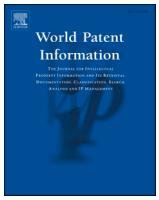
World Patent Information 79 (2024) 102320

Contents lists available at ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin



Leveraging NLP and web knowledge graphs to harmonize locations: A case study on US patent transactions

Grazia Sveva Ascione [a,b,*], Andrea Vezzulli [b]

[a] Bordeaux School of Economics UMR CNRS 6060, University of Bordeaux, 16 Avenue Léon Duguit, 33608, Pessac, France
[b] Università Degli Studi Dell'Insubria Department of Economics - DiECO Via Monte Generoso, 71 - 21100, Varese, Italy

ARTICLE INFO

Keywords:
Patent transactions
Patent assignee
Patent data harmonization
Natural language processing
Knowledge graph

ABSTRACT

In the present study, we introduce a novel methodology for the harmonization and standardization of locations associated with patent transactions recorded at the USPTO from 2005 to 2022. Using natural language processing (NLP) techniques in conjunction with search engine-based web knowledge graphs, our method comprises four phases: data pre-processing, semantic clustering, exploitation of web-knowledge graphs, and API-driven harmonization. Initiating our analysis with a dataset of 63,838 unique locations, our methodology effectively reduces this number by more than 50 %. This approach exhibits an accuracy rate of approximately 92 %. The resulting geolocated dataset of companies' patent transactions offers a valuable resource for fine-grained geographical analyses of the markets for technology; in particular, we provide examples of relevant economic insights which can be learned from looking at the geographical patterns of those transactions.

1. Introduction

Patent transactions, which involve the buying, selling, licensing, or transfer of patent rights, facilitate the transfer of technology and knowledge between different entities, allowing for the dissemination and commercialization of inventions [1]. They enable inventors and organizations to monetize their inventions by licensing or selling their patent rights to other parties, leading to the development of new products, processes, and technologies, driving economic growth and competitiveness [2]. Despite several researchers having noted the importance of utilizing patent assignment data for understanding technology transfer (i.e. [3–5]), research using this kind of data is surprisingly sparse in previous literature and, especially for what concerns the literature on patent transactions in the US, one of the reasons for that is that the available data is not in a format amenable for use by researchers [6–8]. In particular, the way geographical factors affect technological flows and knowledge diffusion has been under-investigated, except for few works [7,9,10]. Adding geographic information to patent transactions could unveil a multitude of insights. It could assist in identifying innovation hubs, recognizing knowledge importers and exporters, and revealing patterns of collaborations across disparate areas and industries. Moreover, the presence of geography-specific sellers, but not buyers, may suggest barriers to commercialization or an inability to leverage the patents effectively.

As such, we see this gap as an opportunity to advance our understanding of the redistribution and value of intellectual property, as well as market dynamics and linking them to geographical patterns. Therefore, in this research we propose a novel methodology which is aimed at harmonizing and geocoding the locations which are related to patent transactions. In particular, we propose a 4 stages multi-layered method which encompasses natural language processing (NLP) and exploits freely available web knowledge graphs. The obtained results, which refer to locations linked to USPTO patent transactions from 2005 to 2022, are encouraging: the proposed method reduces the number of locations by more than 50 %. The validity of these results is confirmed by a manually labeled sample consisting of 5 % of the total locations, which proves an accuracy rate of around 98 % for the present procedure. The rest of the paper is organized as follows: Section 2 presents the background, Section 3 describes the data features, the proposed methodology in detail and its validation procedure, Section 4 presents some insights about patent transactions using the harmonized locations and Section 5 concludes.

* Corresponding author. Bordeaux School of Economics UMR CNRS 6060, University of Bordeaux, 16 avenue Léon Duguit, 33608, Pessac, France.
E-mail addresses: graziasveva.ascione@uninsubria.it (G.S. Ascione), andrea.vezzulli@uninsubria.it (A. Vezzulli).

https://doi.org/10.1016/j.wpi.2024.102320
Received 1 December 2023; Received in revised form 29 October 2024; Accepted 30 October 2024
Available online 7 November 2024

2. Background

2.1. The geography of US patent transactions

The increased number of patents is linked to the recent expansion of the markets for technology, in which patents are increasingly conceived as tradable assets [3,5]. Transactions involving technology packages (patents, patent licensing, and other intellectual property and know-how) can entail knowledge exchange between transacting agents.

There is a wide literature that argues that successful technology transfer relies to some extent on tacit knowledge [6]. Communication of tacit knowledge over long distances is likely to be problematic, since personal interaction is essential to transfer this type of knowledge effectively, especially across firms' boundaries. As a result, potential buyers' ability to evaluate and utilize external patented inventions may be largely dependent on the buyer's research activities and, thus, on their prior knowledge stocks located in the geographical area of the external invention. Therefore, in order to understand the potential for external redeployment of a new invention towards a potential buyer, it is important to assess whether the potential buyer's prior knowledge stocks are also geographically located near the location of the patented external invention. Consistently, it has been proven that geography plays an important role in patent transactions [8–10].

However, because of the difficulty in obtaining clean patent transaction data [7,9], only few papers consider the geographic dimension of US patent trade. [8] examine whether geographic and time factors shape patent transactions across states and sectors in the US and compare their findings with those of citation flows. Their results support that geographic nearness, in terms of distance and contiguity, also matters for patent trade. Further, they confirm that the knowledge generated from innovative ideas, which are patented and traded, is more geographically restricted and, therefore, its effective reach is less far stretched in space compared to knowledge flows based on citations. [10] use US firm data to examine whether patents sold during the application phase are less likely to be sold outside the seller's state than patents that are sold after they are issued: they find that the latter are more likely to be sold outside the state's borders, compared to patent pending applications. Therefore, they claim that patents play a role in mitigating the geographic distance in the market for ideas. [9] study several dimension of patent trades, including whether the geographical proximity of the initial innovator's knowledge stocks matters in the decision to sell a patent; they find that the initial innovator's prior knowledge stocks, that allow innovators to capture value from their own inventions within the firm (as opposed to transferring their ownership to others) are highly co-localized geographically and characterized by technological proximity.

2.2. Correctly geolocating patent assignees

Harmonizing the geographical information about patents holders has been considered an important issue in the literature [11]. In particular, patent assignees at the USPTO are deemed to have addresses that are difficult to geolocate [12]. The first attempts of harmonizing patent related locations date back to the 90s. Examples are the early efforts by Refs. [11,13] and more recent attempts by Refs. [12,14,15]. [11] propose a methodology based on matching scripts that allows allocating the majority of the patentee and inventor addresses of EU-27 Member States to their respective NUTS 2 regions. [12] use highly specific geolocation data to disambiguate assignees and inventors' names. [14] use information available from the patent office registers on the address of patentees to geocode assignees and inventors' locations all over the world since the 1980s. It is not until very recently that experts started leveraging NLP to harmonize patent assignees' locations. In this direction, [15], presented PatentCity, a novel dataset on the location and nature of patentees from the 19th century using information extracted through a combination of NLP techniques (such as NER) for addresses extraction and Google API services for the geocoding part.[1]

3. Data and methodology

3.1. Data and its ambiguities

We make use of a comprehensive dataset covering all granted patents transacted by companies from 2005 to 2022. The source is the USPTO Patent Assignment Database (PAD).[2] The database records, for each transaction, the name of the buyer (i.e. assignee) and of the seller (i.e. assignor), the date at which the assignment was recorded at the USPTO, the date at which the private agreement between the parties was signed, the associated patent number and the location of the buyer at the moment of the transaction. The database also reports the rationale behind the transfer of the patent (i.e. conveyance type). This field is useful for data cleaning because it allows changes in ownership (recorded as "assignment of assignors' interest") to be distinguished from other administrative events (e.g. the union of commercial interests as a "merger", the securitization of a patent as collateral for a "security interest/agreement", the change of name or address of its current owner as a "change of name/address", the corrections of previous mistakes as a "corrective assignment"). For the purpose of this research, we only select transactions where the conveyance type is assignment of assignors' interest, merger and government interest. The total number of patents filed since 2005 is 3′772′221, of which 3′284′931 are involved in at least one transaction. Further, if we consider transactions in which only companies are involved, we have 3′280′163 transactions, which are linked to 63′838 distinct locations.

However, the location information presents significant challenges for reliable analysis and interpretation. These inconsistencies stem from various sources, including data entry errors, differences in naming conventions, and other forms of ambiguity. These inconsistencies can compromise the validity of research findings and lead to misleading or erroneous conclusions if not adequately addressed. Therefore, it is crucial to engage in a systematic approach to clean and standardize these locations before further analysis. In particular, the following list outlines some specific issues that impede the effective geolocation of patent buyers' locations in the aforementioned data.

1. Spacing mistakes and added characters/words: in the original data the spaces are not harmonized and it is quite common to see extra characters added to the location. For instance, in the data we find "BERGAMO, IT", as well as " I BERGAMO IT", which can be both linked to the city of Bergamo, in northern Italy. We also find "F PARIS FR" which refers to the city of Paris, in France and "D BERLIN DE" which points at Berlin. Those first tokens seem to indicate the initial character of the country in the local language. Further, it is possible that entire words are added, such as in "SCOPE COMPLEX LODHI ROAD IN" which points to the location Lodhi Road, in New Delhi, India or "URBANIZACION OBARRIO PA", which refers to the neighborhood of Obarrio in Panamà.
2. Province code or other administrative codes: in some cases, province or administrative codes are added to the name of the city or town for administrative or postal purposes. For instance, we find "ROMA RM IT", which refers to the city of Rome, in Italy or "PARIS EME FR", where "EME" refers to the V arrondissement of the city.
3. Specific district information: in some cases, very specific district information is added to the city information, which, if not standardized, might lead to consider two cities as different locations only because the district information is different. For instance, we find

[1] More information of Google geolocation services is available here: https://developers.google.com/maps/documentation/geolocation/overview.

[2] The database is freely available at https://www.uspto.gov/ip-policy/economic-research/research-datasets/patent-assignment-dataset.

"PARIS L A DEFENSE FR", where la Défense refers to a specific neighborhood in the city of Paris or "BERLIN MITTE DE" where Mitte in a neighborhood of the city of Berlin, Germany.
4. Spelling mistakes: it is very common in the data to find spelling mistakes referring to city, state or countries. Some examples include "BERLINE DE", referring to Berlin, "OSLOW NO" referring to Oslo, Norway, "COPENHAGE DK DK", referring to Copenhagen in Denmark.
5. Same location in different languages: some geographical entities are also present both in English and in their original language, therefore with different spellings. For instance, we can find "KOBENHAVEN N DK" referring to Copenhagen, "WIEN A AT AT", referring to Vienna, Austria and "FIRENZE IT" referring to Florence, Italy.

3.2. Methodology

3.2.1. Stage 1: preprocessing

We start from the "assignee" table from PAD[3] where we have 3′280′163 transaction records of interest involving 310′285 assignees identified as companies in the 2005–2022 time period. The assignee table has four columns: *ref_id*, which reports the number of transaction, *ee_name*, which reports the name of the assignee at a given date, *ee_address1* and *ee_address2* which report the principal and a second address of the assignee in that specific transaction, where present, and then the columns *ee_city*, *ee_state*, *ee_postcode* and *ee_country* which respectively report city, state, postcode and country related to a certain assignee as reported in a certain transaction.[4] As the first step, following [7], we input "US" in the country column when it is empty, but the state column is not. Considering that our goal is to harmonize locations and not to correct addresses, we focus on the city, state and country columns.

Then we create a "location" column which is composed by city, state (where present) and country. We then discard locations shorter than 6 characters, as they are likely to be not accurate enough for geo-localization purposes[5] and we obtain a list of 63′838 unique distinct locations.

The second step of the preprocessing phase is to convert non-ASCII characters (including Chinese and Arabic characters) to interpretable characters for the following semantic clustering phase. Then, each location is fully harmonized with basic operations such as upcasing and space harmonization. After preprocessing, we are left with 55′753 unique locations which are fed to the semantic clustering phase.

3.2.2. Stage 2: semantic clustering

In the preprocessing phase we corrected spacing and other formatting differences across the locations. However, we still have to group locations which are likely to be referring to the same geographical entity and providing their correct form, in order to be able to geolocate them. The first objective is tackled in this step defined as *semantic clustering*. For instance, in our data, "A FUERSTENFELD AT" and "A FURSTENFELD AT" refer to the same location Fürstenfeld in Austria as well as "ABENO KU OSAKA SHI OSAKA JP" and "ABENO KU OSAKA SHI OSASKA JP″ both refer to Abeno-ku, one of the 24 neighborhoods in Osaka, Japan. These kinds of typos can be easily tackled with a string similarity approach, such as the Levenshtein distance. Levenshtein distance (LD) is a measure of the similarity between two strings, the source string (i) and the target string (j) [16]. The distance is the number of deletions, insertions, or substitutions required to transform s into t. The greater the Levenshtein distance, the more different the strings are. Levenshtein distance can be defined as follows:

$$D\begin{pmatrix} \\ i,j \\ \\ \end{pmatrix}\begin{cases} 0 \ \text{if } i=0 \text{ and } j=0, \\ i \quad \text{if } j=0 \text{ and } i>0, \\ j \quad \text{if } i=0 \text{ and } j>0, \\ \min\begin{cases} D(i-1,j)+1, \\ \mathrm{D}(i,j-1)+1, \\ D(i-1,j-1)+\mathrm{sub}(A[i],B[j]) \end{cases} \text{otherwise.} \end{cases} \tag{1}$$

However, we opted not to employ Levenshtein distance calculations for all possible pairwise combinations of elements in the dataset. This decision was driven by considerations of computational efficiency and resource utilization.

Calculating the Levenshtein distance for every possible pair in a large dataset would entail a computational complexity of $O(n^2)$, where n is the number of elements. This quadratic complexity would lead to an exponential increase in both time and computational resources required as the dataset grows.

Moreover, the Levenshtein algorithm itself has a time complexity of $O(m \times p)$, where m and p are the lengths of the two strings being compared. Therefore, the combined computational cost would be prohibitively high for large datasets, making it an impractical choice for our specific research context.

Thus, to maintain the scalability of our study and to allocate computational resources more parsimoniously, we followed the approach exemplified in the following metacode.

1. Import the required libraries and functions.
2. Define a function cluster_strings that takes a DataFrame, a column name, and a threshold as its parameters.
3. Inside the function:
   (a) Alphabetically sort the DataFrame based on the specified column.
   (b) Initialize variables: a numeric variable *cluster* to an initial value 0, a string variable *prev_string* to the first string in the DataFrame, and a numerical vector *clusters* to a list of zeros with the same length as the DataFrame.
   (c) Loop through each row of the DataFrame, starting from the second row:
      i. Compute the Levenshtein distance between the string in the current row of the DataFrame and the previous string *prev_string*.
      ii. If the computed distance is greater than the specified threshold,[6] increment the *cluster* variable by 1.
      iii. Assign the current *cluster* value to the current row in the *clusters* list.
      iv. Update the *prev_string* variable with the current string.
   (d) Add the *clusters* list as a new column in the DataFrame.
4. Return the updated DataFrame.

Through this approach, we achieve a linear time complexity closer to $O(n)$, thereby significantly reducing computational costs while still leveraging the utility of the Levenshtein distance metric. This compromise provides a more favorable balance between accuracy and efficiency, thereby making our research more scalable and computationally feasible.

Following the completion of this process, we are left with a total of 50′754 unique locations. It is important to note, however, that semantic

[3] To download the original assignee table from PAD refer to the link in the previous footnote.

[4] However, it is worth noting the content of each of these fields is not homogeneous among different locations. For instance, for some records the city field is populated with the fraction or the district too.

[5] The majority of locations shorter than 6 characters consist of states and countries with short names. Examples include cases such as "MA US", which might point to the state of Massachusetts in the United States, but we prefer, due to the uncertainty, to leave out this kind of cases from the analysis.

[6] In particular, for this specific set of data the selected threshold was 2.

clustering alone is unable for determining the correct version of a given location name. This limitation is subsequently addressed in the succeeding phase, utilizing web knowledge graphs.

3.2.3. Stage 3: web-knowledge graphs

The third step of this methodology, defined as *Web Knowledge Graph*, is focused on rectifying the grammatical errors and typos present in the location names; these inaccuracies in the data not only hinder the process of harmonization but also, and perhaps more critically, they introduce a notable level of complexity in the subsequent geolocation retrieval process. When subjected to API-based geolocation services, these misreported location names are frequently unresolvable to specific geographic coordinates. This is because API systems typically rely on pre-defined gazetteers or geographic databases that are unable to handle deviations from standardized naming conventions. Consequently, these inaccuracies inhibit the system's ability to assign accurate latitude and longitude values, ultimately compromising the quality and usability of the spatial data.

To automate the correction of these errors, we use the *DuckDuckGo* search engine[7] which, among its features, uses a Web Knowledge Graph as result for geographically related queries.[8]

The Web Knowledge Graph is a technology used by search engines to enhance the user experience by providing a summary of the most relevant information regarding a search query.

This information is typically extracted from a variety of sources across the internet and presented in a structured format, often at the top of the search results [17]. The use of Web Knowledge Graphs is well established in many data mining tasks [18,19], improving also the performance of many intelligent systems [20].

Notably, and quite relevantly in our case, for location-based queries, the Web Knowledge Graph often includes the correctly spelled and formatted name of the location. To retrieve the relevant information contained in the Web Knowledge Graph, we created a web crawler which automatically enters each location name into the DuckDuckGo search engine and then systematically retrieves the corrected location name from the Web Knowledge Graph, if present. This step is crucial as it leverages the vast amount of data available on the internet, accessed via DuckDuckGo, to correct and harmonize the location names. Utilizing the Web Knowledge Graph not only helps in rectifying typos and grammatical mistakes but also aids in standardizing the names, as the names retrieved from the Web Knowledge Graph are in a format that is widely accepted and used. This technique is very effective at tackling very ambiguous locations, which might present not only spelling mistakes but also might involve omissions and a multi-language scenario. For instance, the location "KOBENHAVN DK" is automatically converted to the English "Copenaghen" and the Italian location "SAMARATE FRAZIONE CASCINA COSTA IT", which can be translated as "Samarate, township of Cascina Costa Italy", is automatically reduced to "Samarate". The name "CALGARY T H A ALBERTA CA" is linked to "Calgary". This system is proved to be working with different languages too: for instance, "SOK NO CAYYOLU ANKARA TR" which is in Turkish language where "SOK" is Abbreviation for "Sokak," which means "Street" in Turkish, "NO" refers instead to the number, and CAYYOLU is a neighborhood in Ankara, is correctly linked to Çayyolu. After this step the unique locations are reduced to 33′851.[9]

3.2.4. Stage 4: API harmonization

The final step is to geolocate -namely converting addresses into coordinates-those locations. To do this we exploit the Nominatim API service, which, for each location name inputted, returns a corrected, fully specified location name with latitude and longitude. Nominatim is a search engine for the OpenStreetMap (OSM) database, providing geolocation data for most of the places around the world. The Application Programming Interface (API) service of Nominatim allows users to access this geolocation data programmatically using different programming languages, including Python.[10]

The Nominatim API allows users to perform geocoding and reverse geocoding (converting coordinates into addresses). It is free to use under the Open Database License (ODbL), which promotes the open sharing of data. However, users must abide by the Nominatim usage policy, which includes restrictions on heavy traffic to prevent overloading the servers.

The Nominatim API has been accessed through the *geopy* library,[11] which is a client for several popular geocoding web services, including Nominatim. This library simplifies the process of making requests to the Nominatim API, handling the HTTP requests and responses for the user. Using the API, in this step we obtain for each unique location its coordinates, as well as information about the city, county and state (where present) and country related to each location. After this step, we have 31′321 locations each with latitude and longitude as well as the city, county, state and country related to each location.

3.3. Validation

To assess the quality of the final results, we adopted a validation process using a gold standard criterion. The creation of the gold standard started with the selection of a random sample, constituting 5 % (3′192 locations) of the original location list retrieved from PAD. Each location in this sample was manually assigned a correct geographical designation. We then utilized the Nominatim API to fetch the associated polygon data for each location. Within this context, a "polygon" refers to a series of connected points that form a closed geometric figure, defining the boundaries of geographical entities such as cities, states, or countries. This polygon information is crucial for validating the accuracy of our geolocalization.

We successfully retrieved polygon data for 87 % of the locations. However, for 13 % of the cases, the Nominatim API provided data corresponding to single points or a sequence of points, such as in addresses "Rue de la Couture, Rungis, France", instead of polygons. This discrepancy arises when the API lacks comprehensive polygon data for certain locations or when the location corresponds to linear features like streets. Our gold standard is, therefore, a labeled dataset of 3′192 locations which are assigned to a polygon or a point.

The primary objective of our methodology is not merely to harmonize locations, but also to ensure their precise geolocalization. Hence,

[7] DuckDuckGo is a search engine that distinguishes itself through a user-centric focus on privacy and anonymity. Unlike more conventional search engines, which track user behavior for advertising and data analytics, DuckDuckGo does not collect personal information. It employs a combination of its own web crawler and APIs from various services to generate search results, aiming to deliver unbiased information without the filter bubbles often associated with personalized search algorithms.

[8] In the context of our study, we opted to employ DuckDuckGo as the search engine for data retrieval. This decision was predicated on the engine's consistent result presentation across different users for identical queries, a feature not universally true for other mainstream search engines such as Google. This uniformity in search results is crucial for the replicability and generalizability of our findings. While not academically rigorous, some anecdotal comparisons highlighting these differences can be found at https://www.bluehost.in/tutorials/5-interesting-things-that-duckduckgo-can-do-and-google-cant.

[9] However, in 538 cases the search engine does not return any result for the web knowledge graph. This is possible in cases in which the locations provide to the browser contradictory or highly unclear information, in cases such as "SHANGHAI CA", "I SOVIGLIANA VINCI IL", or "PERSIARAN GURNEY PENANG MY PERSIARAN GURNEY".

[10] Python is a popular choice for accessing the Nominatim API due to its user-friendly syntax and wide range of libraries that make it easier to work with HTTP requests and JSON data, which are essential for interacting with APIs.

[11] The geopy documentation is available at the following url https://geopy.readthedocs.io/en/stable/.

we consider as correctly geolocalized the cases in which the latitude and longitude of the resulting locations fall within the polygon, with a tolerance of approximately 10 km (0.0901°) from the polygon borders or point. This tolerance threshold, while maintaining strict accuracy requirements, does not introduce significant bias into our evaluation mechanism, given our goal of achieving precise geolocalization at the city level or higher. Consequently, we selected accuracy, defined as the share of correctly predicted locations, as our metric for validation, according to the following formula:

$$accuracy(y, \hat{y}) = \frac{1}{n\ samples} \sum_{i=0}^{n\ samples-1} 1(\hat{y}_i = y_i) \quad (2)$$

where *n samples* represents the total number of locations that are part of our gold standard (3′192), $\hat{y}_i$ is the predicted value of the i-th location and y_i is the corresponding true value.[12] Further, we define accuracy in terms of the granularity of the geolocation. For instance, if a location is identified as a city, the geolocation accuracy is verified at the city level, and similarly, state-level accuracy is expected for state level. This approach ensures that each geolocation precisely corresponds to its designated geographical entity.

Following this evaluation mechanism, our methodology achieves an accuracy score of 92 %. Further, we assess the impact of each phase in our methodology on the overall accuracy score, by systematically analyzing the three key stages: preprocessing, semantic clustering, and the Web Knowledge Graph. This approach allows us to discern the individual contributions of each stage to the accuracy.

Our analysis revealed that omitting the preprocessing step results in a lower accuracy of 88 %. This demonstrates that, while preprocessing plays a significant role in enhancing overall accuracy, it is not solely decisive. On the other hand, eliminating semantic clustering does not affect the accuracy significantly, which stays at 92 %. However, we consider this step crucial for the scalability of our methodology. Semantic clustering significantly reduces the volume of unique names that need to be processed in the subsequent Web Knowledge Graph phase, thereby decreasing the time required for this computing intensive step.

The most critical finding was the impact of bypassing the Web Knowledge Graph correction. Omitting this step led to a substantial drop in accuracy, down to 72 %. This stark reduction underscores the importance of correcting spelling mistakes and standardizing names to successfully geolocate addresses. The Web Knowledge Graph phase is evidently pivotal in ensuring high accuracy and reliability of the geolocalization results, confirming its essential role in our methodology (see Fig. 1).

4. Results and discussion

Having clean data on transactions allows us to analyze the geographical patterns related to US patent trade at different scales. In particular, we focus on firm-to-firm reassignments, leaving out from the analysis assignments made from employee-inventors to the firms that hire them. We take this decision as we are interested in the geography of knowledge flows between companies. Fig. 2 represents those different levels: it plots the number of patents, divided in 7 (Fig. 2a) or 5 bins (Fig. 2b and c) transacted among the most active geographical entities; on the x axis we represent the most active patent buyers, while on the y axis we represent the most active patent sellers.[13] In particular, Fig. 2a represents the most active countries in terms of transactions. Consistently with the literature, our results clearly support strong localization of patent transaction flows as countries tend to involve more in exchanging patents within their borders than with other states or countries [10]; it is also important to notice, the United States and Japan are the two most active countries. Further, there is an important flow of patents towards the Cayman Islands, mostly from the United States. This might be explained by a variety of economic, legal, and strategic factors, including tax benefits and a strong asset protection policy. Also, for other jurisdictions, such as China, as local companies seek global expansion, optimizing IP strategy through favorable jurisdictions like the Cayman Islands becomes crucial.

Fig. 2b represents instead the transaction among different US states; from the Figure we see that the most relevant states in terms of transactions are the state of New York, California, New Jersey, Texas and Washington. As in the previous Figure, also in Fig. 2b it is clear that the greatest number of transactions take place within states borders; however, some exceptions include patents sold from California to Texas and vice-versa and from New York to California.

Fig. 2c looks at transactions involving cities at world-wide level. The most important cities are all located within the US, including Waukegan (Illinois), Mountain View and Saint José (both in California). When looking among transactions between different cities, there is a notable flow of patents from Dallas (Texas) to Saint José, and from Greesnboro (North Carolina) and Fremont (California) towards Torrance (California)

5. Final remarks

As underscored by existing literature, patent transaction data has often been underutilized because of the complexity of their format and the need for *a priori* accurate data cleaning and preparation [7–9]. This study addresses this challenge by introducing a novel methodology for the harmonization and standardization of location data associated with such transactions. Specifically, we focus on standardizing the locations for 310′285 patent assignees linked to 3′280′163 patent transactions filed at the USPTO between 2005 and 2022. The methodology yields promising results, obtaining a more than 50 % reduction in the number of unique locations. Notably, the "Web Knowledge Graph" phase emerges as the most impactful, accounting for a 34 % decrease from the preceding semantic clustering stage and having the greatest impact on the accuracy of the methodology itself.

The findings elaborated in Section 4 underscore the pivotal role of geo-referencing patent transactions to gain nuanced insights into the spatial dynamics of innovation. Most compelling is the observation that the majority of transactions, whether at the city, state, or country level, predominantly occur within the same geographical entity. This result is in line with prior academic work [8].

Nevertheless, the methodology here presented does not come without limitations. For instance, ambiguities in some location attributes—such as inconsistencies between city, state, and country information—remain unresolved. Further, it should be noted that the methodology is designed to handle locations with a length of five tokens or more. While the results are extensive, they are not exhaustive of all transactions occurring between 2005 and 2022. Some transactions may be missing from our dataset, as they might not have been reported to the USPTO, whether unintentionally or as part of a strategic decision. In addition, the complexity is further compounded by transactions that involve sellers operating under different names or through subsidiaries,

[12] The reported definition of accuracy is the one proposed by the Python package sk-learn. Further detail is available at the following url: https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score.

[13] Because many recorded assignments represent transactions between inventors-employers and their employees-assignees as of the grant date of the patent, we identify their names and exclude these assignments (i.e., first assignments) from the plots.

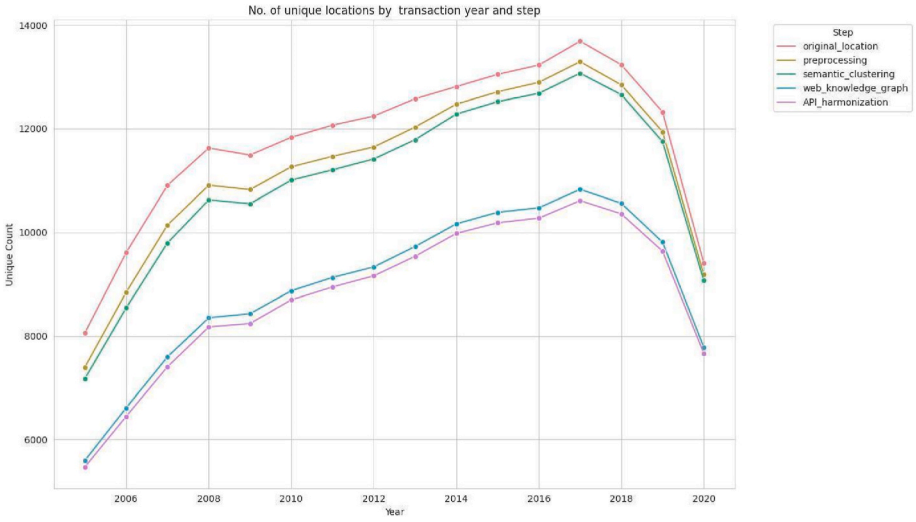
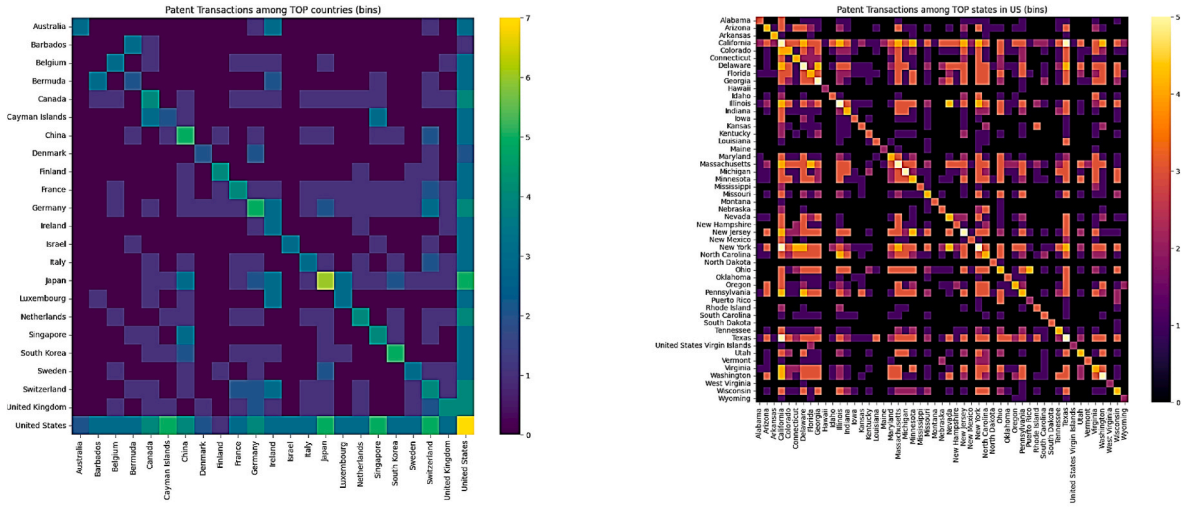
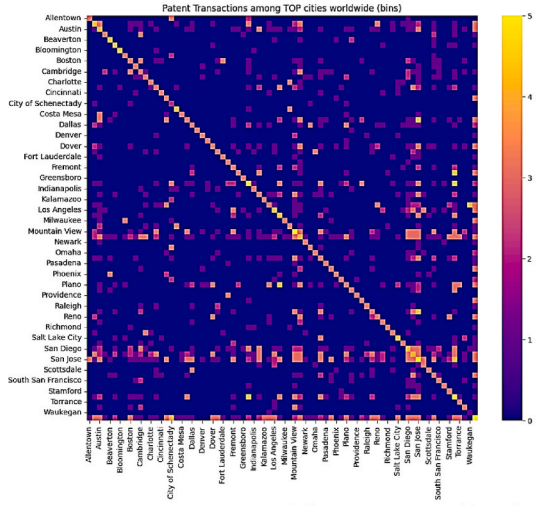


Fig. 1. No. of unique locations by transaction year and methodological step.



(a) Transactions among different countries

b) Transactions among different US states



(c) Transactions among different cities world-wide

Fig. 2. Patent transactions from 2005 to 2022 at different geographical scales.

as these could actually be considered as singular entities operating under multiple aliases. In future research, this problem should be tackled by combining our method with other algorithms that match company names with their subsidiaries, such as the one recently proposed by Ref. [21]. This would also help in identifying firms who may transact patents using a specific subsidiary. Moreover, new research could also expand the scope of the current analysis considering also the cases when the inventors assign the patent to the companies that hire them. This would further expand the present understanding of technological diffusion patterns and geography of innovation.

Finally, we envision this study as an initial step toward a more robust examination of the geographical aspects of patent transactions, facilitated by cleaner and more reliable data. The methodology allows for the integration of additional dimensions into geolocated transaction data, including the characteristics of both patents and assignees. In addition, these transaction data in the future could be linked to other kinds of information, such as the size of buyer and sellers, tax incentives and transportation infrastructure. Areas with dense patent transactions might witness economic spillovers in the form of job creation, increased demand in housing, or the development of auxiliary businesses.

Therefore, we believe that enriching transaction data with geographic information through our methodology could serve as a fertile ground for revisiting traditional research questions—such as the underlying factors driving patent transactions—as well as for pioneering new avenues of inquiry.

CRediT authorship contribution statement

Grazia Sveva Ascione: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andrea Vezzulli:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Andrea Vezzulli reports a relationship with NODES Scarl that includes: non-financial support. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This publication is part of the project NODES which has received funding from the MUR – M4C2 1.5 of PNRR funded by the European Union - NextGenerationEU, Mission 4 Component 1.5 - ECS00000036 - CUP J83B22000050001. Further, Grazia Sveva Ascione thanks the *Centre national de la recherche scientifique* (CNRS) who supported her work through the CNRS Prematuration (IAPLB) project. Andrea Vezzulli gratefully acknowledges funding from the InsIDE Lab and from the MUR (Italian Ministry of University and Research) Grant 'Department of Excellence 2023–2027' (Project no. CUP J37G22000330001).

References

[1] Ioan C. Cucoranu, Anil V. Parwani, Suryanarayana Vepa, Ronald S. Weinstein, Liron Pantanowitz, Digital pathology: a systematic evaluation of the patent landscape, J. Pathol. Inf. 5 (1) (2014) 16.
[2] Sercan Ozcan, Nazrul Islam, Patent information retrieval: approaching a method and analysing nanotechnology patent collaborations, Scientometrics 111 (2017) 941–970.
[3] Ashish Arora, Andrea Fosfuri, Alfonso Gambardella, Markets for technology and their implications for corporate strategy, Ind. Corp. Change 10 (2) (2001) 419–451.
[4] Marco Ceccagnoli, Stuart JH. Graham, Matthew J. Higgins, Jeongsik Lee, Productivity and the role of complementary assets in firms' demand for technology innovations, Ind. Corp. Change 19 (3) (2010) 839–869.
[5] Mario Benassi, Alberto Di Minin, Playing in between: patent brokers in markets for technology, R&d Management 39 (1) (2009) 68–86.
[6] Per Botolf Maurseth, Roger Svensson, The importance of tacit knowledge: dynamic inventor activity in the commercialization phase, Res. Pol. 49 (7) (2020) 104012.
[7] Stuart JH. Graham, Alan C. Marco, Amanda F. Myers, Patent transactions in the marketplace: lessons from the uspto patent assignment dataset, J. Econ. Manag. Strat. 27 (3) (2018) 343–371.
[8] Kyriakos Drivas, Irene Fafaliou, Elpiniki Fampiou, Demetrius Yannelis, The effect of patent grant on the geographic reach of patent trade, J. High Technol. Manag. Res. 26 (1) (2015) 58–65.
[9] Nicolás Figueroa, Carlos J. Serrano, Patent trading flows of small and large firms, Res. Pol. 48 (7) (2019) 1601–1616.
[10] Kyriakos Drivas, Claire Economidou, Is geographic nearness important for trading ideas? evidence from the us, J. Technol. Tran. 40 (2015) 629–662.
[11] Julie Callaert, Mariette Du Plessis, J. Growels, Christophe Lecocq, Tom Magerman, Bruno Peeters, Xiaoyan Song, Bart Van Looy, Charlotte Vereyen, Patent statistics at eurostat: Methods for regionalisation, sector allocation and name harmonisation. Eurostat Methodologies and Working Papers, 2011.
[12] Greg Morrison, Massimo Riccaboni, Fabio Pammolli, Disambiguation of patent inventors and assignees using high-resolution geolocation data, Sci. Data 4 (1) (2017) 1–21.
[13] Naomi R. Lamoreaux, Kenneth L. Sokoloff, Inventors, firms, and the market for technology in the late nineteenth and early twentieth centuries, in: Learning by Doing in Markets, Firms, and Countries, University of Chicago Press, 1999, pp. 19–60.
[14] Gaétan De Rassenfosse, Jan Kozak, Florian Seliger, Geocoding of worldwide patent data, Sci. Data 6 (1) (2019) 260.
[15] Antonin Bergeaud, Cyril Verluise, A New Dataset to Study a Century of Innovation in Europe and in the US, 2022.
[16] Vladimir I. Levenshtein, et al., Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet Physics Doklady, vol. 10, Soviet Union, 1966, pp. 707–710.
[17] Gianluca Demartini, A tutorial on leveraging knowledge graphs for web search, in: Information Retrieval: 9th Russian Summer School, RuSSIR 2015, Saint Petersburg, Russia, August 24-28, 2015, Revised Selected Papers 9, 2016, pp. 24–37.
[18] Nicolas Heist, Sven Hertling, Daniel Ringler, Heiko Paulheim, Knowledge graphs on the web-an overview. Knowledge Graphs for eXplainable Artificial Intelligence, 2020, pp. 3–22.
[19] Evgeniy Gabrilovich, Nicolas Usunier, Constructing and mining web-scale knowledge graphs, in: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2016, pp. 1195–1197.
[20] Antoine Bordes, Evgeniy Gabrilovich, Constructing and mining web-scale knowledge graphs: kdd 2014 tutorial, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, 1967–1967.
[21] Grazia Sveva Ascione and Valerio Sterzi. Presenting Terrorizer: an algorithm to harmonize assignee names. arXiv preprint arXiv:2403.12083.

Grazia Sveva Ascione holds a PhD in Economics of Innovation from the University of Turin. Grazia Sveva works in the area of research related to Intellectual Property, patent analytics and natural language processing (NLP). She is currently a postdoctoral researcher in Economics and Data Science at the University of Insubria.

Andrea Vezzulli holds a PhD in Economics from the University of Milan. He is associate professor of applied economics and coordinator of the PhD Program in Methods and Models for Economic Decisions at the University of Insubria, Department of Economics. He is also a research affiliate at the Invernizzi Center for Research on Innovation, Organization, Strategy and Entrepreneurship (ICRIOS), Bocconi University and at the Responsible Management Research Center (REMARC), University of Pisa. His research interests focus on innovation, IPRs, knowledge diffusion and small business finance.