
19·4	6·8			



(a) Global localization (b) Global dispersion

FIGURE 8

Number of three-digit sectors with global localization and dispersion

localization in Pharmaceuticals (SIC244) might be driven mostly by the strong tendency of Basic pharmaceuticals (SIC2441) to cluster. Alternatively, this finding could be driven by a tendency for firms across different industries that are part of the same sector to co-localize 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.

Hence with regard to the localization of three-digit sectors, we must contemplate three possible explanations. First, there could be a classification problem where the relevant level of analysis is three-digit sectors instead of four-digit industries. Previous findings for four-digit industries would then reflect what happens in sectors. Second, the classification problem may be in the opposite direction and sectoral localization may just reflect localization of four-digit industries. In this case, the relevant level of analysis is the four-digit industry since sectoral localization is driven by localization in one or more industries within the sector. Third, and more subtly, there may be some location differences between industries in the same sector so that the relevant level of analysis is still the four-digit industry, but at the same time, there may also be some interactions happening between these industries leading plants in different industries to opt for locations close to one another.

To assess these three explanations, we look at the location patterns of industries within sectors. In the next subsection we show that localization is still strong in four-digit industries even after controlling for the location of the three-digit sectors. That is, the second and third

explanations are preferred to the first. Later on, we test the second explanation against the third. Our results there offer some support for both of these explanations.

Localization within three-digit sectors

If the relevant level of aggregation is the three-digit sector, we expect plants in localized industries to locate close to other plants in the same sector irrespective of the industry they belong to. Thus after controlling for the location of the three-digit sector, four-digit industries should exhibit no tendency to localize. Alternatively, if the relevant level of aggregation is the four-digit industry, we still expect to observe industry effects after controlling for the location of the sector.

Whether four-digit industries still show localization after controlling for the overall localization 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 firms 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 sector B instead of sampling from S.[20]

There are 184 industries that are part of a three-digit sector containing at least two industries. Of these, we find that 48% show localization. This is very marginally less than with our baseline. However, for all distances under 150 km fewer industries deviate and the amount of localization captured by Γ is somewhat lower than in the baseline (substantially lower for distances under 60 km). With respect to dispersion, the figure is at 18%, much lower than for the baseline (24%).

From this we conclude that when controlling for the location of the parent sector, industries tend to exhibit slightly less localization and dispersion. Interestingly, we note that for Textile (SIC17), Publishing (SIC22) and Basic metals (SIC27) industries, far fewer industries are localized with respect to their parent sector than general manufacturing. For instance, in Publishing only five industries in 13 are localized in their parent sector against all of them in the baseline analysis.[21] Still, five of the 10 most localized industries in Table 2 remain among the 10 most localized industries when controlling for the location of their parent sectors. However, their Γ has on average a value less than half of that in the baseline.

In conclusion, the three-digit sector to which an industry belongs explains some of its location behaviour but not all. Differences between four-digit industries remain substantial suggesting that we must take them as our basic level of analysis. However, further analysis is necessary to understand the strong sector effects we observe. It could be that plants in related industries have a tendency to localize in the same areas but that location decisions are independent across the industries. Thus this localization may be a purely random outcome of clusters happening to be close to each other. Alternatively the factors driving localization in these industries could share some similarities and thus lead their establishments to cluster together. These two types of explanation are similar in that they both suggest that plants in related industries have independent location patterns (*i.e.* the location of one industry does not directly influence that of another). In this case, we can talk of *joint-localization*. The opposite case is where establishments in an industry may decide to locate close to establishments in related industries. In this case location patterns across industries are no longer independent and we can speak of *co-localization*.

20. 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 EG index of the sector and the weighted average of the EG indices of each individual industry.

21. The reason is that most industries in Publishing are based around London. The same holds for Birmingham and Basic metals, Manchester and Textiles, etc.

Localization across industries within three-digit sectors

Despite some attention in the recent literature (Ellison and Glaeser (1997), Devereux *et al*. (2004)) and its crucial importance with respect to many theoretical and policy concerns, very little is known about co-localization.[22] Although it is easy to define co-localization as the tendency of different industries to opt for each other's company and as a result to cluster together, measuring co-localization and distinguishing from joint-localization is much more complex than analysing localization. To see why, consider the following generalized version of equation (1):

$$\widehat{K}_{(A,B)}(d) = \frac{1}{P(n_A, n_B)h} \sum_{i=1}^{n_A} \sum_{\substack{j=1 \\ j \neq i}}^{n_B} f\left(\frac{d - d_{i,j}}{h}\right) \tag{9}$$

where bandwidth (h) and kernel function (f) are chosen as for equation (1), 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 different establishments with one establishment from each subset.[23] The density $\widehat{K}_{(A,B)}(\cdot)$ is a straightforward generalization of $\widehat{K}_A(\cdot)$ which allows us to calculate the density of bilateral distances between establishments in any two subsets from a population.

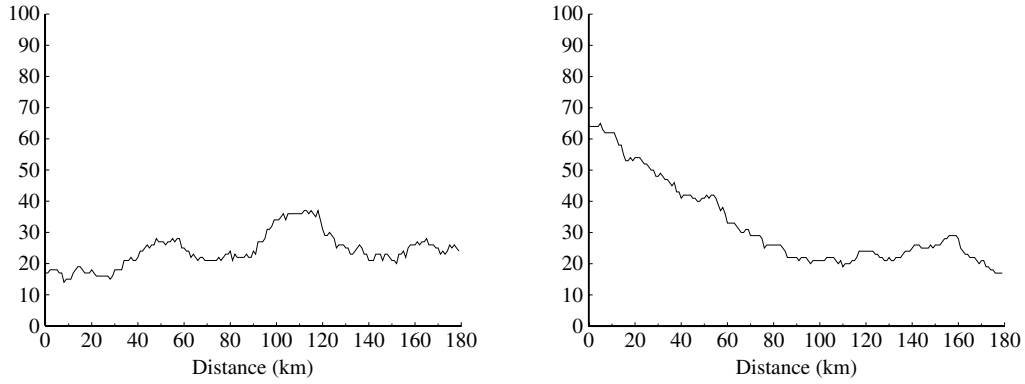
Then, it remains to define the counterfactuals which this distribution should be compared to. In this respect, note that tests over $\widehat{K}_{(A,B)}(\cdot)$ may involve counterfactuals $\tilde{A}$ and $\tilde{B}$ drawn from any subset of S. *A priori* this implies a considerable number of possible tests. Many of these tests are not very informative with respect to co-localization. For instance calculating the bilateral distance density using equation (9) for two 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 localization in two industries can also generate some form of joint-localization across the same two industries. A measure based on sampling from the overall population cannot distinguish between co-localization and joint-localization and thus confuses the two explanations.

Among all the possible tests one could construct using (9), we believe one exercise is of particular interest here—to see whether there is some tendency for localization across different four-digit industries after controlling for the overall tendency of the establishments of both industries to cluster. To investigate this, we can apply equation (9) to any two four-digit industries, A and B that are part of the same three-digit industries, and sample our counterfactuals from the set of all sites occupied by an establishment in either of these industries, $A \cup B$. The intuition behind this test is very simple. It allows us to determine whether plants in an industry (A) have a tendency to be closer to plants in related industries (B) that are part of the same sector than to plants in the same industry (A). For instance after controlling for the localization tendency of the partition composed of Basic Pharmaceuticals and Pharmaceutical Preparations (SIC2441 and 2442), it allows 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 upper deviations from randomness mean that establishments in these industries are attracted to each other even after controlling for whatever tendency they have to cluster. At the same time, being unable to reject randomness is no rejection of industries being co-localized in some sense.

There are 317 possible pairs of four-digit industries that are part of the same three-digit sectors. Local co-localization is detected in 56% of the pairs but the proportion falls to 34%

22. The term co-agglomeration is also used in the literature. For consistency with respect to the terminology used so far, we speak of co-localization. Note however that the existing literature does not distinguish between joint-localization and co-localization.

23. 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 ($A \cap B = \oslash$) then $P(n_A, n_B) = n_A n_B$.



(a) Number of industry pairs with co-localization

(b) Number of industry pairs with co-dispersion

FIGURE 9

Global co-localization and separation among industry pairs in the same sectors

when controlling for global confidence bands. Furthermore, the extent of co-localization, as measured by Γ is never very large. Hence there seems to be a fairly widespread tendency for related industries to co-localize but the tendency is not of overwhelming intensity. Turning to lower deviations, or co-dispersions, they occur locally in 31% of industry pairs and 29% when controlling for global confidence bands. The tendency for pairs of industries to separate is thus slightly less widespread than their tendency to co-localize but it takes place with greater intensity when it does (reflecting related industries which cluster in very different areas).

The spatial scales at which these phenomena take place are highly revealing. Figure 9 plots the number of pairs of industries with co-localization and co-dispersion for each level of distance. Co-localization peaks between 100 and 120 km whereas co-dispersion declines continuously between 0 and 60 km and then remains stable at a very low level. These results are fully consistent with the facts that four-digit industries seem to cluster mostly at fairly low spatial scales whereas three-digit industries also show some localization at medium scales.

The final picture we obtain is thus one where the relevant unit of industrial aggregation is at least four-digit industries (if not something more disaggregated). These industries have a tendency to localize at small spatial scales. At the same time however, there are some interactions taking place across industries that are part of the same branches with a tendency for plants to locate closer to plants in other (related) industries than to plants in their own industry.

7. CONCLUSION

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

This approach, to our knowledge, is entirely new in economics. The use of K-functions (*i.e.* the cumulatives of our K-densities) is widespread in quantitative geography (Ripley (1976), Cressie (1993)). The novelty here lies in the use of smoothing techniques (Silverman, 1986) allowing us to look at distance densities rather than their cumulatives. We believe the K-densities are more informative than K-functions with respect to the scale of localization. The standard

approach in quantitative geography also uses homogeneous spatial Poisson processes to generate counterfactuals.[24] We differ from this by sampling randomly from a set of sites which allows us to control for the overall distribution of manufacturing.[25] Finally, we construct proper confidence intervals and confidence bands instead of taking the envelope of a small number of simulations.

We applied our tests to an exhaustive U.K. manufacturing data-set. Our main findings are:

- 52% of four-digit industries exhibit localization at a 5% confidence level and 24% of them show dispersion at the same confidence level.
- Localization in four-digit industries takes place mostly between 0 and 50 km.
- The extent of localization and dispersion are very skewed across industries.
- Four- and five-digit industries follow broad sector and branch patterns with respect to localization.
- In some industrial branches, localization at the industry level is driven by larger establishments, whereas in others it is smaller establishments which have a tendency to cluster.
- Localization and dispersion are as frequent in three-digit sectors as in four-digit industries for distances below 80 km. Three-digit sectors also show a lot of localization at the regional scale (80–140 km) due, at least in part, to the tendency of four-digit industries to co-localize at this spatial scale.

Some of these results 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. For instance, we do not find as many industries to be localized as previously claimed. Along the same lines, we find the propensity for 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 more nuanced than earlier studies have led us to believe. Finally our results about the scale and the scope of localization 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 localization, dispersion, and co-localization. For instance, one may wish to compare the behaviour of independent plants with that of affiliated plants or that of foreign plants vs. domestic ones. Also, much remains to be learnt about co-localization in vertically linked industries, etc. We hope to be able to shed light on these questions in future research. Furthermore, we would like to see our approach replicated for other countries. Data availability is certainly an issue here. Note however that several countries including Canada and France are currently developing geo-coded data similar to those available in the U.K. For many countries including the U.S., such developments are less likely in the very near future.[26] However, in these countries establishment level data are often available at a very fine level of geographic aggregation like census tracts

24. That is they assume that under the null hypothesis of randomness the probability of a point occurring in any region R is proportional to the area of R. Marcon and Puech (2003) use these standard approaches based on homogeneous spatial Poisson processes to study industrial location in France. Their paper clearly demonstrates the problems with applying standard techniques directly to the study of localization. For example, their results suggest that all industries are concentrated—hardly surprising given the fact that aggregate activity in France is very concentrated around Paris!

25. With our emphasis on Monte Carlo simulations to allow for the uneven distribution of aggregate activity, our approach shares some similarities with very recent work on the use of case-control counterfactuals for the application of spatial point patterns in epidemiology. In that literature, however, counterfactuals are constructed by taking the location of individuals as given and then randomly deciding whether individuals are healthy. We believe that our approach of taking a firm's sector as given and treating location as random is a more suitable thought experiment when studying localization. See Diggle (2003).

26. See Holmes and Stevens (2004) for a detailed discussion of North American data sources.

or zipcode areas in the U.S. Our methodology can easily be adapted to such data by randomly allocating establishments within the smallest available spatial units. In this case, the error made on the location is of the same magnitude as the square root of the area of the smallest spatial unit. Provided the areas are relatively small, we suspect that this error will be far outweighed by the improvement with respect to meeting the five requirements that we outline above.

In future work, we also wish to develop our distance-based approach to test more sophisticated models of industrial location. So far we have only tested the simplest non-trivial model of industrial location: randomness conditional on overall manufacturing agglomeration. It is obvious that different counterfactuals could be generated by using more sophisticated theoretical frameworks. These theories could be tested as here, by looking at deviations from confidence bands computed from counterfactuals (again to be generated from the theories to be tested). Alternatively, one could use our indices of localization as endogenous variables and regress them on a set of industry characteristics.

As a final point, note that distance-based analyses can be applied beyond industrial geography. Any data with detailed geographical information readily lends themselves 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 address new questions that could not previously be tackled.

APPENDIX: A MODEL

Consider N establishments and Z, a set of points on the Euclidean plane whose cardinal is much larger than N. The elements of Z are called potential sites. Denote by $d_{j,k}$ the Euclidean distance between the potential sites j and k. Z is such that

$$\forall j \in Z, d_{O,j} \leq \overline{d}, \tag{A.1}$$

where $\overline{d}$ is the maximum distance between a potential site and the origin O. Each potential site j for establishment i has a value equal to

$$V_{j(i)} = v_j + \epsilon_{i,j} + g_{j,A(i)}, \tag{A.2}$$

where v_j is the intrinsic value of the potential site, $\epsilon_{i,j}$ is a positive i.i.d. random component, and $g_{j,A(i)}$ is an industry-specific component which applies to all establishments which belong to industry A. The component v_j reflects the fact that some potential sites are more desirable irrespective of one's industry. The random component $\epsilon_{i,j}$ captures all the factors that are idiosyncratic to establishment i with respect to the potential site j. Finally, the industry-specific component is such that

$$g_{j,A(i)} = \sum\nolimits_k G(d_{j,k}), \tag{A.3}$$

where the k are attraction points in the Euclidean space for all establishments in industry A. The decay function $G(\cdot)$ is such that $G' < 0$ and $G'' > 0$. Attraction points capture the effect of localized natural endowments. Establishments gain from being closer to these attraction points. The location decisions among the potential sites depend thus on three factors, their intrinsic values which apply symmetrically to all establishments, their idiosyncratic values which apply differently to all establishments and the location value g which applies symmetrically to all establishments in the same industry.

Potential sites are allocated sequentially. The equilibrium is such that the first establishment occupies the site it values most. The second establishment occupies the site it values most among the remaining ones, etc. Denote by S the set of all occupied sites and S_A the set of occupied sites by the establishments of industry A in equilibrium. Note that the industry-specific location value is defined by fixed natural endowments. The model can however be easily generalized to spatial distortions generated by local externalities (see Ellison and Glaeser (1997), for a justification and details on the observational equivalence between fixed natural endowments and local externalities). The only complication is that local externalities, which make relative (and no longer absolute) location matter, would leave some room for multiple equilibria.

The null hypothesis for industry A is

$$H_0 : g_{\cdot,A} = 0. \tag{A.4}$$

This null is equivalent to S_A being a random sample of Z in equilibrium. To reject this, it is sufficient to have the K-density for industry A to be different from that of an industry with the same number of establishments occupying sites randomly sampled from Z. Unfortunately, we do not observe Z, the set of potential sites.

Our empirical strategy relies on the fact that Z can be proxied by S. For this approximation to be valid we need (i) all industries to be small with respect to overall manufacturing so that we have a large number of them, (ii) that the location of attraction points must be independent across industries and (iii) $\mathrm{Var}(g) \ll \mathrm{Var}(v + \epsilon)$ in almost all industries. The last condition requires that the effects of natural endowments must not be so strong as to make some industries cluster very tightly. Otherwise, the proportion of very short bilateral distances between elements of S would be larger than that of Z. Conditions (i)–(iii) indicate that mild clustering in industries is not a worry provided industries are small and clustering across industries is independent. In this case, the law of large number applies and S is approximately a random sample of Z. It is then possible to sample counterfactuals from S rather than Z.

Acknowledgements. Thanks to Oriana Bandiera, Roger Bivand, Ian Gordon, Hidehiko Ichimura, Barry McCormick, Tomoya Mori, Diego Puga, Danny Quah, Steve Redding, Stuart Rosenthal, Bernard Salanié, Tony Venables, Qiwei Yao, two anonymous referees, and to seminar participants at a CEPR network meeting in Villars, a CEPR workshop in Barcelona, the RSAI 2001 annual meetings in Charleston, the RGS 2003 annual meetings in London, The Norwegian School of Economics (Bergen), University of Glasgow, Université de Lyon, Université de Namur, University of Nottingham, University of Sussex, University of Toronto, and the London School of Economics for comments and discussions which greatly helped us clarify our minds on crucial steps of the paper. We are also grateful to Rachel Griffith, Felix Ritchie, and Helen Simpson for helping us with the data and to Nick Gill for his first-class research assistance. Finally, this project could not have started without pump-priming financial support from STICERD. Financial support from the Economic and Social Research Council (ESRC Grant R000239878) and the Leverhulme Trust is also gratefully acknowledged. This paper uses data that are Crown Copyright and reproduced with the permission of the HMSO Controller and the Queen's Printer for Scotland. However, the use of these data does not imply ONS endorsement of either the interpretation or the analysis in this paper.

REFERENCES

COMBES, P.-P. and LAFOURCADE, M. (2005), "Transport Costs: Measures, Determinants, and Regional Policy Implications for France", *Journal of Economic Geography*, **5** (3), 319–349.
CRESSIE, N. A. C. (1993) *Statistics for Spatial Data* (New York: John Wiley).
DAVISON, A. C. and HINKLEY, D. V. (1997) *Bootstrap Methods and their Application* (Cambridge: Cambridge University Press).
DEVEREUX, M. P., GRIFFITH, R. and SIMPSON, H. (2004), "The Geographic Distribution of Production Activity in the UK", *Regional Science and Urban Economics*, **35** (5), 533–564.
DIGGLE, P. J. (2003) *Statistical Analysis of Spatial Point Patterns* (New York: Oxford University Press).
DURANTON, G. and OVERMAN, H. G. (2005), "Detailed Location Patterns of UK Manufacturing Industries" (Processed, London School of Economics).
EFRON, B. and TIBSHIRANI, R. J. (1993) *An Introduction to the Bootstrap* (New York: Chapman and Hall).
ELLISON, G. and GLAESER, E. L. (1997), "Geographic Concentration in US Manufacturing Industries: A Dartboard Approach", *Journal of Political Economy*, **105** (5), 889–927.
GRIFFITH, R. (1999), "Using the ARD Establishment Level Data: An Application to Estimating Production Functions", *Economic Journal*, **109** (456), F416–F442.
HOLMES, T. J. and STEVENS, J. J. (2002), "Geographic Concentration and Establishment Scale", *Review of Economics and Statistics*, **84** (4), 682–690.
HOLMES, T. J. and STEVENS, J. J. (2004), "Spatial Distribution of Economic Activities in North America", in V. Henderson and J.-F. Thisse (eds.) *Handbook of Regional and Urban Economics*, Vol. 4 (Amsterdam: North-Holland).
HOOVER, E. M. (1937) *Location Theory and the Shoe and Leather Industries* (Cambridge, MA: Harvard University Press).
MARCON, E. and PUECH, F. (2003), "Evaluating the Geographic Concentration of Industries using Distance-Based Methods", *Journal of Economic Geography*, **3** (4), 409–428.
MARSHALL, A. (1890) *Principles of Economics* (London: Macmillan).
MAUREL, F. and SÉDILLOT, B. (1999), "A Measure of the Geographic Concentration of French Manufacturing Industries", *Regional Science and Urban Economics*, **29** (5), 575–604.
QUAH, D. T. (1997), "Empirics for Growth and Distribution: Stratification, Polarization, and Convergence Clubs", *Journal of Economic Growth*, **2** (1), 27–59.
RAPER, J. F., RHIND, D. W. and SHEPHERD, J. W. (1992) *Postcodes: The New Geography* (Harlow, Essex: Longman Scientific and Technical).
RIPLEY, B. D. (1976), "The Second-Order Analysis of Stationary Point Processes", *Journal of Applied Probability*, **13** (2), 255–266.
ROSENTHAL, S. A. and STRANGE, W. C. (2001), "The Determinants of Agglomeration", *Journal of Urban Economics*, **50** (2), 191–229.
SILVERMAN, B. W. (1986) *Density Estimation for Statistics and Data Analysis* (New York: Chapman and Hall).
YULE, G. U. and KENDALL, M. G. (1950) *An Introduction to the Theory of Statistics* (London: Griffin).