Measuring the Localisation of Knowledge Spillovers with a Text-based Matching Estimator

April 22, 2025

Abstract

I re-examine the geographic localisation of knowledge spillovers using patent citations and a text-based matching approach. Using data from the United States Patent and Trademark Office (USPTO) from 2005 to 2010, I encoded patent abstracts as vectors representing their meaning using a contextual embeddings model. Following Jaffe, Trajtenberg and Henderson (1993), I compared the distance of patents that cite each other to a control group of those that do not. My primary contribution is to use these embeddings to proxy for unobserved characteristics in the matching estimator. I find strong evidence of localisation across all Cooperative Patent Classification (CPC) sections, with spillovers decaying quickly. I also observe secondary localisation patterns at medium distances (500-1000km), potentially reflecting broader agglomeration clusters described in theoretical works.

1. Introduction

Researchers have conceptualised innovation as a key economic force throughout the discipline's history. It is most prominent as a central component of economic growth, where technological change drives sustained increases in living standards. However, measurement and methodological barriers hampered research that directly evaluates the empirical aspects of innovation. While some innovation inputs, such as R&D spending, are directly observed, knowledge is not easily quantified, and the complexity of the networks governing its creation and dissemination poses a significant identification challenge.

The literature has attempted to overcome these empirical challenges in two primary ways: with measures of investment in innovation and through patent data. The second, which I focus on here, comprises a complete census of all patenting activity under a patent office and contains extensive information about inventions, inventors, and intellectual property rights holders. Although patents are only intermediate products in the innovation process, they offer a direct insight into how knowledge spreads. Each patent must cite prior art to demonstrate how it stands above it.[1]

One focal point for research using citation data has been testing whether knowledge spillovers are geographically localised. In the case of patent citations, this is akin to asking if inventors who cite each other also tend to be near each other once we control for other drivers of spatial concentration. I re-test this hypothesis with data from the United States Patent and Trademark Office (USPTO) but incorporate newer tools from natural language processing.

[1] For example, see the Patent and Trademark Law Amendments Act (1980), which codifies conditions for the citation of prior art in the US.

These tools help improve matching approaches previously used in the patent literature while avoiding some common criticisms.

After reviewing relevant literature on the localisation of knowledge spillovers, I discuss my conceptual framework in Section 2, data and methods in Section 3, and results in Section 4. Finally, I conclude with a summary and potential extensions in Section 5.

1.1. Patent citations and knowledge spillovers

Jaffe, Trajtenberg and Henderson (1993) introduced many key ideas underlying the analysis of knowledge spillovers through patent data. They argue that patent citations can accurately reflect these spillovers once we exclude commercial relations between inventors or assignees. The key argument is that inventors do not include citations that do not reflect spillovers since it would unnecessarily restrict the scope of an invention. Conversely, not citing a patent reflecting a knowledge flow would be challenged by the patent examiner, an expert over the relevant technologies.

However, findings could still reflect other agglomerative forces even if we exclude commercial transactions. Firms and inventors can benefit from sharing inputs to production and improve matching in labour markets when they co-locate (Carlino and Kerr, 2015, Section 6.4). Jaffe, Trajtenberg and Henderson (1993)'s insight was to create a control group of patents that mimic the technological and temporal characteristics of the citing patents. For each cited-citing patent pair, they find a control patent that does not cite the cited patent and has the same application year and three-digit United States Patent Classification (USPC) class as the citing patent.

They find localisation at the Standard Metropolitan Statistical Area (SMSA), state, and national levels. However, these results depended on how

well the three-digit classes proxied endogenous factors. This issue led Thompson and Fox-Kean (2005a) to use the six-digit classes for matching controls to reassess these results, finding no evidence of intranational localisation. They argued that three-digit controls hid significant intra-class heterogeneity, to which Henderson, Jaffe and Trajtenberg (2005) commented that boundaries between six-digit classes were arbitrary. Thompson and Fox-Kean (2005b) wrote another reply, but the issue remained unsettled.

Murata *et al.* (2014) followed in the spirit of these earlier papers but proposed important methodological advances. Its primary contribution was to adapt the localisation test of Duranton and Overman (2005) for the context of patent citations. Jaffe, Trajtenberg and Henderson (1993) and Thompson and Fox-Kean (2005a) used a discrete localisation measure. They compared the frequency at which cited and citing patents originated in the same discrete spatial unit (i.e., SMSA, state, and country) to that of cited and control. Not only did two patents in neighbouring units have the same impact as those across the country in the final results, it also implied that the results were sensitive to the modifiable areal unit problem (Wong, 2009).

In contrast, the Duranton and Overman (2005) test, which I use and describe in more detail in Section 3, treats observations as points in continuous space. This approach addresses the aforementioned issues and incorporates a richer information set in estimated parameters. Murata *et al.* (2014) find evidence supporting localisation in 70% of all three-digit classes when using three-digit controls and in a third of all three-digit classes when using six-digit controls. Additionally, more than 10% of classes showed dispersion when using six-digit controls.

The latter point implies that results from aggregate spatial units might fail to show localisation as the opposing forces would cancel out, which explains the different results in the original papers. However, it does not address the quality of either control. To do so, Murata *et al.* (2014) conducted a sensitivity analysis which generalises the controls to include three and six-digit as limiting cases. Their simulations show that tests will correctly find localisation in most classes unless the matching procedure introduces extremely high selection bias.

1.2. Developments in patent text analysis

Although results from Murata *et al.* (2014) show that the evidence for localisation is robust, developments in patent analysis tools have created an opportunity to re-examine the matching approach. Natural language processing (NLP) has progressed incredibly in the past few years, and patents contain rich textual data in their abstracts, claims, and invention descriptions. Economists have already begun incorporating textual data sources in research, but only in a limited capacity. For a general review of these cases and an introduction to text data in economics, see Gentzkow, Kelly and Taddy (2019).

Arts, Cassiman and Gomez (2017) use the Jaccard similarity coefficient, the size of the intersection of words divided by the size of the union of words in two documents, to identify similar patents using their titles and abstracts. They found that patents matched with this method were likelier to have the same assignees and inventors, technological classification, and cite one another. The results were also validated by a panel of experts, highlighting that the index had weak matching power for patents with little text. As an application, they conduct a discrete space version of the matching approach, finding evidence for localisation.

However, Arts, Cassiman and Gomez (2017) and other uses of textual data within innovation economics (Kelly *et al.*, 2021; Kalyani *et al.*, 2025, for example) have focused on simple text-based statistics, far behind the current state-of-the-art. Since Vaswani *et al.* (2017) introduced the transformer architecture, deep learning has dominated much of the natural language domain. A leading use of the architecture has been for contextual embeddings, which are vector representations of the meaning of documents. These have since outperformed previous approaches in tasks like clustering, similarity, and information retrieval (Reimers and Gurevych, 2019), and researchers can easily access pre-trained models through libraries like Sentence Transformers in Python.

Once we encode a patent's text as an embedding, we can obtain a similarity measure between patents by calculating the vectors' angle (i.e., the cosine). More similar patents would have a smaller angle between their encoded texts (Jurafsky and H. Martin, 2025, Section 6.4), and we can explore the similarity space using nearest neighbours algorithms. Sveva Ascione and Sterzi (2024) evaluate the performance of embedding models for patent similarity and find that transformer-based models outperform static measures. They use patent interferences, the case of distinct inventors submitting nearly identical claims simultaneously, as the benchmark for these tests.

I use a state-of-the-art contextual embedding model to match control to citing patents. However, due to time constraints, I have not fine-tuned the model. Despite this limitation, the model I use performs significantly better on benchmarks than the baseline contextual embedding model used by Sveva Ascione and Sterzi (2024). Feng (2020) is the only application of contextual embeddings in a matching localisation estimator that I could find. However,

because it uses a discrete measure of space, it also suffers from spatial aggregation problems. They generally find weaker localisation evidence than Jaffe, Trajtenberg and Henderson (1993).

1.3. Alternative approaches and issues

Many authors have used alternative approaches to identify various aspects of agglomeration, including the localisation of knowledge spillovers. Buzard *et al.* (2017) find that R&D labs are highly concentrated. Buzard *et al.* (2020) match patents from these R&D clusters to show that these citations are also highly localised. They find that discrete agglomeration measures are likely to understate the degree of localisation and that knowledge spillovers operate over short distances.

The latter point matches with evidence from Arzaghi and Henderson (2008), who use granular location data of advertising agencies in Manhattan and find that spillovers decay quickly. Kerr and Kominers (2015), further discussed in Section 3, provide a theoretical model showing that these small interaction distances form larger clusters than these distances. They find that patent citation data generally supports their model's predictions. Other research supports localisation as measured from the introduction of new words to trademarks (Graevenitz, Graham and Myers, 2021), patent interferences (Ganguli, Lin and Reynolds, 2020), and patent and research citations amongst universities (Belenzon and Schankerman, 2013).

Although inventors generally add patent citations, examiners and third parties might include them. Thompson (2006) compared variation within a patent in matching rates between inventor-added and examiner-added citations, finding that inventor-added ones are more localised. However, he uses a discrete space estimator. Buzard *et al.* (2020) re-tested their hypothesis

with and without examiner-added citations, finding that the exclusion has no significant impact on the final results. Interestingly, Chen (2017) finds that examiner-added citing patents' texts are more similar to their cited patents than inventor-added.

Chen (2017) also finds that either case is more similar than between non-citing patents, but some have raised concerns about citation data in economics. Roach and Cohen (2013) compare patent citations with a survey of research reports from firm's R&D lab managers. They find that patents tend not to cite basic research important to their development, pointing towards understanding citations as capturing specific elements of knowledge spillovers. Kuhn, Younge and Marco (2020) show that the nature of patent citations has also been changing over time using measures of textual similarity. Hence, economists should consider citations alongside other measures and be careful when comparing across periods.

Researchers have also made significant methodological advances. Buzard *et al.* (2020) use coarsened exact matching (Iacus, King and Porro, 2017), which improves the balance of groups compared to my approach. This estimator and other balancing methods, such as the Covariate Balancing Propensity Score (Imai and Ratkovic, 2014), could efficiently incorporate even more information into creating balanced patent analysis samples.

Continuous space estimators have also been more prevalent in economics since Duranton and Overman (2005). Marcon and Puech (2017) have introduced a typology of these measures. Some measures, such as Lang, Marcon and Puech (2019), might provide more interpretable estimators with better properties, but economists have not generally adopted them.

2. Conceptual framework

Kerr and Kominers (2015) provide one of the few microfoundations of agglomeration patterns from individual interactions. I adapt the language of this model for the context of individual inventors but keep the notation the same. Each inventor $i \in N$ sequentially chooses their location $j(i)$ from a possible location set $\mathrm{Z} \subset \mathbb{R}^2$. Locations are drawn from a uniform distribution, with more locations than inventors. The distance between inventors i and i' is $d_{j(i),j(i')}$. The benefit of interaction at cost c for these inventors is $G\left(d_{j(i),j(i')}\right)$. We assume $G(d) > 0$ and $G'(d) < 0$.

Inventors will choose the location with the highest potential benefit, i.e., $\sum_{i' \neq i} G\left(d_j, d_{j(i')}\right)$. Then, they might choose to interact with other inventors. Inventors interact if and only if the benefits of interacting exceeds the costs; formally, $G\left(d_{j(i),j(i')}\right) \geq c$. Hence, there must be a maximum distance $\rho \equiv \max\{d : G(d) \geq c\}$ at which interactions happen.

An agglomeration cluster is a set of locations interconnected by inventor interactions. For example, if inventor A interacts with inventor B, which interacts with inventor C, these form a cluster even if A does not interact with C. The latter is implied by the distance between A and C being larger than ρ. They also define inventor agglomeration as a normalised measure of the number of agglomeration clusters for a fixed number of inventors. The fewer clusters there are, the more agglomerated inventors will be.

From these assumptions, Kerr and Kominers (2015) generate a series of predictions I describe informally. First, if ρ increases, inventors who could not previously interact will now be able to. These additional interactions might result in neighbouring clusters merging. An increase in ρ will, therefore, weakly increase agglomeration. Second, these clusters will be larger, so for

fixed N, the density of inventors in a cluster decreases. Third, larger, less dense clusters imply that the average bilateral distance between inventors increases.

These predictions help link short interaction distances to observed characteristics of agglomeration clusters. However, the key proposition from the model is that, for a fixed ρ, an increase in $|G'(d)|$ weakly increases the number of inventors concentrated at short distances. Hence, the degree to which inventors are localised at different distances will indicate how fast benefits from knowledge spillovers decay. This observation serves as a justification for continuous density estimators. We compare the density of the treated group with that of the control, in which knowledge spillovers do not exist.

This approach is the central idea of Duranton and Overman (2005)'s estimator. The original approach estimates the density of bilateral distances between firms within an industry and compares it with the distances between all possible locations, i.e., those in Z. However, since these distances are not independent, they cannot create confidence intervals of the control density analytically. Hence, they use simulations to generate these intervals. In Murata *et al.* (2014)'s adaptation, we compare the distances between cited and citing patents with that of cited and control. We randomly draw controls for each cited-citing pair to estimate intervals.

Kerr and Kominers (2015)'s model also might help explain why some evidence of inventor-added citations being more localised than examiner-added exists, even though both are localised. Inventor-added could better reflect direct interactions between inventors, which operate under maximum interaction distance ρ. However, examiner-added directly influences these citing patents through a sequence of individual connections. Since these connections,

by definition, form an agglomeration cluster, either citation source indicates localisation.

3. Data and methods[2]

3.1. Sample construction

I used data from the USPTO's PatentsView platform, which helps disseminate intellectual property data. It contains quarterly updated datasets created from raw patent information. The datasets undergo an entity (e.g., inventors, assignees, and locations) disambiguation process. Locations are then standardised with place names from the OSMNames project[3], which further simplifies the pre-processing required for the density estimator. Finally, the dataset contains information on five different patent classification systems, including the USPC and its successor, the Cooperative Patent Classification (CPC).

The platform serves data through an API and bulk downloads files. I used the API to fetch all utility patents granted between 2005 and 2025 with at least one US-based inventor and one US-based corporate assignee. I then excluded all patents with multiple assignees, any missing information, or citations not added by inventors or examiners. Finally, I removed those patents with abstracts in the bottom and top 0.01 per cent of the character count. These conditions were satisfied by around 2 million unique patents and 11,500 unique inventors. I collected citations, locations, and CPC class information separately from the bulk download files.

I chose utility patents and corporate assignees to match the sample construction choices of Jaffe, Trajtenberg and Henderson (1993) and Thompson

[2] All code is available at https://github.com/gabriellob0/lse-diss.

[3] The OSMNames database contains place names from OpenStreetMap and geographic information. It is available at https://osmnames.org/.

and Fox-Kean (2005a). However, these papers and Murata *et al.* (2014) included patents granted between the mid-70s and 1990 or 2000, so my sample covers a different period. This fact would likely imply weaker localisation effects if we consider the increased importance of long-distance collaboration. For example, Kerr and Kerr (2018) highlight the increase in cross-country inventor teams.

Another difference is that Jaffe, Trajtenberg and Henderson (1993) exclusively analysed assignees which are top corporations or universities, but they did not find any differences between groups. Hence, I only restricted the number of assignees. I did so to simplify my code and since multiple assignees could reflect complex commercial arrangements with inventors. For example, as noted by Jaffe, Trajtenberg and Henderson (1993), an invention might be assigneed to the US and a foreign office of a same company, but the internalised flow of knowledge within the broader organisation might not reflect this legal distinction. This may or may not be reflected in the choice of assignee for other patents, introducing a potential source of bias.

Missing observations are likely missing randomly and amount to less than a thousand observations. I observed that summary tables describing missingness varied between requests for the same data, likely indicating API issues. The applicant, the examiner, and third parties can include citations. I only kept the first two since the third, which comprises a small group, might not reflect spillovers. Although we could use the same argument to favour removing examiner-added citations, previous research has not used this restriction. An interesting extension would have been to examine both samples, but given time restrictions, I opted for the larger one.

Following the results of Arts, Cassiman and Gomez (2017), I restricted the abstract size to improve the matching quality. Short abstracts, the shortest

being three characters, likely do not contain enough information to generate an appropriate embedding. Long abstracts would either require trimming prior to encoding or a model with a longer sequence length (i.e., the length of the text it can encode). Models that satisfy the latter condition require more compute, but given that the right tail of character sizes is thin, the impact of including them is likely nil.

3.2. Matching strategy

The general matching procedure is as follows. First, we select a period from which we define a set of cited patents. We then find all patents that cite any of the patents in the originating set. The cited-citing pairs then correspond to our treatment group. For each cited-citing pair, we define a set of multiple control patents that do not cite the cited patent of the pair but have characteristics similar to those of the citing. The cited-control pairs correspond to the control group.

I intended to define the originating set as patents granted between 2005 and 2010 and consider all citations in my raw data. However, the matching process became too computationally expensive, so I had to create a reduced set to satisfy my time constraints. I defined all patents granted in the first month of 2005 with at least one citation in my data over the next five years as belonging to the cited set. I then joined each citing patent in these five years to their respective citation. As in previous papers, I exclude any cited-citing pairs with the same assignee or any inventor in common. These citations likely reflect commercial arrangements between assignees and inventors or continued work, so they are not true externalities.

To select the control patents, I proceeded in two steps. First, for each cited-citing pair, I select all patents granted between 2005 and 2010 with an

application date within 180 days of the citing patent's application date, removing those with the same assignee or inventors as their respective cited patent. Then, I encoded all unique citing and potential control patents as embeddings, and for each citing patent, I found the patents in the 99.9th similarity quantile. The intersection between patents within the date range and the similarity quantile corresponds to the set of admissible controls. Table 1 shows the count of unique patents before and after the embeddings — only a single cited patent had no citing patent with an appropriate control patent.

Table 1

Sample sizes before and after similarity matching

		Unique patents		
	Observations	Cited	Citing	Control
Pre-similarity matching	214,886,441	1,199	2,758	306,966
Post-similarity matching	147,088	1,198	2,754	70,231

Notes: The pre-similarity matching data includes controls with an application date within 180 days of the citing patent's application date that do not cite the cited patent or have any inventors or assignees in common. The post-similarity matching data consists of the intersection of the first group with the 308 nearest neighbours over all citing and control patents.

To encode the patent abstracts, I used the "nomic-embed-text-v2-moe" embedding model (Nussbaum and Duderstadt, 2025). This model has open-sourced data, weights (analogous to regression coefficients), and code and is available on HuggingFace. Then, I queried the nearest neighbours, as determined by the cosine of the embeddings, with an approximate nearest neighbour algorithm. The similarity quantile corresponded to the 308 nearest neighbours.

I also matched the final set of cited patents to their respective CPC sections. There are nine sections, A to H and Y, and patents might have more than one section. Table 2 lists the number of patents in each section alongside their description. Although the number of cited patents is proportional to the

number of cite-citing-control triples, it is imperfect. This fact might indicate class heterogeneity in citations and matching rates. Coupled with the points I make in Section 1.3, we must be careful when interpreting differences in localisation across classes.

Table 2

Counts of cited patents, citing pairs, and patent triples by CPC section

CPC section	Description	Counts		
		Cited patents	Citing pairs	Patent triples
A	Human necessities	184	501	29,284
B	Performing operations; transporting	220	490	25,055
C	Chemistry; metallurgy	139	357	17,491
D	Textiles; paper	10	14	501
E	Fixed constructions	33	93	6,495
F	Mechanical engineering; lighting; heating; weapons; blasting engines or pumps	76	144	7,879
G	Physics	492	1,559	76,343
H	Electricity	448	1,385	68,480
Y	General tagging of new technological developments; general tagging of cross-sectional technologies spanning over several sections of the IPC; technical subjects covered by former USPC cross-reference art collections [XRACs] and digests	286	727	35,987

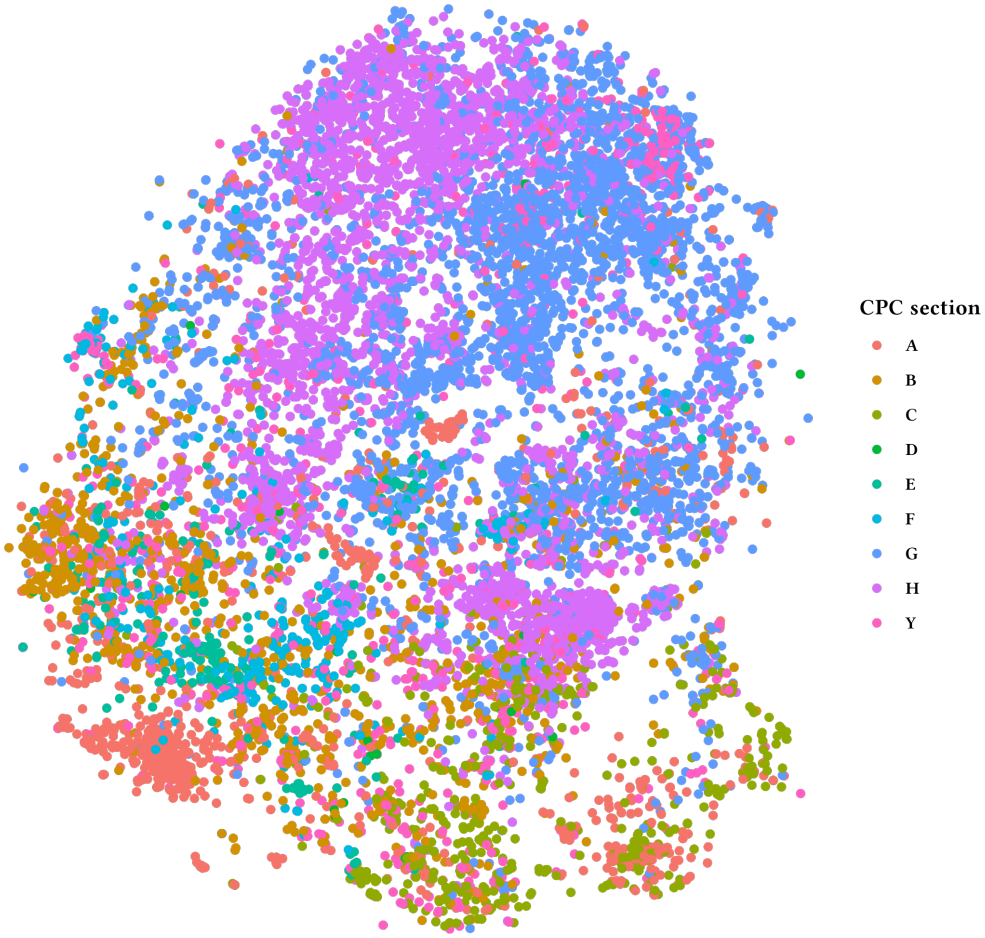
Notes: Cited patent counts refer to the count of unique cited patents in the matched sample. A citing pair is a unique cited-citing patent combination. A triple is a cited, citing, and control patent. Hence, multiple triples exist for each cited patent and each cited-citing pair. The definitions of the CPC section are from the European Patent Office website.

Figure 1 shows a random sample of the 256-dimensional embeddings represented in 2 dimensions using the t-SNE dimensionality reduction algorithm. Proximity denotes semantic similarity. Colours denote the patent class, and we see evidence of clustering based on these colours. These patterns highlight

how the text and the classes of the patents are connected, which is captured by the embeddings.

Figure 1

Patent embeddings and CPC classes



Notes: The figure shows a two-dimensional visualisation of 50,000 citing and control patent embeddings from a random sample. I keep only a unique CPC section selected at random per patent. I reduce dimensions using the t-SNE algorithm with a perplexity of 30 and theta of 0.5. Closer observations have similar meanings.

As an illustrative example, the first patent within the set of embedded patents is patent number 6844531, titled "Conduit ready electric belly-band heater and method of use." Its nearest neighbour, regardless of whether it is an admissible control, is patent 7372007, which describes a "Medium voltage heater element."

Both share the same CPC subclass and only differ in the group, which is the most granular classification. Also, while the former has a single classification, patent 7372007 has four classifications. This multiplicity has generated ambiguity in the literature, which the embedding approach avoids.

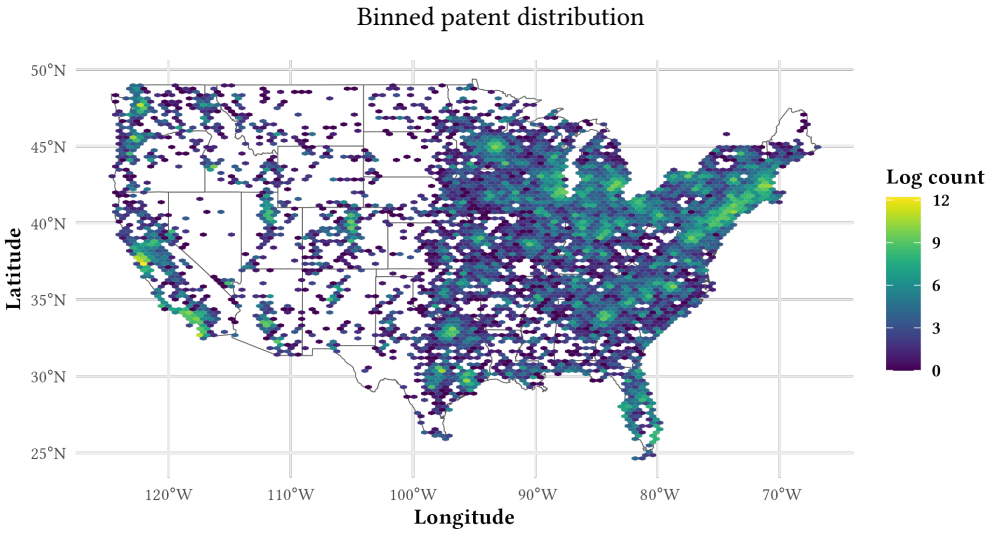
3.3. Localisation test

For the localisation test, I calculated the geodesic distance between all cited-citing and cited-control pairs using their latitudes and longitudes. However, patents do not have a location, so most previous research constructed one from inventor locations. Since patents have multiple inventors who may work in different parts of the country, I followed Kerr and Kominers (2015)'s procedure to construct them.

First, I selected areas where the most inventors live for each patent. If this location was not unique, I selected the one corresponding to the lowest inventor sequence within the restricted set. The inventor sequence variable lists the order in which the patent lists its inventors. Generally, the first inventor has contributed the most to the invention, and the order between the others matters less. Therefore, the second step should capture the location that contributed the most or at least be random.

Figure 2 shows the matched sample's binned log count of all cited, citing, and control patents. As expected, the unconditional spatial distribution of patents shows geographic concentration. A clear example is Silicon Valley in California. Even without knowledge spillover, other agglomerative forces affect inventors and assignees. As highlighted by Ellison and Glaeser (1997), natural advantages and pure change could alone drive these patterns in industrial localisation. This plot helps illustrate the need for a matching estimator.

Figure 2



Notes: The figure shows the spatial distribution of all patents in the matched sample in 100 hexagonal bins. The colours are in log scale. I determine patent location first by selecting the most common inventor locations and then by selecting the location with the lowest inventor sequence. I only consider locations in the US.

For each random draw, I estimated the density $\hat{K}(d)$, where d is the distance, with a Gaussian kernel. I selected bandwidths using the Sheather and Jones (1991) method. Formally, for each distance d the, the density $\hat{K}(d)$ is

$$\hat{K}(d) = \frac{1}{2hN} \sum_{i=1}^{n} \sum_{j=1}^{n^i} f\left(\frac{d - d_{i,j}}{h}\right),$$

where, for each subsample, n is the number of cited patents, n^i the number of citing patents for cited patent i, and N is the sum of n^i for all $i \in \{1, ..., n\}$. Additionally, h is the bandwidth, $f()$ kernel, and $d_{i,j}$ the distances between patents i and j.

Since the density should be zero for negative distances, we use the reflection method employed by Duranton and Overman (2005). I estimated the density, denoted $\tilde{K}(d)$, over an augmented set that included the original distances and minus the original distances. Then, I set $\hat{K}(d) = 0, \ \forall d < 0$, and $\hat{K}(d) = 2\tilde{K}(d), \ \forall d \geq 0$.

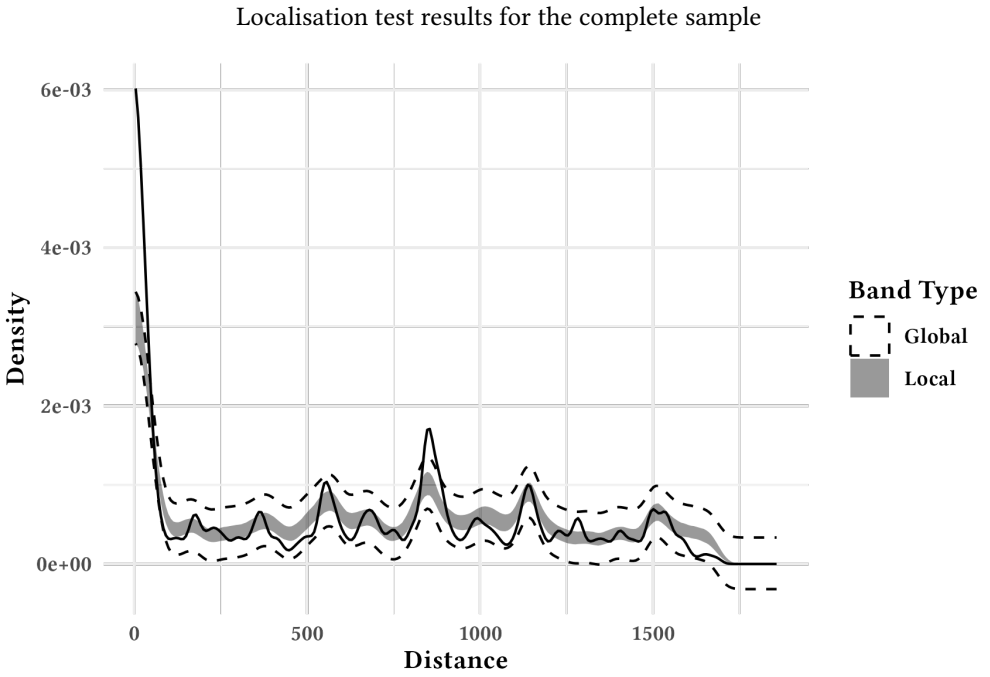
I create the local confidence intervals by selecting the 95th and 5th quantiles over the densities at each distance from the 1,000 random control draws. These quantiles are the lower confidence interval $\underline{K}(d)$ and the upper confidence interval $\overline{K}(d)$. If $\hat{K}(d) \geq \overline{K}(d)$, we say this subsample shows localisation at distance d. Conversely, if $\hat{K}(d) \leq \underline{K}(d)$, we say this subsample shows dispersion at distance d.

The local tests have two issues. First, dispersion and location at a distance generally imply the converse at a separate distance. Second, we are conducting multiple hypothesis tests by considering all distances. Hence, we construct global confidence intervals so that only 5% of simulated densities will cross these intervals, with the probability constant across all distances. We denote these bands $\overline{\overline{K}}(d)$ and $\underline{\underline{K}}(d)$. A subsample displays global localisation if $\hat{K}(d) \geq \overline{\overline{K}}(d)$ for at least one d and global dispersion if $\hat{K}(d) \leq \underline{\underline{K}}(d)$ for at least one d.

4. Results

Figure 3 shows the results of the localisation test for the full sample. The line in the middle is the estimated density, with the shaded band representing the local confidence intervals and the dashed lines the global confidence intervals. The most striking feature is that there is a significant spike at zero. This spike is higher than the local and global bands, which indicates localisation in the patent citations. We also see local dispersion at specific distances but no evidence of global dispersion.

Figure 3



Notes: I estimated densities with a Gaussian kernel and selected bandwidths with Sheather and Jones (1991) method. The shaded area represents local confidence intervals, and the dashed line represents global confidence intervals. I estimate these intervals through a Monte Carlo simulation with 1,000 draws.

The spike is consistent with previous evidence indicating that benefits from knowledge spillovers decay very quickly. The data itself had many patents at the same OSMNames place, the lowest level of aggregation, corresponding to a distance of zero. This pattern indicates that a potential avenue for research would be to study these spillovers with extremely disaggregated data at short distances, offering more precise insights into how they operate.

Table 3 shows the global test results for the complete sample and each CPC section. All CPC sections show localisation results, i.e., the density crosses the global bands at least once. Some classes also show evidence of dispersion. However, if localisation is very strong, it can imply dispersion since the density must sum to one. Hence, Duranton and Overman (2005) only consider industries with dispersion and no localisation as truly dispersed.

Table 3

Global localisation test results

CPC section	Localisation	Dispersion
All	Yes	No
A	Yes	No
B	Yes	Yes
C	Yes	Yes
E	Yes	No
F	Yes	Yes
G	Yes	Yes
H	Yes	Yes
Y	Yes	No

Notes: I define a patent class as localised when the density crosses the upper band of the global confidence intervals. Conversely, a class is dispersed when it crosses the lower bands.

Figure 4 shows the localisation test densities for each CPC section. The patterns are similar to those of the full sample. Generally, there is a spike around zero in the density estimates. There also tends to be localisation between 500km and 1000km for most classes. This distance could correspond to the modal bilateral distance of clusters in the Kerr and Kominers (2015)'s model, excluding direct interactions.

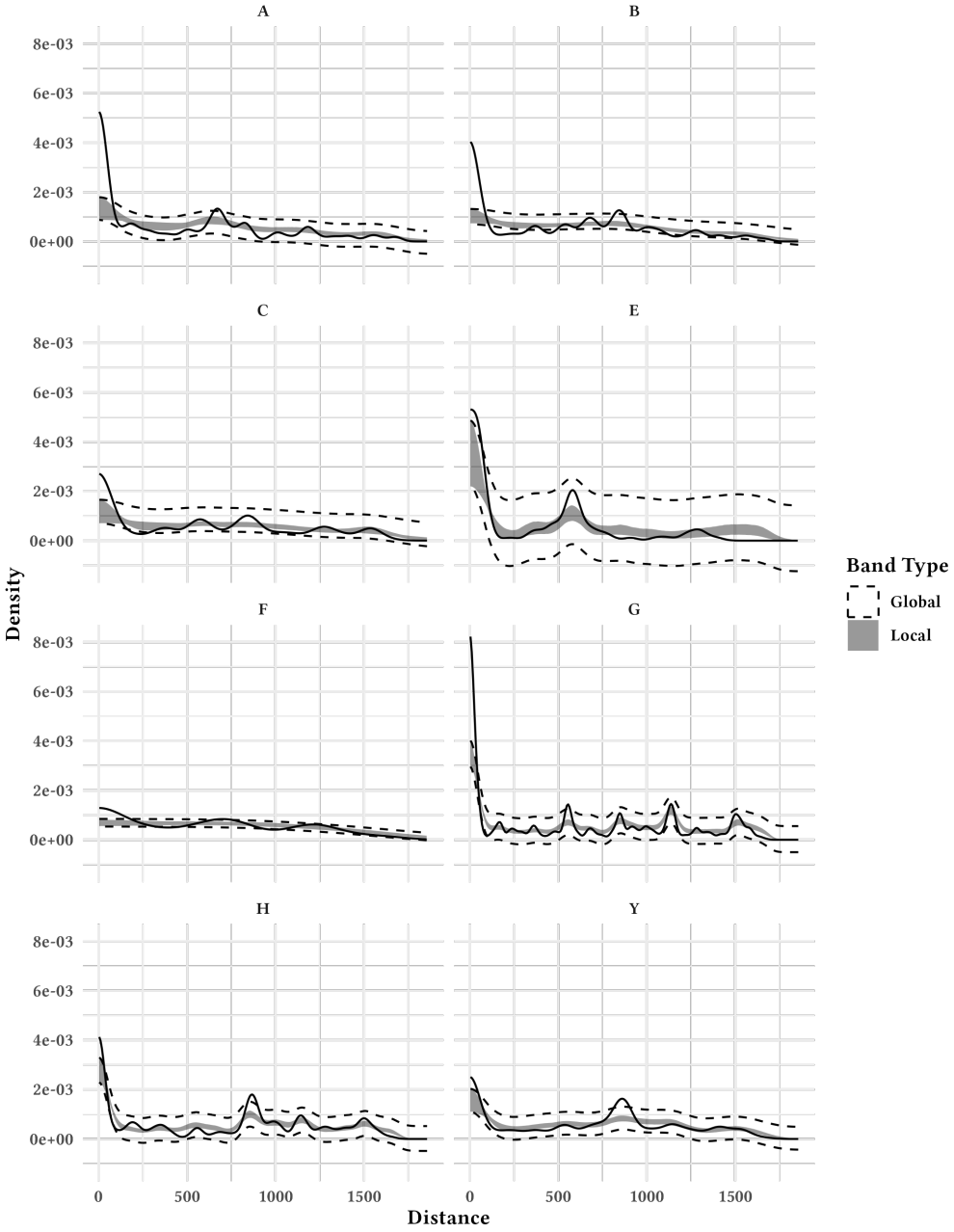
For this latter point, we expect a smoother distribution within the agglomeration cluster area followed by significant dispersion. However, this unconditional distribution pattern (i.e., a separate spike) might reflect the underlying characteristics of inventor location choices—for example, the distance between two very important cities within clusters. The matching approach would control for these patterns.

One interesting point is that previous papers used Silverman's rule of thumb to select bandwidths. It requires stronger assumptions about the data, but it is easier to compute. However, when I tested these bandwidths, the data was over-smoothed compared to when using Sheather and Jones (1991). This smoothing hid the second peak between 500km and 1000km, which provides an interesting link to the conceptual framework I used.

One important distinction from Murata *et al.* (2014) approach is that they replace zero values (i.e., inventors that belong to the same spatial unit) with the expected distance between two random points within that unit's borders. This method reduces bias introduced when we compare large places to smaller ones. However, this data is unavailable from PatentsView, so I could not reproduce the approach. If the area size is proxied by the control group, the bias shifts the distribution but does not affect the localisation results. Otherwise, the direction of the bias on the localisation results is ambiguous.

Figure 4

Localisation test results for CPC classes



Notes: I estimated densities with a Gaussian kernel and selected bandwidths with Sheather and Jones (1991) method. The shaded area represents local confidence intervals, and the dashed line represents global confidence intervals. I estimate these intervals through a Monte Carlo simulation with 1,000 draws. CPC section D was excluded from the data since it had too few observations.

5. Conclusion

These results show evidence of global localisation in knowledge spillovers. The observed patterns are consistent with previous research in the area. In particular, the benefits from knowledge spillovers decay quickly. My main contribution was to use an alternative matching procedure for the empirical strategy first developed by Jaffe, Trajtenberg and Henderson (1993) with less biased localisation tests. Other papers using text analysis in this context only treated the localisation tests as a secondary concern, so they focused on discrete measures. Murata *et al.* (2014) created the adaptation of the Duranton and Overman (2005) test which I use here.

Several techniques could have further improved performance within the patent similarity approach. The leading method would be fine-tuning pre-trained models on patent texts or using domain-specific models. Although these incur additional computing and time requirements, Sveva Ascione and Sterzi (2024) results show that this could create significant improvements. I could also have applied these models with other pieces of text in the patent, particularly their claims. However, this data is not yet fully available at PatentsView.

The latter is particularly interesting for future research since we could create even higher-dimensional datasets with patent information that can be easily analysed. Not only can the full patent text be encoded, but multimodal embedding models also work with images of inventions. This additional data might be better incorporated with newer matching approaches, which can create balanced samples in a data-driven way.

Finally, other dimensions of knowledge spillovers can still be explored within this framework. In particular, with more computation power, we could

use larger samples to study more detailed classes and variations in the level of localisation across time. The nature of the process underlying inventions is changing, and this data can help uncover these changes. The text approach is particularly important in this context since it can easily be matched with external data, such as from basic research or final outputs of the invention process.

However, all these analyses can only characterise specific aspects of knowledge spillovers. Citations only capture spillovers within the patenting element of innovation and, even so, exclude important sources of knowledge such as basic research. Natural language processing tools can be useful for diversifying our information set to characterise knowledge spillovers. With tools like embeddings, we might easily quantify text data that was previously unexplored. Research in this direction can help create a richer interpretation of the mechanisms behind knowledge transmission.

Bibliography

Arts, S., Cassiman, B. and Gomez, J.C. (2017) "Text matching to measure patent similarity," *Strategic Management Journal*, 39(1), pp. 62–84. Available at: https://doi.org/10.1002/smj.2699.

Arzaghi, M. and Henderson, J.V. (2008) "Networking off Madison Avenue," *Review of Economic Studies*, 75(4), pp. 1011–1038. Available at: https://doi.org/10.1111/j.1467-937X.2008.00499.x.

Belenzon, S. and Schankerman, M. (2013) "Spreading the Word: Geography, Policy, and Knowledge Spillovers," *Review of Economics and Statistics*, 95(3), pp. 884–903. Available at: https://doi.org/10.1162/REST_a_00334.

Buzard, K. *et al.* (2017) "The agglomeration of American R&D labs," *Journal of Urban Economics*, 101, pp. 14–26. Available at: https://doi.org/10.1016/j.jue.2017.05.007.

Buzard, K. *et al.* (2020) "Localized knowledge spillovers: Evidence from the spatial clustering of R&D labs and patent citations," *Regional Science and Urban Economics*, 81. Available at: https://doi.org/10.1016/j.regsciurbeco.2019.103490.

Carlino, G. and Kerr, W.R. (2015) "Chapter 6 - Agglomeration and Innovation," in G. Duranton, J.V. Henderson, and W.C. Strange (eds.) *Handbook of Regional and Urban Economics*. Elsevier, pp. 349–404. Available at: https://doi.org/https://doi.org/10.1016/B978-0-444-59517-1.00006-4.

Chen, L. (2017) "Do patent citations indicate knowledge linkage? The evidence from text similarities between patents and their citations," *Journal of Informetrics*, 11(1), pp. 63–79. Available at: https://doi.org/10.1016/j.joi.2016.04.018.

Duranton, G. and Overman, H.G. (2005) "Testing for localization using micro-geographic data," *Review of Economic Studies*, 72(4), pp. 1077–1106. Available at: https://doi.org/Doi 10.1111/0034-6527.00362.

Ellison, G. and Glaeser, E.L. (1997) "Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach," *Journal of Political Economy*, 105(5), pp. 889–927. Available at: https://doi.org/10.1086/262098.

Feng, S. (2020) "The proximity of ideas: An analysis of patent text using machine learning," *PLoS One*, 15(7), p. e234880. Available at: https://doi.org/10.1371/journal.pone.0234880.

Ganguli, I., Lin, J. and Reynolds, N. (2020) "The Paper Trail of Knowledge Spillovers: Evidence from Patent Interferences," *American Economic Journal: Applied Economics*, 12(2), pp. 278–302. Available at: https://doi.org/10.1257/app.20180017.

Gentzkow, M., Kelly, B. and Taddy, M. (2019) "Text as Data," *Journal of Economic Literature*, 57(3), pp. 535–574. Available at: https://doi.org/10.1257/jel.20181020.

Graevenitz, G. von, Graham, S.J.H. and Myers, A.F. (2021) "Distance (still) hampers diffusion of innovations," *Regional Studies*, 56(2), pp. 227–241. Available at: https://doi.org/10.1080/00343404.2021.1918334.

Henderson, R., Jaffe, A. and Trajtenberg, M. (2005) "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment: Comment," *American Economic Review*, 95(1), pp. 461–464. Available at: https://doi.org/10.1257/0002828053828644.

Iacus, S.M., King, G. and Porro, G. (2017) "Causal Inference without Balance Checking: Coarsened Exact Matching," *Political Analysis*, 20(1), pp. 1–24. Available at: https://doi.org/10.1093/pan/mpr013.

Imai, K. and Ratkovic, M. (2014) "Covariate Balancing Propensity Score," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1), pp. 243–263. Available at: https://doi.org/10.1111/rssb.12027.

Jaffe, A.B., Trajtenberg, M. and Henderson, R. (1993) "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *The Quarterly Journal of Economics*, 108(3), pp. 577–598. Available at: https://doi.org/10.2307/2118401.

Jurafsky, D. and H. Martin, J. (2025) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd ed. Available at: https://web.stanford.edu/~jurafsky/slp3/.

Kalyani, A. *et al.* (2025) "The Diffusion of New Technologies," *The Quarterly Journal of Economics*, 140(2), pp. 1299–1365. Available at: https://doi.org/10.1093/qje/qjaf002.

Kelly, B. *et al.* (2021) "Measuring Technological Innovation over the Long Run," *American Economic Review: Insights*, 3(3), pp. 303–320. Available at: https://doi.org/10.1257/aeri.20190499.

Kerr, S.P. and Kerr, W.R. (2018) "Global Collaborative Patents," *The Economic Journal*, 128(612), pp. F235–F272. Available at: https://doi.org/10.1111/ecoj.12369.

Kerr, W.R. and Kominers, S.D. (2015) "Agglomerative Forces and Cluster Shapes," *Review of Economics and Statistics*, 97(4), pp. 877–899. Available at: https://doi.org/10.1162/REST_a_00471.

Kuhn, J., Younge, K. and Marco, A. (2020) "Patent citations reexamined," *The RAND Journal of Economics*, 51(1), pp. 109–132. Available at: https://doi.org/10.1111/1756-2171.12307.

Lang, G., Marcon, E. and Puech, F. (2019) "Distance-based measures of spatial concentration: introducing a relative density function," *The Annals of Regional Science*, 64(2), pp. 243–265. Available at: https://doi.org/10.1007/s00168-019-00946-7.

Marcon, E. and Puech, F. (2017) "A typology of distance-based measures of spatial concentration," *Regional Science and Urban Economics*, 62, pp. 56–67. Available at: https://doi.org/10.1016/j.regsciurbeco.2016.10.004.

Murata, Y. *et al.* (2014) "Localized Knowledge Spillovers and Patent Citations: A Distance-Based Approach," *Review of Economics and Statistics*, 96(5), pp. 967–985. Available at: https://doi.org/10.1162/REST_a_00422.

Nussbaum, Z. and Duderstadt, B. (2025) *Training Sparse Mixture Of Experts Text Embedding Models*. Available at: https://doi.org/10.48550/arXiv.2502.07972.

Reimers, N. and Gurevych, I. (2019) "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (Emnlp-Ijcnlp 2019)*, pp. 3982–3992.

Roach, M. and Cohen, W.M. (2013) "Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research," *Manage Sci*, 59(2), pp. 504–525. Available at: https://doi.org/10.1287/mnsc.1120.1644.

Sheather, S.J. and Jones, M.C. (1991) "A Reliable Data-Based Bandwidth Selection Method for Kernel Density-Estimation," *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 53(3), pp. 683–690.

Sveva Ascione, G. and Sterzi, V. (2024) *A comparative analysis of embedding models for patent similarity*. Available at: https://doi.org/10.48550/arXiv.2403.16630.

Thompson, P. (2006) "Patent Citations and the Geography of Knowledge Spillovers: Evidence from Inventor- and Examiner-added Citations," *Review of Economics and Statistics*, 88(2), pp. 383–388. Available at: https://doi.org/10.1162/rest.88.2.383.

Thompson, P. and Fox-Kean, M. (2005a) "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment," *American Economic Review*, 95(1), pp. 450–460. Available at: https://doi.org/10.1257/0002828053828509.

Thompson, P. and Fox-Kean, M. (2005b) "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment: Reply," *American Economic Review*, 95(1), pp. 465–466. Available at: https://doi.org/10.1257/0002828053828617.

Vaswani, A. *et al.* (2017) "Attention Is All You Need," *Advances in Neural Information Processing Systems 30 (Nips 2017)*, 30.

Wong, D.W. (2009) "Modifiable Areal Unit Problem," in R. Kitchin and N. Thrift (eds.) *International Encyclopedia of Human Geography*. Oxford: Elsevier, pp. 169–174. Available at: https://doi.org/https://doi.org/10.1016/B978-008044910-4.00475-2.

"Patent and Trademark Law Amendments Act" (1980).