WORKING PAPER

A method, data, and quality note introducing the Dynamic Global Microgeographic Patent Database

Nicholas P. Sweeney[a]

April 17, 2026

Abstract

The spatial concentration of innovation is increasingly understood as a driver of economic growth, yet its microgeography remains poorly understood. Inventor address-level patent records are valuable for studying the geography of innovation, but they remain underused because raw records require substantial retrieval, cleaning, harmonisation, deduplication, and geocoding. Utilising improvements in programmatic bulk retrieval, the Dynamic Global Microgeographic Patent Database (DGMPD) addresses this gap. As a 'living' dataset, continuously updated from patents available through trusted international and domestic patent office APIs, the DGMPD provides *daily, technology-specific (IPC/CPC) data on the microgeography of innovation*. Given this unique position, my note is primarily methods focused, explaining the creation of the DGMPD. I nevertheless evidence preliminary external validation tests and results for a fixed subsample of the DGMPD. This includes 281,115 deduplicated and cleaned raw patents priority filed between 01-01-2022 to 31-12-2022, reflecting patenting activity from 117 countries by 630,299 inventors postcode-matched with a c.86% success rate to 199,201 unique matched addresses (anonymised to 40,143 unique 1km^2 grids). Back- and forward-filling of the DGMPD is ongoing for 1980 to April/May 2026. The DGMPD demonstrates how new data infrastructures can refine understanding of evolving technological and spatial patterns of innovation.*

[a]Sir Terry Leahy Fellow, Alliance Manchester Business School

*This is a first pass working paper intended for feedback on method and outputs/analysis that would be useful for researchers interested in the DGMPD. The objective of this database is that it is a trusted source which makes studying the microgeography of innovation far more accessible than at present. The design and ease of export of data for researchers is included with that. Please email `nicholas.sweeney@manchester.ac.uk`

Contents

1. Introduction

There is growing interest in the geography of innovation because many of the most important contemporary research and policy questions are now explicitly spatial.[1–3] This reflects two linked developments. First, innovation has become increasingly central to economic development and growth,[4;5] as well as wider societal challenges such as climate change.[6–9] Second, spatial inequalities within and across countries remain pronounced, with opportunities and vulnerabilities distributed unevenly across places.[10;11]

Across disciplines, inequalities are increasingly understood as multiscalar rather than confined to a single territorial level.[12;13] Neighbourhood trajectories are shaped by regional contexts, regional outcomes are embedded in national systems, and national patterns are themselves structured through uneven local concentrations of activity.[14;15] This is reinforced by the fact that high-technology spillovers are often highly clustered and operate at fine spatial scales.[16;17] In this setting, the relevant unit of analysis is often not only a country or region, but also an inventor location, applicant address, postcode, city, or metropolitan corridor.[18;19]

Address-level patent records are therefore an increasingly valuable source for studying how inventive activity is distributed across space.[20;21] As legal and administrative documents, they contain rich information on inventors, applicants, technologies, citations, and family links.[22;23] The challenge lies not in the absence of analytical value, but in organising that information into forms that can support rigorous spatial analysis at scale.[24] Fragmentation across jurisdictions, variation in disclosure practices, and heterogeneous document formats mean that inventor and applicant addresses are often incomplete, inconsistently formatted, duplicated, or difficult to harmonise.[25;26] As a result, a substantial share of the work required for patent geography still lies in retrieval, cleaning, standardisation, deduplication, and geocoding before substantive analysis can begin.

This challenge has long shaped the empirical frontier of patent geography. OECD REGPAT considerably advanced the field by making patent data more usable for regional analysis and comparative research.[24] This is useful for understanding within country distributions, however suffers from concerns over potentially misguided (policy) interpretations given the modifiable areal unit problem[27†] given that innovative activity may not be evenly distributed within regions, often being urban centric.

More recent contributions have extended this frontier further in terms of global scope and spatially finer units of analysis. Morrison et al.[21] uses microgeolocated patent data to disambiguate name inconsistencies in individual patent records in 8.5 million patents. De Rassenfosse et al.[26] demonstrated the value of geocoding worldwide first filings and recovering missing locations through patent-family information, while Bergeaud and Verluise[29] showed how large-scale extraction from historical patent documents can recover long-run urban innovation patterns from otherwise difficult archival materials. However, these datasets are characterised by one off

† (The modifiable areal unit problem specifically concerns spatial aggregation error. See ecological fallacy for a more generalised discussion in other literatures[28].

or periodic (annual) releases.[30] This, of course, is not how innovation occurs.[25‡]

The rise of big data and improved computational power brings with it significant opportunities, but also the drawbacks must not be underestimated. For instance, a growing body of research relies on OpenAlex citations[4§] to measure knowledge creation. This is encouraging, and must be seen as complimentary to measuring innovation through patenting however, the accuracy and representativeness[32] of OpenAlex, as an emerging technology, is moot. Patent registries, on the other hand, represent the universe of legal protection of new ideas worth.

This note builds on these insights by documenting the *Dynamic Global Microgeographic Patent Database* (DGMPD). The DGMPD's objective is to be a transparent, research-grade spatial data infrastructure for global patent (inventor or applicant) geography. For this illustrative working paper, I focus on 2022, which is the most recent full year in which we can reliably analyse patent data due to differences between patenting authority approval times.[20¶]

Drawing on multiple official patent registry APIs and related document sources,[33–35] the DGMPD brings together advances in data access, OCR-assisted document processing,[36;37] geocoding,[38] and reproducible workflow design within a continuously updatable pipeline. In benchmarking against the three most recognised current datasets with global span,[24;26;29] it presents advancements in the conversion of messy, fragmented patent records into usable (anonymised) microspatial evidence while preserving provenance, making uncertainty visible, and supporting extension as source access, enrichment, and matching improve.

This technology (IPC/CPC) specific micro-geocoded database would be useful for work on uneven innovation and development,[5] green technological change,[39] complexity[40–43] and relatedness.[44–46]

The remainder of the note describes the source material and pipeline architecture, outlines the stages of retrieval, deduplication, address extraction, and geocoding, and presents evidence on the improvements in spatial precision, validation, and stage-specific coverage compared to earlier approaches.

2. Current Geography of Patenting Approaches

2.1. Patent data and innovation

Patenting data is an imperfect proxy for innovation.[47;48] This long recognised limitation means that any discussion of the data measuring patent activity is necessarily limited by this partial account of innovation.[49;50] Moreover, even taking a narrow, technological-digital driven view of innovation, more recently this disparity between patenting and innovation has arguably been

‡It is important that this is not understood as a criticism of existing approaches, where the constraints on retrieval and processing were greater. For instance previous researcher-led efforts on geocoding innovation document that while they were able to use an API to analyse the Korean Intellectual Property Office,[26] they were required to write WIPO and other countries' domestic offices, which is a far more cumbersome. In contrast, as this methods approach is underpinned by programatic tools to bulk-process raw data from official sources. Where De Rassenfosse et al.[26] had to require written permission and release from officials in German patenting, the process is simple as of 17 April 2026: German patents are available through an API on the DPMA[31]. After a routine licence fee is paid. So, rather than a criticism, this method identifies a significant opportunity for well managed and organised data collection based on previous cutting edge research to improve understanding on innovation.

§OpenAlex is an open access library carrying the metadata of '474 million scholarly works'.

¶However, these data run daily until (nearly)-present. A separate avenue of research that these data present is therefore the potential now/forecasting based *bias-adjusted present estimates* on this highly granular, historically rich, data.

exacerbated through the rise in open source coding improvements.[51] For a more extensive recent assessment of the drawbacks of patenting as a proxy for innovation in the literature, see Spreafico et al.[52], yet it is broadly recognised that patenting remains an important, quantifiable, measure of innovative activity.[53]

In any case, the focus of this methods piece is not to discuss the merits of this, but instead to improve the internal validity of the spatial analysis of patent modelling. This note therefore documents the construction of DGMPD from official patent data sources. The database currently draws on programmatic and document-level sources including OPS,[33] PATENTSCOPE,[35] and USPTO materials.[34] These sources differ in cost, rate limits, document structure, and territorial scope. Where sources are freely available, this is documented, and if individuals are comfortable using APIs this could be good given case specific measures.

2.2. Limitations of current patent data sources

Geographical unit The geography of patenting at national scales is well understood, with coverage being comprehensive at that level, meaning that there are a wealth of datasets that currently document the geography of patenting at this aggregation: see PatSeq for a visual summary.[54] It is the sub-national level where these addresses become patchier.

For example, the 2025 release of PATSTAT[55] p.2 notes 'NUTS codes of *some* person records in table `TLS206_PERSON` are enhanced by values from the OECD REGPAT database' (emphasis added). Even where NUTS information is available, the finest NUTS 3 region aggregation delineates populations that range from 150,000-800,000.[56] Given the concentrated microfoundations of innovative activity,[57] these aggregate spatial scales represent somewhat of a limitation.

Alternative sources such as Orbis IP allow for geolocation if they are linked to domestic business registers, for instance through a Companies House number in the UK.[58] However, not is the possibility to do this restricted across countries, they only reveal the administrative addresses of businesses, as opposed to the inventor address. This is problematic because of the well documented 'head quarter bias', i.e. that a (larger) company may centralise their administrative procedure, yhus concentrating applications in one area, not truly reflecting hte geography of innovative activity (inventor address).[59;60]

Privacy There are, indeed, privacy concerns over address level data, especially in the case of inventors and current legal approaches vary across offices. For instance, since 2008, World Intellectual Property Organization[61] has not 'displayed for individual applicants and inventors on the Bibliographic Data tab in the PATENTSCOPE search service', whereas this is available in biliographic XML from EPO, under conditions (specifically Article 3.14) that preclude the use of the information 'for advertising, promotional or spamming purposes'[62]. In the case of World Intellectual Property Organization[61] they note that addresses 'continue to be accessible on the front page of published PCT international patent applications and other documents available through the Documents tab in the PATENTSCOPE search service'.

Given the public availability of this data, and the proliferation of APIs allowing more programatic analysis,[63] there is a large gap in current understanding of the spatial concentration of innovation activity if analysis remains predominantly at the regional level of administrative addresses. Nevertheless, these discordant approaches and that micro-level analysis is an evolving terrain, I take a cautious approach, ensuring that I respect the standards of minimum disclosure that are applied by the ONS, only disclosing unit which contain > 10 investors or applicants respectively.

3. DGMPD variables

Table 1 indicates the variables that are being built in the DGMPD, and the variables that I intend to offer as part of the DGMPD exporting tool.

Table 1: Availability of selected variables in the DGMPD

Variable	Raw	Export
Party-level variables		
Inventor name	Yes	No
Inventor address (full raw address string)	Yes	No
Inventor postcode	Yes	No
Inventor party-level geocode (latitude, longitude, precision)	Yes	No
Applicant name	Yes	No
Applicant address (full raw address string)	Yes	No
Applicant postcode	Yes	No
Applicant party-level geocode (latitude, longitude, precision)	Yes	No
Patent- and family-level variables (by inventor or applicant location)		
Patent identifiers (publication/application identifiers)	Yes	No
Family identifier	Yes	No
Priority year	Yes	Yes
Publication date / publication year	Yes	Aggr./Anon.
Patent family size / patent count	Yes	Aggr./Anon.
Route and authority markers (e.g. EP, WO, PCT)	Yes	Aggr./Anon.
IPC / CPC classifications	Yes	Aggr./Anon.
Backward citation measures	Yes	Aggr./Anon.
Forward citation measures	Yes	Aggr./Anon.
Co-patenting / collaboration links	Yes	Aggr./Anon.
Materialised export-layer variables		
Inventor geography-year patent counts	Yes	Yes
Applicant geography-year patent counts	Yes	Yes
Technology taxonomy assignments (OECD green, digital, AI, etc.)	Yes	Yes
Relatedness indices and percentiles	Yes	Yes
Complexity indices and percentiles	Yes	Yes
Geography identifiers and labels (city, region, country, global)	Yes	Yes
Geography centroids (latitude / longitude)	Yes	Yes
Reliability / QA flags and score missingness	Yes	Yes
Source count used	No	Yes

Notes: *Aggr./Anon* indicates aggregated or anonymised where applicable. All patents have been deduped across jurisdictions/patent family. The export layer is geography-year oriented and does not currently expose party-level addresses, names, or citation/link records as direct export columns. In line with standard statistical practice observations are only disclosed where >10 observations are present.

4. Retrieval pipeline

4.1. Sources

The DGMPD will use multiple APIs because inconsistencies and missingness arise across authorities. For example, although the EPO's OPS provides worldwide bibliographic coverage, the country geography of these patents is in fact much more limited. I list the current sources for which I have been granted access or have begun cleaning the data in Table 2.

Table 2: Core source inputs and wider API access conditions

Source	What it provides	Access / cost	Operational constraint	Importance
EPO OPS[33]	Programmatic access to patent data, including XML-based bibliographic and register materials	Free up to 4 GB/week; EUR 2,800/year above that threshold	Weekly data cap for non-paying users	Core large-scale retrieval and enrichment
WIPO PATENTSCOPE Webservice[35]	Bibliographic XML and TIFF images for published international applications	CHF 600/year	Conditional-use API; PCT applications only	Official PCT-layer source
USPTO ODP API[34]	US Patent File Wrapper Documents	Free	Weekly limit of 1,200,000 calls	Major US document source

Beyond the data already being processed, the WIPO API Catalog identifies a growing ecosystem of patent-related APIs, including national services for Japan,[64] Australia,[65] Germany,[31] Poland[66], and Kazakhstan.[67] These represent promising avenues for improving the representativeness of the DGMPD but must be evaluated against their country specific delivery systems, e.g. the German DMPA offers patents in weekly, rather than daily updates.

4.2. Retrieval and extraction of addresses

As this is a source-specific harvesting of patent records, the exact details of retrieval vary. For example, the USPTO data required an additional stage of preprocessing (PDF OCR) to convert into usable input data.‖ Alternatively, the OPS data required calls to two separate APIs provided by EPO to populate the geography and the technologies. The full details of these specific processes and their coverage and selection are available on request, and will be published in the final full Appendix.

I provide a summary of the commonalities of each approach and how they relate to each other here.

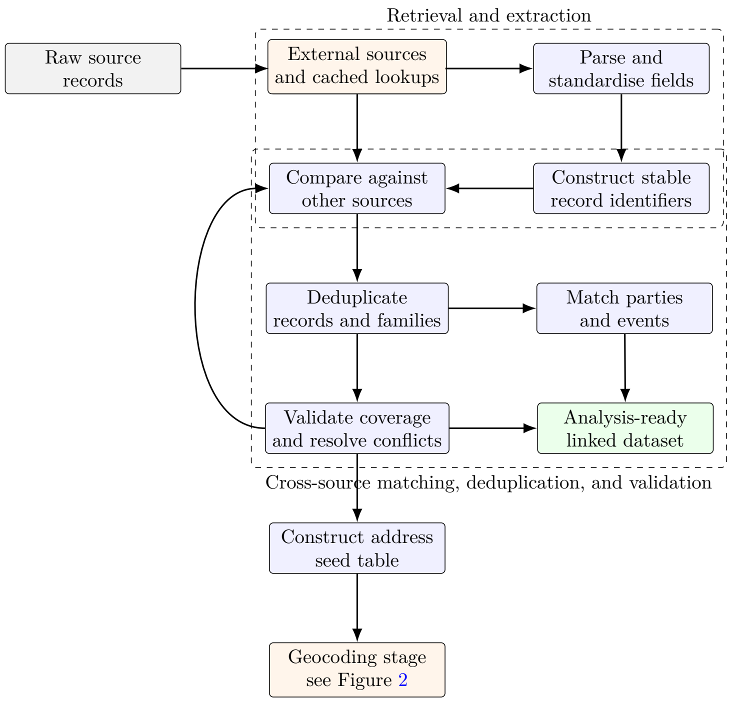
Figure 1 shows the intuition of how records are retrieved, parsed, standardised, compared with auxiliary sources, deduplicated using stable identifiers, and matched across parties, events, addresses, and patent families. A key issue is that patent protection is jurisdiction-specific, so inventive activity linked to one country may appear in multiple domestic and international filings[23;53]. This can lead to imprecision in capturing patenting activity - either missing patents

‖The OPS data is available in XML,[33] whereas the WIPO data comes in TIFF,[35] and the USPTO data is PDF.[34]

that were filed by , or, more likely, double counting if the inventors file in multiple jurisdictions.[25;68] It is therefore important that analysis concentrates patent families rather than a one-to-one mapping between inventions and documents.[69]

For each priority-date cell, patent records are retrieved from the relevant source systems and, where needed, split into smaller filing-date partitions. Records are then parsed into a common schema covering application and publication identifiers, priority claims, procedural routes, parties, classifications, citations, and address fields. Publication records are standardised to stable DOCDB identifiers and compared across sources. Additional enrichment calls recover missing fields such as classifications, citations, family identifiers, publication metadata, and party details. Payloads are cached by publication key to avoid duplication, and family identifiers are used for deduplication where available, with publication- and application-level keys retained as fallbacks.

Figure 1: General workflow for transforming raw source records into addresses



5. Address matching

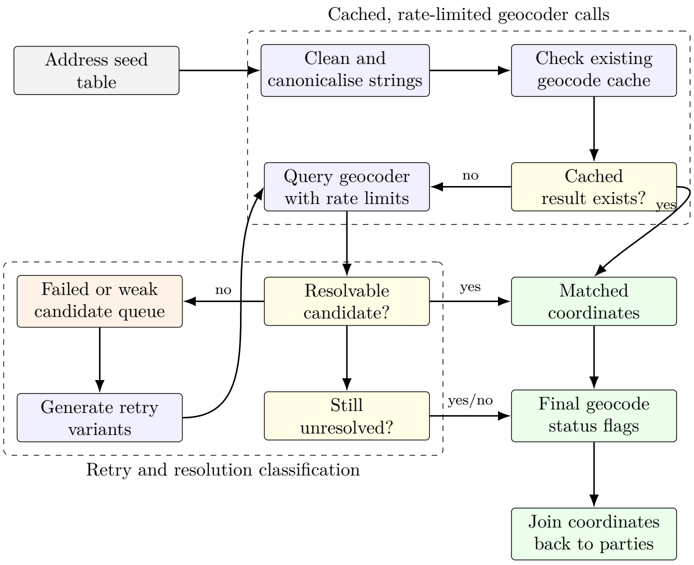
5.1. Method

In the next stage I use a cached adddress-resolution geocoding, implemented as a separate cached address-resolution stage, the logic of which is summarised in Figure 2. Applicant and inventor address strings are first cleaned,[**] standardised, and reduced to canonical cache keys so that repeated addresses are only queried once. Existing cache entries are reused where available; otherwise, unresolved addresses are submitted to Nominatim[38] a rate-limited geocoder based

[**]For instance, many addresses include the Floor number of the office of the inventor, which offers no value in estimating the coordinate location.

on OpenStreetMap data that is the standard open source approach.[70;71††] Returned candidates are classified by match quality and stored with both coordinates and resolution status.

Figure 2: Geocoding workflow for address resolution and retry handling



The geocoding acceptance rule follows the geocoding-quality literature, which treats match type and positional precision as central dimensions of geocode reliability.[74–76‡‡] Prior work shows that automated geocoding error varies systematically with the spatial reference used, with address- or street-level matches generally more precise than postal-code, city, county, or other areal centroids.[77;78] This matters for the microgeography of innovation because positional error can attenuate or distort spatial relationships, particularly when analysis is conducted at fine spatial scales.[21] I therefore accept Nominatim results as matched coordinates only when the returned precision is postcode-level or finer and the textual evidence indicates medium or high confidence. Broader city-, state-, or country-level candidates are retained as candidate-only fallback evidence rather than treated as final point locations. This provides a sensible balance when analysing innovation locations as it recognises that there may be useful partial information while avoiding the assumption that coarse administrative centroids are equivalent to precise address geocodes.[79]

Addresses that fail to resolve, or return only weak candidates, are placed in a retry queue. Retry variants progressively scale the geography up: the pipeline first attempts the most specific available address string, then relaxes the query by dropping street-level detail, using city-region-country combinations, and finally broader regional or country-level forms where necessary. This preserves high-precision matches when possible while still recovering approximate locations for incomplete or noisy address records. Final outputs distinguish precise matches, broader fallback matches, candidate-only cases, and unresolved addresses before joining coordinates back to the patent-party records.s

[††]Nominatim is free, but slow, limiting calls to 1 per second, which presents a medium-term delay to scaling this analysis. Alongside Nominatim, over time, I would like to develop DGMPD's address validation to analyse the same address keys using paid variants such as Google for Developers[72] and/or HERE[73], to help allay any concerns over the precision accuracy of Nominatim's co-ordinate selection.

[‡‡]What constitutes a strong or weak match is inherently subjective. Full details on matching logic, and robustness, will be available online as I build out the addresses across patent offices.

5.2. Results

There are few control groups to compare this data set (discussed more directly below), but this approach represented an improvement compared to previous approaches in terms of granularity and match rate. The focus of Morrison et al. [21] was less on geographic precision and more on individual linking, and so . De Rassenfosse et al. [26]'s geocoded analysis of 7 million patents stages at the country level, and I could not find details on the scale of matching in the main text nor the supplementary information. It is unfair to compare Bergeaud and Verluise [29]'s historical account of 5 cities in the PatCity database, as their focus was to chart crosscity change over decades rather than having microspatial trends as their object, but as an indication their 5 city patent panel achieved a below-city matching precision of 17.5%, and thus the DGMPD presents a signifcant improvement in below-city understanding of innovation.

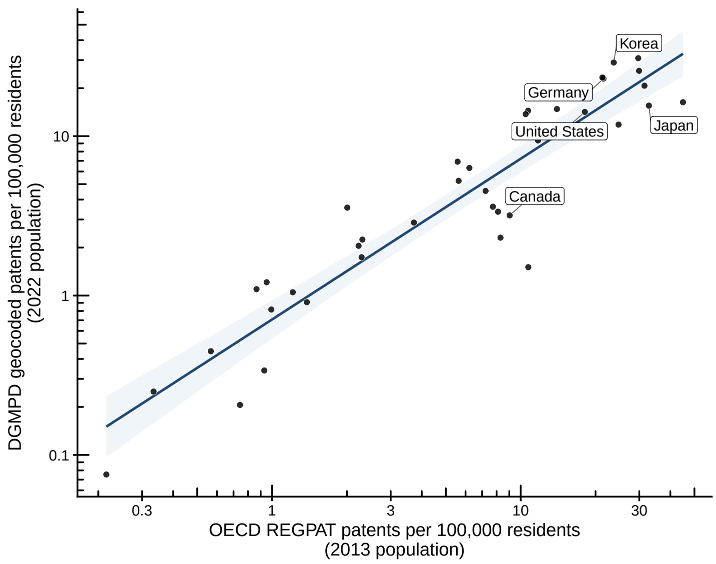
Table 3: Percentage summary of extracted and resolved EPO OPS records

Component	EPO OPS
Years covered	2022
Raw document rows	281 115
Raw document rows with any address	100.000%
Deduplicated patent documents	279 072
Deduplicated patent documents with any address	100.000%
Applicant rows	308 996
Applicant rows with any address	100.000%
Inventor rows	938 448
Inventor rows with any address	99.330%
Unique address keys	199 201
Input composition among resolved address keys	
Country present in input key	99.997%
Postcode present in input key	79.069%
City / locality present in input key	86.695%
State / province present in input key	11.469%
County / district present in input key	0.000%
Resolution outcomes among resolved address keys	
Matched	87.307%
Candidate only	7.215%
Unmatched	5.478%
Matched at postcode precision	58.952%
Matched at address precision	21.976%
Matched at street precision	5.149%
Matched at city precision	1.101%
Matched at state precision	0.106%
Matched at county precision	0.021%
Matched at country precision	0.001%
Technological coverage among document rows	
Documents with IPC codes	99.838%

6. Validation against existing geocoded patent sources

As a provisional external validation exercise, I compare the DGMPD against the most recent public release of the OECD REGPAT database. The most up-to-date publicly available version I could locate appears to extend to 2013, which necessarily limits direct comparability with the 2022 DGMPD subsample used here. Over the intervening decade, the economic geography of innovation changed substantially. More recent comparison data are either access-restricted through the OECD or available via paid PATSTAT products.[55;82] These comparisons should therefore be treated as provisional rather than definitive.

Figure 3: DGMPD vs. OECD REGPAT (national)

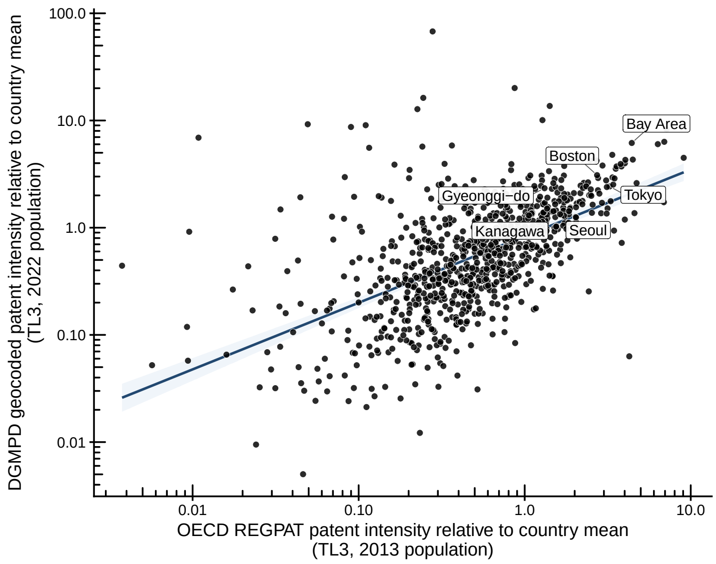


The national comparison is nevertheless encouraging. Even with differences in period, source construction, and coverage, the DGMPD geocoded patent counts per 100,000 residents track the OECD REGPAT benchmark reasonably closely across countries, with the relationship appearing broadly positive in the country-level comparison figure.

I also present a within-country regional comparison using population-adjusted patent intensity measures in both datasets. Specifically, both measures are normalised by population, and the regional figure expresses patent intensity relative to the relevant country mean. This choice makes the comparison less sensitive to differences in national patenting levels and better suited to assessing whether the DGMPD captures subnational spatial structure. The resulting relationship is again positive, with major innovation centres such as Tokyo, Kanagawa, Seoul, Gyeonggi-do, Boston, and the Bay Area appearing in plausible positions in both sources.

Between 2013 and 2022, global patenting increased from 2.57 million[80] to 3.46 million,[81] with this growth itself being spatially uneven, including strong concentration in China.

Figure 4: DGMPD vs. OECD REGPAT (regional)



I emphasise that these results are encouraging, but still provisional and well short of a long-run validation standard. If the contribution of the DGMPD is the dynamic measurement of the microgeography of innovation through APIs, then it requires deeper comparison with existing within-region geocoded sources wherever such benchmarks can be assembled, despite their scarcity. I am therefore exploring more robust validation strategies, including patent-level matching approaches in the spirit of De Rassenfosse et al. [26] and Bergeaud and Verluise [29], as compute capacity and data infrastructure scale.

7. Results

This section documents the potential of the DGMPD while adhering to standard statistical approaches on aggregation and anonymisation of records.

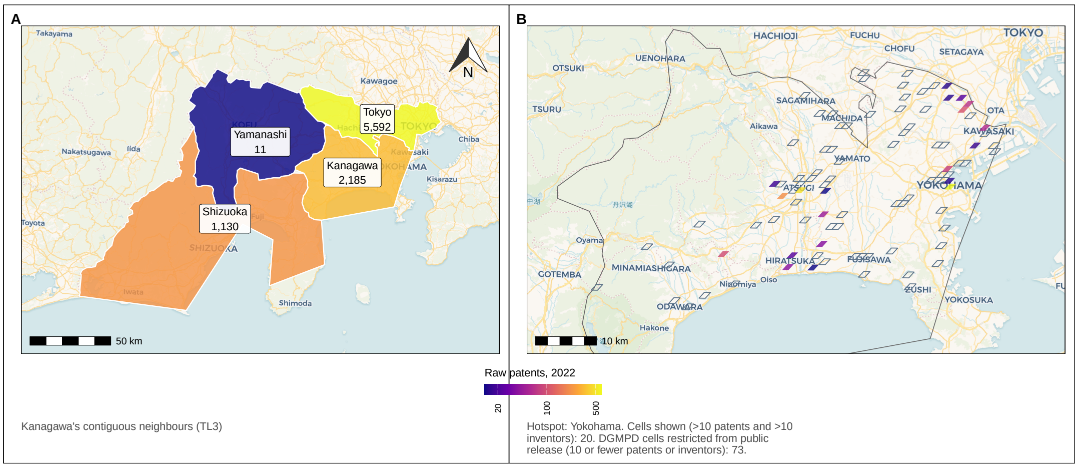
7.1. Global patenting

This is not surprising (nor at all novel) given the aggregate consistencies with the OECD data. What is more interesting about this data is illustrated in the following case studies. I pick Kanagawa, a prefecture near Tokyo, and the Bay Area, San Francisco, as illustrative examples of the empirical potential of this microdataset. This analysis can be applied worldwide.

7.2. Case studies

Kanagawa Figure 5 illustrates both the value of the DGMPD for conventional regional comparison and its additional value at finer spatial scale. Panel A places Kanagawa within its contiguous TL3 context, showing that it is a major innovative region in its own right, though still smaller in aggregate terms than neighbouring Tokyo. Panel B then zooms within Kanagawa itself, revealing that inventive activity is far from evenly distributed across space. Instead, it is concentrated in a limited number of cells, with a clear hotspot around Yokohama and additional pockets extending across the eastern urban corridor. The associated Moran's I statistic indicates that this pattern is not random, but clustered, which is consistent with the idea that innovative activity is organised through dense local environments rather than spread uniformly across administrative space.

Figure 5: Geography of innovation in Kanagawa, Japan



Bay Area Figure 6 provides a parallel case for the Bay Area. Panel A shows that the TL3 geography here covers a much larger and more internally heterogeneous space than in Kanagawa, highlighting a broader problem with administrative units: they are often not directly comparable across countries or even especially meaningful for the immediate geography of innovation. Such units may still be useful for studying broader objects such as labour markets, commuting systems, or regional production structures, but they are less well suited to identifying the sites at which innovative activity actually concentrates, where highly localised spillovers may operate at the level of the office, street, or neighbourhood.

Panel B therefore shifts to the DGMPD's finer spatial resolution, making it possible to compare innovative intensity across places on a more equal footing. At this scale, inventive activity is again strongly clustered, with a pronounced hotspot around Mountain View and a wider corridor running through Palo Alto, Sunnyvale, Santa Clara, and San Jose. The comparison shows that, although both regions are innovative, the Bay Area exhibits a denser and more extensive microspatial concentration of patenting activity.

Microspatial sectoral profiles Because the DGMPD is technology-specific, it also allows place–sector comparisons across and within regions. Figure 7 illustrates this by comparing Kanagawa and the Bay Area at both the TL3 and grid levels across the 35 WIPO technology fields dseigned for cross country comparison (which are constructed from raw IPC, available in the DGMPD).[83] Panel A shows that the two regions differ materially in their broad technological composition: the Bay Area is more heavily oriented towards computer technology and digital communication, while Kanagawa has relatively stronger shares in semiconductors and transport-related activity. Panel B shows what is gained from moving to 1km^2: not just regional specialisation, but within-area variation in how innovation is organised. Palo Alto, for example, appears as a highly specialised computer-technology hotspot, whereas Kanagawa innovates through a more sectorally mixed local structure. The DGMPD therefore offers greater precision when analysing hyper-local innovation.

Figure 6: Geography of innovation in the Bay Area

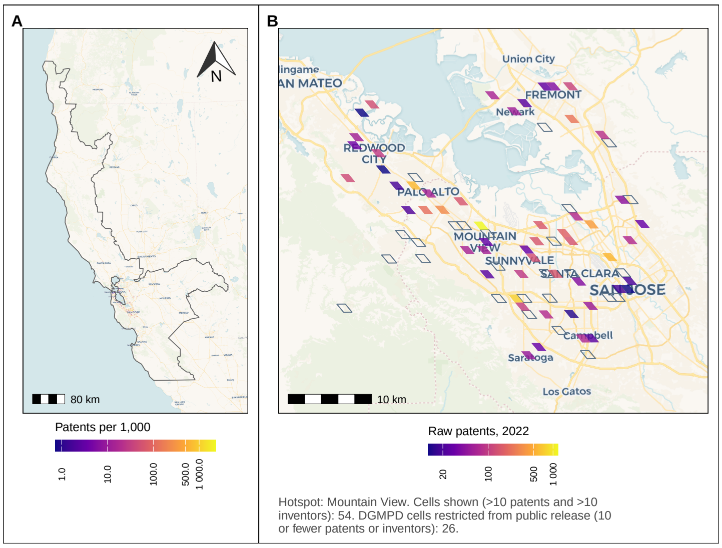
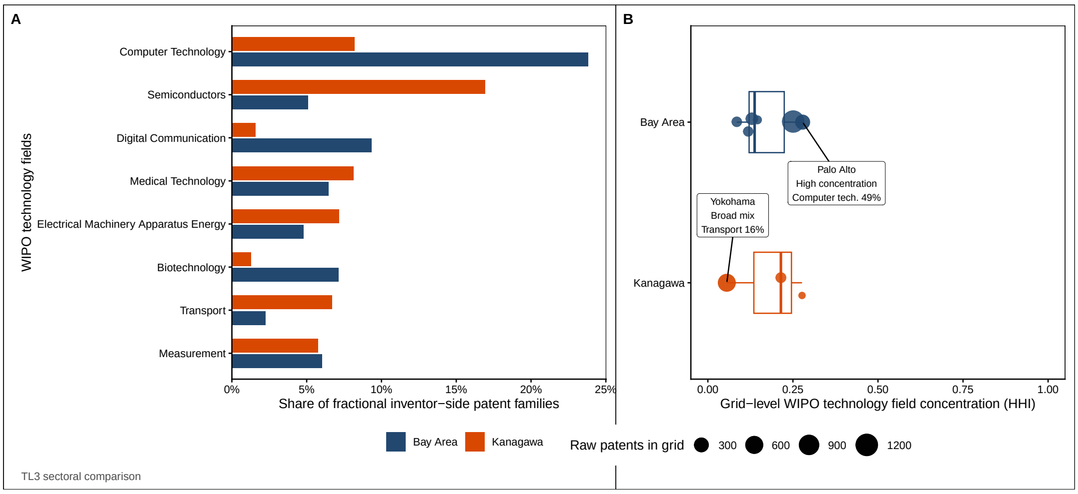


Figure 7: TL3 to 1km2 specialisation in Kanagawa, JP vs. the Bay Area, SF, USA



8. Access to the DGMPD

I would like the DGMPD to be used by researchers. I am therefore in the process of building and hosting an export platform where individuals can access this data and filter by location (inventor/applicant), geography (e.g. a certain country or region), scale (public release only regions, for now), years, indicators (e.g. total patents, patents by technology class, green patents, including within-area variations in aggregations, etc.). These files are often incredibly larger De Rassenfosse et al. [26]'s raw data is +200 GB, and often for useful research only a small slice of this.

Currently, as I scale up this analysis, access will be on an email request basis, where users will be able to understand the shape of these data, in the hopes that they contribute feedback and improvements while the project is in its 'beta' mode. Please do get in touch.

9. Future directions for the DGMPD and research

9.1. Developing the DGMPD

An important part of the DGMPD is trust. That is why I document my processes openly, underscoring that this pipeline is reproducable by anyone with the funds for the start up costs required in terms of licences and a computer or server large enough to process and host this data. Information on the pricing and licences is available above in Table 2. As the number of APIs increase in this centralised microgeolocated database, these burdens will however intensifiy. For transparency on the computational burden of this data, as an indication, a VPS server with 32GB RAM, using 8-12 workers took, around 3 days to process the USPTO data for 2022.

There is an obvious tradeoff between granularity/frequency and noise in the data. If the DGMPD is to be a useful research tool then it must be auditable and comparable. Moving out of beta stage of the DGMPD, I would therefore suggest a format similar to the current Overture Maps Foundation [84] releases, which collates large place data, with (roughly) monthly releases providing comparablity across researchers while also catering for the rapidly growing availability (requirement) that scale scale high frequency data brings for analysis. I am open to feedback on this though.

9.2. Future directions for research

For research, this opens up new possibilities for analysing the microgeography of innovation. Given the location and technologies specific data that the DGMPD presents there are numerous ways to advance existing research. There are clear extensions on finer grained spatial analysis of technological complexity, [41] relatedness [85;86] and embedding. [87] For economic geography there is the possiblity to use INADOC (same broad patent families) groupings over time to analyse the extent to which regions become victim of 'lock in' effects. [88;89] Work on spatial spillovers – and the co-incidence of global pipelines and local knowledge clusters [90] – measured through co-patenting is another clear development.

I propose a few initial thoughts that depart from these standard literatures here.

Including the varying definitions based on OECD ENV-TECH or CYCFavot et al. [39]

nicholas.sweeney@manchester.ac.uk

Greater granularity and representativeness in increased richness in microspatial data analysis also allows researchers to ask new quantitative questions on broader processes of economic change and social equity. For instance, proposed and now plausible line of future research may be to analyse the extent to which knowledge that is produced by inventors rests in their area over time, analysing the centralisation of the appropriation of knowledge (proxied by the disparity in inventor vs. applicant address). Informed by the literature global cities and the wider human geography literature,[91;92] this would provide an insight into the spatial division of labour in more detail and the reorganisation of society through broad economic trends.[93;94] In doing so, his brings to the fore questions of value capture in an area which were previously only analysed at the national scale.[95;96]

Daily variation and noise in patenting data means that for most, there is little interest in these trends, with annual trends providing a more practical summary of useful insights. However, as an economic geographer increased in the role of time, as well as space, in answering questions of location, I also would argue, on the dismissal of 'noise', here that potentially useful data is being ignored. For instance, given growing concerns over mental health and anxiety in society,[97] this data could offer insights on the role of 'involution' (collective burnout), as concept explored more but not necessarily limited to China,[98] controlling for area-industry fixed effects, if a certain area began increasing in the % of weekend patenting, being largely an administrative, 'overspill' weekend task, this could be argued to proxy certain groups are put under greater temporal stress.

As a word of caution, greater granularity presents not only computing and methodological challenges, but also epistemological ones. Considering the MAUP, there are problems that are more acute at finer spatial scales. For instance, there is uncertainty around whether inventor addresses are the site of activity, or the residential address of the inventors themselves.[99;100] There are ways of possibly handling this, for instance assuming that if co-invention occurs at the same address, that is reliably an office, or modelling by aggregate built environment classification using something like the UN's Urban Centre Database containing data on over 11,000 cities based on the Global Human Settlement Layer.[101] However, these are modelling assumptions, and by their nature imperfect. While they represent possible avenues for further research, this is a moot point. The greater granularity that is presented in this article should therefore not be considered a 'solution' to these problems surrounding measurement of the geography of innovation.

10. Conclusion and call for interest

The DGMPD is 'living' in the sense that given the backbone of the database is API-based, it evolves with daily updates and reflects a growing landscape of patenting.

However it is also 'living' in its methodological approach. I would welcome comments from researchers on improvements to this pipeline, and also what they would find useful as outputs. The objective of this is that the dataset provides the most useful and holistic empirical backbone for microspatial analysis while complying with appropriate confidentiality standards.

References

[1] Maryann P Feldman and Dieter F Kogler. Stylized facts in the geography of innovation. *Handbook of the Economics of Innovation*, 1:381–410, 2010.

[2] Chris Forman and Avi Goldfarb. Concentration and agglomeration of it innovation and entrepreneurship: Evidence from patenting. Technical report, National Bureau of Economic Research, 2020.

[3] OECD. *Place-Based Policies for the Future*. OECD Regional Development Studies. OECD Publishing, Paris, 2025. doi: 10.1787/e5ff6716-en. URL https://doi.org/10.1787/e5ff6716-en.

[4] Ricardo Hausmann, Muhammed Ali Yildirim, Christian Chacua, Matte Hartog, and Shreyas Gadgin Matha. Innovation policies under economic complexity. *World Intellectual Property Organization (WIPO) Economic Research Working Paper Series*, (79), 2024.

[5] Cassandra Sweet and Dalibor Eterovic. Do patent rights matter? 40 years of innovation, complexity and productivity. *World Development*, 115:78–93, 2019.

[6] Riccardo Crescenzi and Andrés Rodríguez-Pose. *Innovation and regional growth in the European Union*. Springer Science & Business Media, 2011.

[7] OECD. *OECD Science, Technology and Innovation Outlook 2025: Driving Change in a Shifting Landscape*. OECD Publishing, Paris, 2025. ISBN 9789264399426. doi: 10.1787/5fe57b90-en. URL https://doi.org/10.1787/5fe57b90-en.

[8] Patricia Peñalosa and Lukas Kleine-Rueschkamp. The geography of green innovation hubs in oecd regions. Technical report, OECD, Paris, 2024.

[9] Nguyen T Tran, Tung D Nguyen, My Nguyen, and Phuong MH Pham. Climate vulnerability and green innovation: A cross-country analysis. *Journal of Open Innovation: Technology, Market, and Complexity*, 11(2):100548, 2025.

[10] Andres Rodríguez-Pose and Federico Bartalucci. Regional vulnerability to the green transition. Technical report, Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs . . . , 2023.

[11] Flavio L Pinheiro, Pierre-Alexandre Balland, Ron Boschma, and Dominik Hartmann. The dark side of the geography of innovation: relatedness, complexity and regional inequality in europe. *Regional Studies*, 59(1):2106362, 2025.

[12] Harald Bathelt, Maximilian Buchholz, and Michael Storper. The nature, causes, and consequences of inter-regional inequality, 2024.

[13] Amy E Thompson, Jessica Munson, Scott G Ortman, Andrés G Mejía Ramón, Gary M Feinman, Gabriela Cervantes Quequezana, Pablo Cruz, Adam S Green, Dan Lawrence, and Paul Roscoe. Assessing neighborhoods, wealth differentials, and perceived inequality in preindustrial societies. *Proceedings of the National Academy of Sciences*, 122(16): e2400699121, 2025.

[14] David B Audretsch and Maryann P Feldman. R&d spillovers and the geography of innovation and production. *The American economic review*, 86(3):630–640, 1996.

[15] Diego D'Adda, Donato Iacobucci, and Francesco Perugini. Regional technological profiles and collaborations: An empirical analysis of joint patents and eu-funded projects. *Papers in Regional Science*, 104(1):100073, 2025.

[16] Gabriel M Ahlfeldt, Stephan Heblich, and Tobias Seidel. Micro-geographic property price and rent indices. *Regional Science and Urban Economics*, 98:103836, 2023.

[17] Melanie Krause and Sebastian Kripfganz. Regional dependencies and local spillovers: Insights from commuter flows. *Journal of Regional Science*, 65(3):565–585, 2025.

[18] Martin Andersson, Johan Klaesson, and Johan P Larsson. How local are spatial density externalities? neighbourhood effects in agglomeration economies. *Regional studies*, 50(6): 1082–1095, 2016.

[19] Edward L Glaeser, Hedi D Kallal, Jose A Scheinkman, and Andrei Shleifer. Growth in cities. *Journal of political economy*, 100(6):1126–1152, 1992.

[20] Riccardo Crescenzi, Max Nathan, and Andrés Rodríguez-Pose. Do inventors talk to strangers? on proximity and collaborative knowledge creation. *Research policy*, 45(1): 177–194, 2016.

[21] Greg Morrison, Massimo Riccaboni, and Fabio Pammolli. Disambiguation of patent inventors and assignees using high-resolution geolocation data. *Scientific data*, 4(1):1–21, 2017.

[22] EPO. Epo worldwide legal event data (inpadoc), 2026. URL https://www.epo.org/en/searching-for-patents/data/bulk-data-sets/inpadoc.

[23] WIPO. Pct faqs, 2026. URL https://www.wipo.int/en/web/pct-system/faqs/faqs.

[24] Stéphane Maraut, Hélène Dernis, Colin Webb, Vincenzo Spiezia, Dominique Guellec, et al. The oecd regpat database. *OECD Science, Technology and Industry Working Papers). Paris: Organisation for Economic Co-operation and Development. Retrieved September*, 2:2013, 2008.

[25] Eric W Bond and Zachary R Luther. International patent families: When to apply. *Review of International Economics*, 33(2):450–465, 2025.

[26] Gaétan De Rassenfosse, Jan Kozak, and Florian Seliger. Geocoding of worldwide patent data. *Scientific data*, 6(1):260, 2019.

[27] Jared Hewko, Karen E Smoyer-Tomic, and M John Hodgson. Measuring neighbourhood spatial accessibility to urban amenities: does aggregation error matter? *Environment and Planning A*, 34(7):1185–1206, 2002.

[28] Sharon Schwartz. The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. *American journal of public health*, 84(5):819–824, 1994.

[29] Antonin Bergeaud and Cyril Verluise. A new dataset to study a century of innovation in europe and in the us. *Research Policy*, 53(1):104903, 2024.

[30] Bernardo Caldarola, Dario Mazzilli, Lorenzo Napolitano, Aurelio Patelli, and Angelica Sbardella. Economic complexity and the sustainability transition: A review of data, methods, and literature. *Journal of Physics: Complexity*, 5(2):022001, 2024.

[31] German Patent and Trade Mark Office. Dpmaconnectplus, 2026. URL https://www.dpma.de/english/search/data_supply_services/dpmaconnect/index.html. Accessed 16 April 2026.

[32] Mengxue Zheng, Lili Miao, Yi Bu, and Vincent Larivière. Understanding discrepancies in the coverage of openalex: The case of china. *Journal of the Association for Information Science and Technology*, 76(11):1591–1601, 2025.

[33] EPO. Open patent services (ops). https://www.epo.org/en/searching-for-patents/data/web-services/ops, 2026. Accessed 15 April 2026.

[34] USPTO. Getting started. https://data.uspto.gov/apis/getting-started, 2026. USPTO APIs documentation. Accessed 15 April 2026.

[35] WIPO. Patentscope data. https://www.wipo.int/en/web/patentscope/data/index, 2026. Accessed 15 April 2026.

[36] Ray Smith. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE, 2007.

[37] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

[38] Nominatim. Nominatim. https://nominatim.org/, 2026. Open-source geocoding with OpenStreetMap data. Accessed 15 April 2026.

[39] Marinella Favot, Leyla Vesnic, Riccardo Priore, Andrea Bincoletto, and Fabio Morea. Green patents and green codes: How different methodologies lead to different results. *Resources, Conservation & Recycling Advances*, 18:200132, 2023.

[40] Ricardo Hausmann, Daniel P Stock, and Muhammed A Yıldırım. Implied comparative advantage. *Research Policy*, 51(8):104143, 2022.

[41] Leon Mewes and Tom Broekel. Technological complexity and economic growth of regions. *Research Policy*, 51(8):104156, 2022.

[42] Christian Chacua, Shreyas Gadgin Matha, Matte Hartog, Ricardo Hausmann, and Muhammed A Yildirim. *Innovation Policies Under Economic Complexity*. Centre for Economic Policy Research, 2025.

[43] Ernest Miguelez, Julio D Raffo, Christian Chacua, Massimiliano Coda-Zabetta, Deyun Yin, Francesco Lissoni, and Gianluca Tarasconi. Tied in: The global network of local innovation. *World Intellectual Property Organization (WIPO) Economic Research Working Paper Series*, (58), 2019.

[44] Si Joo and Yeonbae Kim. Measuring relatedness between technological fields. *Scientometrics*, 83(2):435–454, 2010.

[45] Seung Hwan Kim, Bogang Jun, and Jeong-Dong Lee. Technological relatedness: how do firms diversify their technology? *Scientometrics*, 128(9):4901–4931, 2023.

[46] David L Rigby. Technological relatedness and knowledge space: Entry and exit of us cities from patent classes. *Regional Studies*, 49(11):1922–1937, 2015.

[47] Marc Baudry and Béatrice Dumont. *Patents: prompting or restricting innovation?*, volume 12. John Wiley & Sons, 2017.

[48] Amy Kapczynski and Talha Syed. The continuum of excludability and the limits of patents. *Yale LJ*, 122:1900, 2012.

[49] Yuxuan Lan, Ziyue Yuan, Ruiqi Tang, Shu-Chien Hsu, and Hsi-Hsien Wei. Green innovation and the esg disconnect: Evidence from green patenting in the construction industry in china. *Journal of Management in Engineering*, 41(1):04024066, 2025.

[50] Donald F Turner. Patents, antitrust and innovation. *U. Pitt. L. Rev.*, 28:151, 1966.

[51] Yu-Kai Lin and Likoebe M Maruping. Open source collaboration in digital entrepreneurship. *Organization Science*, 33(1):212–230, 2022.

[52] Christian Spreafico, Nils Thonemann, Massimo Pizzol, Rickard Arvidsson, Bernhard Steubing, Stefano Cucurachi, Giuseppe Cardellini, and Matteo Spreafico. Using patents to support prospective life cycle assessment: opportunities and limitations. *The International Journal of Life Cycle Assessment*, 30(2):201–220, 2025.

[53] Srinivasan Ananthraman, Bart Cambré, Markus Kittler, and Henry Delcamp. Divide and conquer: Relating patent quality and value in a conceptual framework based on a systematic review. *International Journal of Management Reviews*, 26(2):285–311, 2024.

[54] The Lens. Patseq data. https://support.lens.org/knowledge-base/patseq-data/, 2023. Accessed 15 April 2026.

[55] EPO. *PATSTAT Global Data Catalog*, sep 2025. URL https://link.epo.org/web/business/patstat/en-data-catalog-patsat-global-autumn-2025.pdf. Autumn 2025 edition, accessed 16 April 2026.

[56] European Parliament and Council of the European Union. Regulation (ec) no 1059/2003 of the european parliament and of the council of 26 may 2003 on the establishment of a common classification of territorial units for statistics (nuts). Official Journal of the European Union, L 154, 21 June 2003, pp. 1–41, 2003. Establishing a common classification of territorial units for statistics (NUTS).

[57] Martina Pardy. Multinationals and intra-regional innovation concentration. *Research Policy*, 54(6):105235, 2025.

[58] Bureau van Dijk. Orbis intellectual property, 2026. URL https://login.bvdinfo.com/R1/OrbisIntellectualProperty. Accessed 16 April 2026.

[59] David Antons and Frank T Piller. Opening the black box of “not invented here”: Attitudes, decision biases, and behavioral consequences. *Academy of Management perspectives*, 29(2):193–217, 2015.

[60] Anand Nandkumar and Kannan Srikanth. Right person in the right place: H ow the host country ipr influences the distribution of inventors in offshore r&d projects of multinational enterprises. *Strategic Management Journal*, 37(8):1715–1733, 2016.

[61] World Intellectual Property Organization. Patentscope® search service: Removal of address data. https://www.wipo.int/en/web/patentscope/w/news/2008/news_0013, dec 2008. 15 December 2008. Accessed 15 April 2026.

[62] European Patent Office. Terms and conditions for the delivery and use of patent information products and registration for events organised by the epo. https://www.epo.org/en/service-support/ordering/pi-terms-and-conditions, 2026. Accessed 15 April 2026.

[63] Grazia Sveva Ascione and Andrea Vezzulli. Leveraging nlp and web knowledge graphs to harmonize locations: A case study on us patent transactions. *World Patent Information*, 79:102320, 2024.

[64] Japan Patent Office. Provision of patent information using api on a trial basis, may 2025. URL https://www.jpo.go.jp/e/system/laws/koho/internet/api-patent_info.html. Accessed 16 April 2026.

[65] IP Australia. Apis, 2026. URL https://www.ipaustralia.gov.au/tools-and-research/professional-resources/apis. Accessed 16 April 2026.

[66] Patent Office of the Republic of Poland. Pueup api zewnętrzne, 2026. URL https://api.uprp.gov.pl/. Accessed 16 April 2026.

[67] National Institute of Intellectual Property of the Republic of Kazakhstan. About the institute, 2026. URL https://kazpatent.kz/en/qazpatent-turaly/institut-turaly. Official institute page; accessed 16 April 2026.

[68] P Dimas, D Stamopoulos, and A Protogerou. A longitudinal dataset of sector-level patent data for europe. *Data in Brief*, page 112095, 2025.

[69] Antoine Dechezleprêtre, Yann Ménière, and Myra Mohnen. International patent families: from application strategies to statistical indicators. *Scientometrics*, 111(2):793–828, 2017.

[70] Konstantin Clemens. Geocoding with openstreetmap data. *GEOProcessing 2015*, 10, 2015.

[71] Lukas Kriesch and Sebastian Losacker. A geolocated dataset of german news articles. *Scientific Data*, 12(1):1128, 2025.

[72] Google for Developers. Geocoding api overview. https://developers.google.com/maps/documentation/geocoding/guides-v3/overview, 2026. Accessed 15 April 2026.

[73] HERE. Here developer portal. https://www.here.com/developer, 2026. Accessed 15 April 2026.

[74] Sunil Bisnath and Yang Gao. Precise point positioning. *GPS world*, 20(4):43–50, 2009.

[75] Mario Schootman, David A Sterling, James Struthers, Yan Yan, Ted Laboube, Brett Emo, and Gary Higgs. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Annals of epidemiology*, 17(6):464–470, 2007.

[76] Zhengcong Yin, Daniel W Goldberg, Binbin Lin, Bing Zhou, Diya Li, Andong Ma, Ziqian Ming, Heng Cai, Zhe Zhang, Shaohua Wang, et al. Toward building next-generation geocoding systems: a systematic review. *arXiv preprint arXiv:2503.18888*, 2025.

[77] Michael R Cayo and Thomas O Talbot. Positional error in automated geocoding of residential addresses. *International journal of health geographics*, 2(1):10, 2003.

[78] Paul A Zandbergen. Geocoding quality and implications for spatial analysis. *Geography Compass*, 3(2):647–680, 2009.

[79] Andreas Behr, Christoph Schiwy, and Lucy Hong. Overcoming the maup: impact of agglomeration economies on regional performance in germany. *Applied Economics*, 57 (40):6244–6257, 2025.

[80] World Intellectual Property Organization. Global intellectual property filings up in 2013, china drives patent application growth, dec 2014. URL https://www.wipo.int/pressroom/en/articles/2014/article_0018.html. Press release PR/2014/771, accessed 16 April 2026.

[81] WIPO. *World Intellectual Property Indicators 2023*. WIPO, Geneva, 2023. doi: 10.34667/tind.48541. URL https://www.wipo.int/edocs/pubdocs/en/wipo-pub-941-2023-en-world-intellectual-property-indicators-2023.pdf. Accessed 16 April 2026.

[82] EPO. Patent information products, 2026. URL https://www.epo.org/en/service-support/ordering/patent-information-products. Accessed 16 April 2026.

[83] Ulrich Schmoch. Concept of a technology classification for country comparisons. Technical report, World Intellectual Property Organization, jun 2008. URL https://www.wipo.int/documents/2948119/3215563/wipo_ipc_technology.pdf. Final report to the World Intellectual Property Organization, accessed 16 April 2026.

[84] Overture Maps Foundation. Releases. https://docs.overturemaps.org/release-calendar/, 2026. Overture Maps Documentation. Last updated 30 March 2026. Accessed 15 April 2026.

[85] Frank Neffke, Martin Henning, and Ron Boschma. How do regions diversify over time? industry relatedness and the development of new growth paths in regions. *Economic geography*, 87(3):237–265, 2011.

[86] Frank MH Neffke, Anne Otto, and César Hidalgo. The mobility of displaced workers: how the local industry mix affects job search. *Journal of Urban Economics*, 108:124–140, 2018.

[87] Tasos Kitsos, Simone Maria Grabner, and Andre Carrascal-Incera. Industrial embeddedness and regional economic resistance in europe. *Economic Geography*, 99(3):227–252, 2023.

[88] Ron Boschma. Proximity and innovation: a critical assessment. *Regional studies*, 39(1): 61–74, 2005.

[89] Gernot Grabher. The weakness of strong ties; the lock-in of regional development in ruhr area. *The embedded firm; on the socioeconomics of industrial networks*, pages 255–277, 1993.

[90] Harald Bathelt, Anders Malmberg, and Peter Maskell. Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. *Progress in human geography*, 28 (1):31–56, 2004.

[91] AnnaLee Saxenian. Regional networks and the resurgence of silicon valley. *California management review*, 33(1):89–112, 1990.

[92] Doreen Massey. Towards a critioue of industrial location theory. *Antipode*, 5(3):33–39, 1973.

[93] Doreen Massey. In what sense a regional problem? *Regional studies*, 13(2):233–243, 1979.

[94] Doreen Massey. *Spatial Divisions of Labor*. Routledge, New York, 2 edition, 1995. ISBN 9781003572619. doi: 10.4324/9781003572619.

[95] John Cantwell and Lucia Piscitello. The location of technological activities of mncs in european regions: The role of spillovers and local competencies. *Journal of International management*, 8(1):69–96, 2002.

[96] Michele Cincera, Bruno van Pottelsberghe de la Potterie, and Veugelers Reinhilde. Assessing the foreign control of production of technology: The case of a small open economy. *Scientometrics*, 66(3):493–512, 2006.

[97] Health Foundation. Tackling health inequalities in mayoral and city regions: Impact report 2022, 2022. URL https://www.health.org.uk/sites/default/files/2022-06/tackling_health_inequalities_in_mayoral_and_city_regions_impact_report_0.pdf.

[98] Da Yi, Jingwen Wu, Minqiang Zhang, Qing Zeng, Jinqing Wang, Jingdan Liang, and Yashi Cai. Does involution cause anxiety? an empirical study from chinese universities. *International journal of environmental research and public health*, 19(16):9826, 2022.

[99] Francesco Lissoni and Ernest Miguelez. Migration and innovation: Learning from patent and inventor data. *Journal of Economic Perspectives*, 38(1):27–54, 2024.

[100] Keith Pennington and J Myles Shaver. It matters where people live: innovative collaborations and the proximity of inventors' residences. *Strategy Science*, 2026.

[101] Martino Pesaresi, Marcello Schiavina, Panagiotis Politis, Sergio Freire, Katarzyna Krasnodębska, Johannes H. Uhl, Alessandra Carioli, et al. Advances on the global human settlement layer by joint assessment of earth observation and population survey data. *International Journal of Digital Earth*, 17(1), 2024. doi: 10.1080/17538947.2024.2390454. Recommended generic citation for the latest GHSL release.