



**HAL**  
open science

# The Geography of Retracted Papers: Showcasing a Crossref–Dimensions–NETSCITY Pipeline for the Spatial Analysis of Bibliographic Data

Guillaume Cabanac, Alexandre Clause, Laurent Jégou, Marion Maisonobe

## ► To cite this version:

Guillaume Cabanac, Alexandre Clause, Laurent Jégou, Marion Maisonobe. The Geography of Retracted Papers: Showcasing a Crossref–Dimensions–NETSCITY Pipeline for the Spatial Analysis of Bibliographic Data. STI 2023: 27th International Conference on Science, Technology and Innovation Indicators, CWTS, Leiden University, Sep 2023, Leiden, Netherlands. hal-04731543

**HAL Id: hal-04731543**

**<https://hal.science/hal-04731543v1>**

Submitted on 10 Oct 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# The Geography of Retracted Papers: Showcasing a Crossref–Dimensions–NETSCITY Pipeline for the Spatial Analysis of Bibliographic Data

Guillaume Cabanac<sup>\*,\*\*</sup>, Alexandre Clausse<sup>\*</sup>, Laurent Jégou<sup>\*\*\*</sup>, and Marion Maisonobe<sup>\*\*\*\*</sup>

<sup>\*</sup>*guillaume.cabanac@univ-tlse3.fr; alexandre.clausse@irit.fr*  
0000-0003-3060-6241; 0009-0004-7215-6247  
IRIT UMR 5505 CNRS, Université Toulouse 3 – Paul Sabatier, France

<sup>\*\*</sup>Institut universitaire de France (IUF), Paris

<sup>\*\*\*</sup>*laurent.jegou@univ-tlse2.fr*  
0000-0003-4304-679X  
LISST-Cieu UMR 5193 CNRS, Université Toulouse-Jean Jaurès, France

<sup>\*\*\*\*</sup>*marion.maisonobe@cnrs.fr*  
0000-0002-2968-9038  
Géographie-cités, CNRS – EHESS - Université Paris 1 - Université Paris Cité, France

A surge of retractions has been observed during the last two decades. Last year, flagship publishers even retracted hundreds of journal articles in bulk on the grounds of research misconduct. Press releases list the manipulated journals, point to offending paper mills but seldom comment on the places producing unreliable papers. We present two contributions in this paper. First, we introduce a Crossref–Dimensions–NETSCITY workflow that enables the geographic exploration of any bibliographic dataset. Second, we apply this workflow to the corpus of all 12k retracted publications extracted from Crossref (open data) and enriched with affiliation data (provided for free) exported from Dimensions. The analysis of the geographic distribution of retracted papers reveals a concentration in Asia that is increasing over time.

## Abstract

## 1. Introduction

Global scientific output measured by the number of peer-reviewed publications is growing annually. Many sources reflect this growth and the observation holds true regardless of the source used (Web of Science, Scopus, Dimensions). According to Fanelli & Larivière (2016), this growth is less related to the increase in individual productivity than it is to the increase in the overall demography of scholars authoring in international journals. Over the last 50 years, an increase in the number of researchers and lecturers occurred in a large number of countries; and this demographic growth has been in conjunction with the spatial deconcentration of world scientific production, i.e. the emergence of new production sites such as universities and research centres (Grossetti *et al.*, 2014; Maisonobe *et al.*, 2018).

In this globalised context, strong incentives to publish have developed at the level of institutions and national research evaluation systems; sometimes fueled by the desire or necessity to appear in international rankings. This quest for visibility leads to a race to publish (especially when publication practices are rewarded with cash for the authors), which explains the recourse to predatory journals or so-called “paper mills” (Hvistendahl, 2013; Else & Van Noorden, 2021).

As a result, the number of retractions increases (Oransky, 2022a), to correct errors or whole papers, especially the forged/computer-generated publications resulting from widespread fraudulent practices, which are then the subject of bulk retraction campaigns (see, e.g., Oransky, 2021; Oransky, 2022b). These retractions *en masse* stem from the increased vigilance of scientists and hobbyists determined to act against scientific misconduct. Open post-publication peer review platforms such as PubPeer (Barbour & Stell, 2020) are now frequently used to report questionable scientific content, which can then lead to retractions by journal editors or the authors themselves.

Both the global growth of publications and the effects of scientific misconduct can explain that we are seeing an increase in article retractions as recorded in the Retraction Database (Oransky, 2022a). However, the literature on scientific misconduct does not categorically conclude that a greater pressure to publish leads to more errors and fraud (Fanelli *et al.*, 2015). According to Fanelli *et al.*: “scientific misconduct is more likely in countries that lack research integrity policies, in countries where individual publication performance is rewarded with cash, in cultures and situations where mutual criticism is hampered, and in the earliest phases of a researcher’s career”. Drawing upon this, we assume that the rate of retracted publications is not evenly distributed across the globe.

This article introduces a method to analyse the geographical distribution of retracted articles at a fine spatial resolution: the level of urban areas.

The Crossref–Dimensions–NETSCITY pipeline that we use enables performing geographic bibliographic study for free, with public metadata (Crossref), free metadata (Dimensions), and free geocoding processing (NETSCITY). This method is generic: one may study the geography of any other subject outside of the presented case study (worldwide retractions). Our contribution includes the integration of a new feature in NETSCITY: the import of bibliographic data from the Dimensions bibliographic database.

## **2. Method: Geographic analysis of a bibliographic corpus at the country and city levels**

The Retraction Database (retractiondatabase.org) tabulates 39,500 retractions as of April 2023. These data collected by curators from the Center For Scientific Integrity are subject to licensing. Since we wished to explore the geography of retractions based on open data, we opted for an alternative source: Crossref (Hendricks *et al.*, 2020). Crossref is a DOI (Digital Object Identifier) registration agency that mints DOIs for publishers that are Crossref members. The contract binding Crossref and publishers requires them to update their records’ metadata, notably by registering retractions. Crossref releases metadata to the public for free and no copyright is applicable since these metadata are facts (Hendricks *et al.*, 2020, p. 415). We queried the Crossref API with a URL (<https://api.crossref.org/works?filter=update-type:retraction>) that retrieved 12,437 publication records identified by a DOI together with bibliographic metadata (e.g., title, year, authors, venue).

The affiliation of authors is not always available from Crossref records. And yet, this information is key to unveil the geography of retractions. We tackled this pitfall by querying the Dimensions database (Herzog *et al.*, 2020) to get the affiliation metadata of the retracted publications retrieved from Crossref. Registered users of the free version of Dimensions can download up to 5k records per query. We split the 12,437 list of DOIs into bulks of 300 DOIs. This bulk size of 300 was found to yield less than 5k records for each query. We submitted  $[12,437 / 300] = 42$  queries iteratively and downloaded each result with the CSV export feature. In brief, the corpus of retractions comprises 42 CSV files that store the publication records extracted from Dimensions based on the DOIs of the 12,437 Crossref records marked as ‘retracted.’

The processing of geographic information conveyed in the authors' affiliation was performed with NETSCITY, a web application that geocodes and enriches geographic metadata at the level of urban areas, countries and world regions (Maisonobe et al., 2019). NETSCITY ingests bibliographic data from Scopus, Web of Science, and tabulated files. We developed an import feature for Dimensions data in 2023 that runs the following workflow.

First, it extracts authors' raw affiliations (column W) in each Dimensions CSV file. Since these data are heterogeneous (Figure 1), we designed a cleaning pipeline illustrated in Figure 2. First, irrelevant and noisy contents are removed: email addresses, incomplete affiliations, postal codes, extra white-spaces and commas (step 1). Next, the pipeline separates distinct affiliations. For each affiliation, we retain the contents separated by the last three commas: they define a triplet consisting of a city, a province (or the first admin sub-level), and a country (CPC, step 2). The pipeline then separates these CPC and normalises them: extra white-spaces, dashes, accents, full stops and parentheses are discarded. The remaining character strings are capitalised, ampersands are replaced with the word "AND", spaces and slashes are replaced by a hyphen. Extra care is taken to homogenise abbreviated and non-abbreviated forms of spatial areas: "SAINT-" are replaced with "ST-". Regularly omitted parts of toponyms are discarded, such as "REPUBLIC-OF-", "-OF-AMERICA" and "-CITY" to increase the likelihood of the matching procedure that searches these toponyms in the NETSCITY geographical database (step 3). When a match is found, the spatial coordinates and the identifier of the matching urban area (IDCOMPOSITE) are collected from NETSCITY and attached to each CPC (step 4). This geocoding process fails when the CPC does not exist in either the NETSCITY database or the geocoding web service LocationIQ (step 5), but a manual correction of addresses can be performed.

The successful geocoding of a publication record is stored in the user's web browser. When all inputted records are processed, NETSCITY supports the user with features to export, display and analyse the geocoded corpus (step 6).

Figure 1: Examples of raw affiliations extracted from Dimensions exports. We focus on the cities (in yellow), provinces (in green) and countries (in blue). Some affiliations lack some items (such as provinces and countries). Some publications list multiple addresses. We masked author names, laboratories and universities for anonymity concerns.

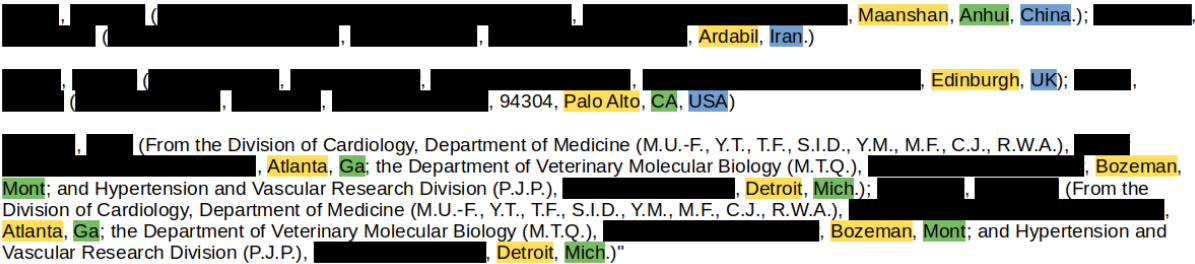
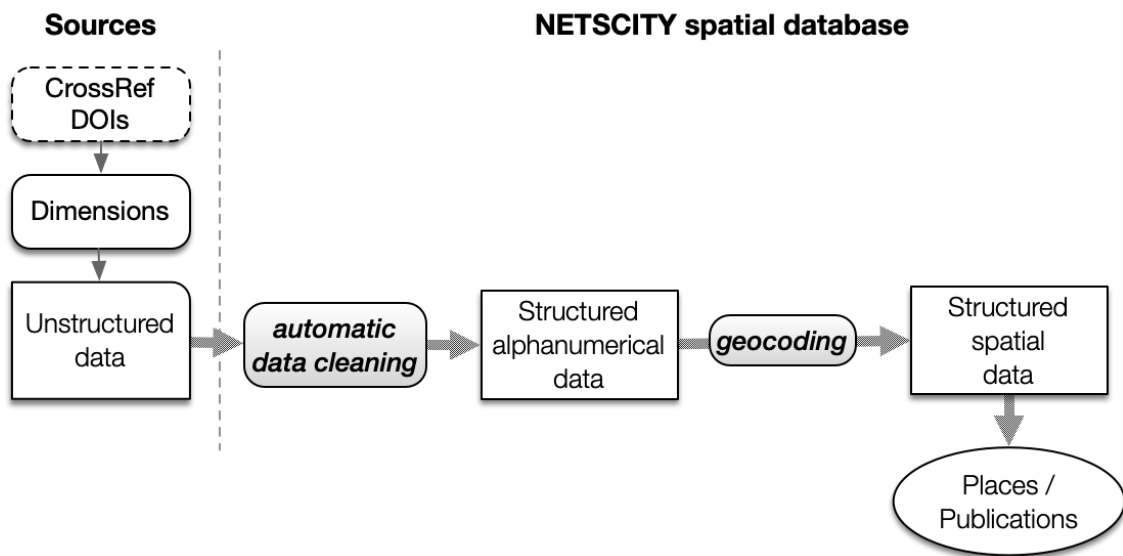


Figure 1 shows several examples of raw affiliations extracted from Dimensions exports. The text is partially masked with black boxes for anonymity. Geographic information is highlighted: cities in yellow, provinces in green, and countries in blue. The examples include:  
1. A masked name, followed by a masked location, then "Ardabil, Iran." (Ardabil is yellow, Iran is blue).  
2. A masked name, followed by a masked location, then "94304, Palo Alto, CA, USA" (Palo Alto is yellow, CA is green, USA is blue).  
3. A masked name, followed by "(From the Division of Cardiology, Department of Medicine (M.U.-F, Y.T., T.F., S.I.D., Y.M., M.F., C.J., R.W.A.), [masked], Atlanta, Ga, the Department of Veterinary Molecular Biology (M.T.Q.), [masked], Bozeman, Mont; and Hypertension and Vascular Research Division (P.J.P.), [masked], Detroit, Mich.); [masked] (From the Division of Cardiology, Department of Medicine (M.U.-F, Y.T., T.F., S.I.D., Y.M., M.F., C.J., R.W.A.), [masked], Atlanta, Ga, the Department of Veterinary Molecular Biology (M.T.Q.), [masked], Bozeman, Mont; and Hypertension and Vascular Research Division (P.J.P.), [masked], Detroit, Mich.)"

Figure 2: Overview of the affiliation city processing pipeline.



We fed the NETSCITY workflow with the Crossref–Dimensions retraction corpus collected earlier. In total, 59,848 addresses were imported, with a geocoding success rate of 93.8%. The resulting Address Table (Figure 3) was exported for further analysis with a R script.

Figure 3: Screenshot of the top rows of the “addresses table” produced by NETSCITY.

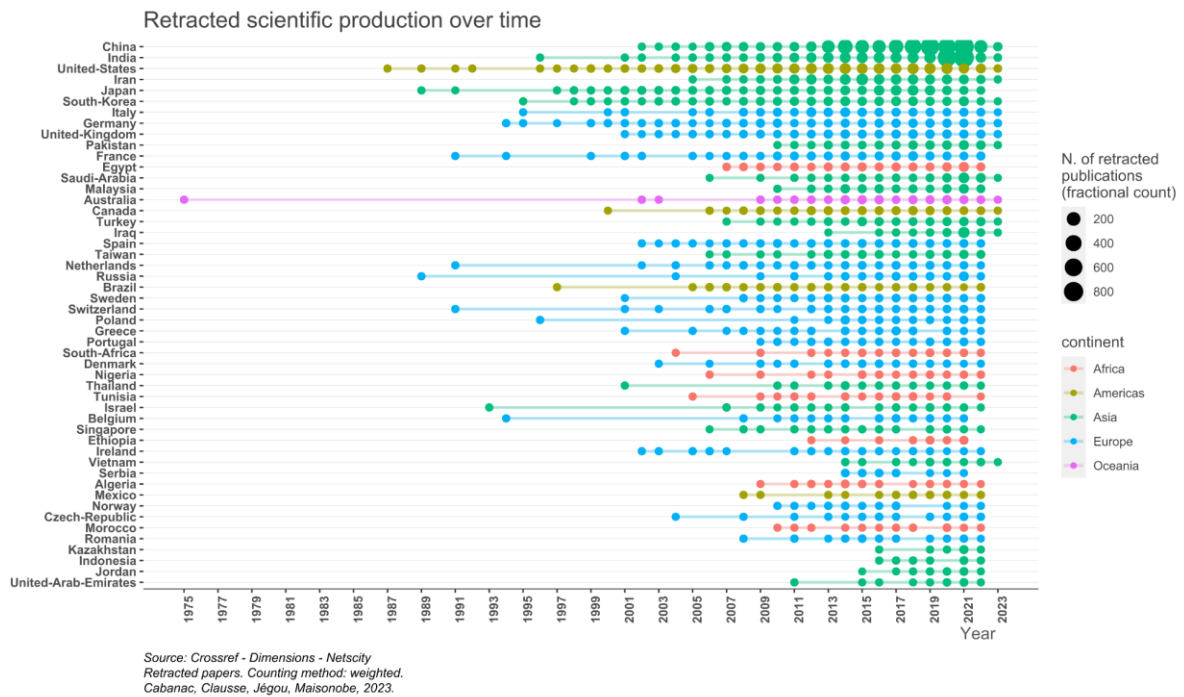
identifier	year	address	city	agglomeration	province	code	country	sub-region	region	continent	latitude	longitude	idcomposite	confidence
15859097	2011	DEPARTMENT OF PATHOLOGY I, KANSAI MEDICAL UNIVERSITY, OSAKA, JAPAN	OSAKA	KYOTO		JP	JAPAN	Eastern Asia	Asia	Asia	34.69	135.3928	AD7568	SANS_API
15859099	2011	DEPARTMENT OF PATHOLOGY I, KANSAI MEDICAL UNIVERSITY, OSAKA, JAPAN	OSAKA	KYOTO		JP	JAPAN	Eastern Asia	Asia	Asia	34.69	135.3928	AD7568	SANS_API
15859101	2011	DEPARTMENT OF PATHOLOGY I, KANSAI MEDICAL UNIVERSITY, OSAKA, JAPAN	OSAKA	KYOTO		JP	JAPAN	Eastern Asia	Asia	Asia	34.69	135.3928	AD7568	SANS_API
15859103	2011	DEPARTMENT OF PATHOLOGY I, KANSAI MEDICAL UNIVERSITY, OSAKA, JAPAN	OSAKA	KYOTO		JP	JAPAN	Eastern Asia	Asia	Asia	34.69	135.3928	AD7568	SANS_API
15859105	2011	DEPARTMENT OF PATHOLOGY I, KANSAI MEDICAL UNIVERSITY, OSAKA, JAPAN	OSAKA	KYOTO		JP	JAPAN	Eastern Asia	Asia	Asia	34.69	135.3928	AD7568	SANS_API
15859107	2011	DEPARTMENT OF PATHOLOGY I, KANSAI MEDICAL UNIVERSITY, OSAKA, JAPAN	OSAKA	KYOTO		JP	JAPAN	Eastern Asia	Asia	Asia	34.69	135.3928	AD7568	SANS_API
15859108	2014	DEPARTMENT OF PATHOLOGY, AFFILIATED HOSPITAL OF SHANDONG ACADEMY OF MEDICAL SCIENCES, JINAN, SHANDONG, CHINA	JINAN	JINAN	SHANDONG	CN	CHINA	Eastern Asia	Asia	Asia	36.685	117.0595	AD2181	SANS_API
15859109	2014	SHANDONG UNIVERSITY SCHOOL OF MEDICINE, JINAN, SHANDONG, CHINA	JINAN	JINAN	SHANDONG	CN	CHINA	Eastern Asia	Asia	Asia	36.685	117.0595	AD2181	SANS_API
15859110	2014	DEPARTMENT OF OBSTETRICS AND GYNECOLOGY, PEKING UNION MEDICAL COLLEGE HOSPITAL, CHINESE ACADEMY OF MEDICAL SCIENCES AND PEKING UNION MEDICAL COLLEGE, BEIJING, CHINA	BEIJING	BEIJING		CN	CHINA	Eastern Asia	Asia	Asia	40.008	116.269	AD2074	SANS_API

### 3. Results : Main locations of retracted publications at the global scale and over time

#### 3.1. Demographics of retracted publications per country and city

In the following figure (Figure 4), countries are ranked in decreasing order of their number of retracted papers (fractionated counts). China and India stand out with a remarkable growth in their total number of retracted papers between 2010 and 2020. The growth of retractions affiliated to China in the last decade is especially distinctive.

Figure 4: Retracted scientific production over time by country of publication.



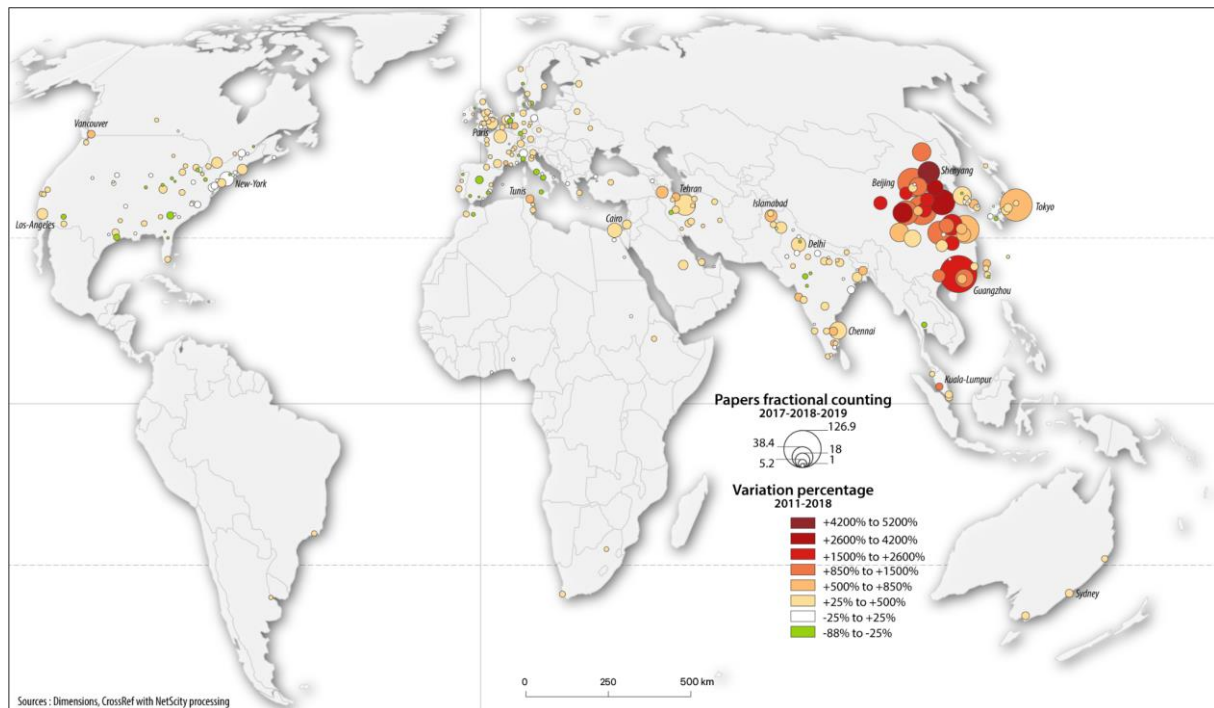
To analyse more precisely the evolution of this geography, we focus on two time periods of three years: 2010-2011-2012 and 2017-2018-2019. From the first to the second period, we observe that:

- The share of international collaborations found in the bylines of the retracted publications decreased from 19% to 15%.
- China and India's growth in number of retracted publications overtook that of the US and Iran. It is even truer for domestic publications only ('domestic publications' feature authors from urban areas all part of the same country).
- China's growth largely surpassed that of India: the number of retracted publications of China has increased by 20 times, and that of India by 3 times only.
- The number of countries involved in retracted papers increased from 55 to 89 and the number of urban areas involved increased from 409 to 900.

Observing the ranks, the cities involved in the largest number of retracted publications changed in favour of Chinese cities. Iranian, South-Korean, American and European urban areas that were in the top 10 urban areas have been replaced by Chinese cities. The only urban area being both in the 2010-2012 top 10 and in the 2017-2019 top 10 is Tokyo. Japan showed an important growth over the period as it has increased by 5 times between 2010-2012 and 2017-2019.

Figure 5 shows the rates of change in the number of retracted papers by urban area between the two periods for all locations found in the affiliations of a retracted paper in 2017-2018-2019. The growing importance of Chinese urban areas over this period is outstanding, as it is for the overall number of publications in the world.

Figure 5: The evolution of retracted scientific production by urban area between 2010-2011-2012 and 2017-2018-2019.



### 3.2. Networks of cities where retracted publications were produced

Specific interurban clusters stand out, suggesting that there could exist groups of places whose authors tend to co-author problematic-then-retracted publications together (Figure 3 and 4). Most of the interurban co-publications are domestic only, but a few international clusters or cities also stand out. This observation is consistent with the known fact that there is some organisation in the manufacture of problematic articles. For example, there is evidence on PubPeer of “affiliation trafficking” to boost the number of internationally cooperative papers either to raise the scores of certain institutions in the rankings or to improve authors’ productivity indicators (Bhattacharjee, 2011).

To highlight these problematic assemblages, we isolated the groups of urban areas involved in at least two retracted papers over each 3-year period that we analysed. In order to identify not only dyads but also combinations of places whose size can go from 2 to  $n$ , we worked with hypergraphs. A hypergraph “is a generalisation of a graph in which an edge can join any number of vertices. In contrast, in an ordinary graph, an edge connects exactly two vertices”<sup>1</sup>. Since journal articles are co-authored from 1 to  $n$  separate places, working with hyperedges allowed us to preserve the complexity of the original information.

Applying a threshold of at least 2 retracted co-publications, we found 27 hyperedges during the 2010-2012 period and 100 hyperedges during the 2017-2019, that is 4 times more. Most hyperedges counts 2 vertices, but a few hyperedges are of size 3 ({Houston, Washington, New-York} in the first time period) and 4 ({Jaipur, Rio-de-Janeiro, Indore, Udaipur}, {Genoa, Murcia, Madrid, Grenade} in the first time period). Most hyperlinks group together urban areas belonging to the same countries, but a few hyperlinks are international: 9 out of 27 between 2010 and 2012 and 23 out of 100 between 2017 and 2019.

<sup>1</sup>Source: <https://en.wikipedia.org/wiki/Hypergraph>

Figure 6: Groups of urban areas between which at least 2 retracted articles were co-authored in 2010-2012.

Groups of urban areas between which at least 2 retracted articles were co-authored in 2010-2012

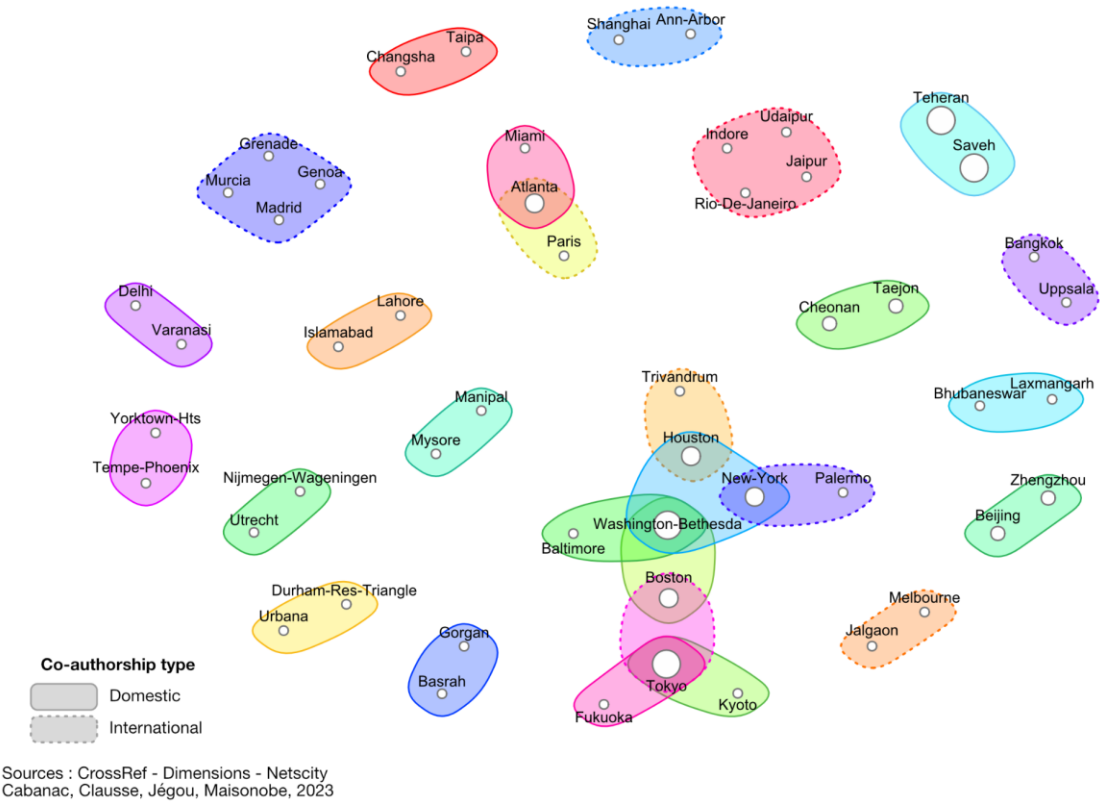
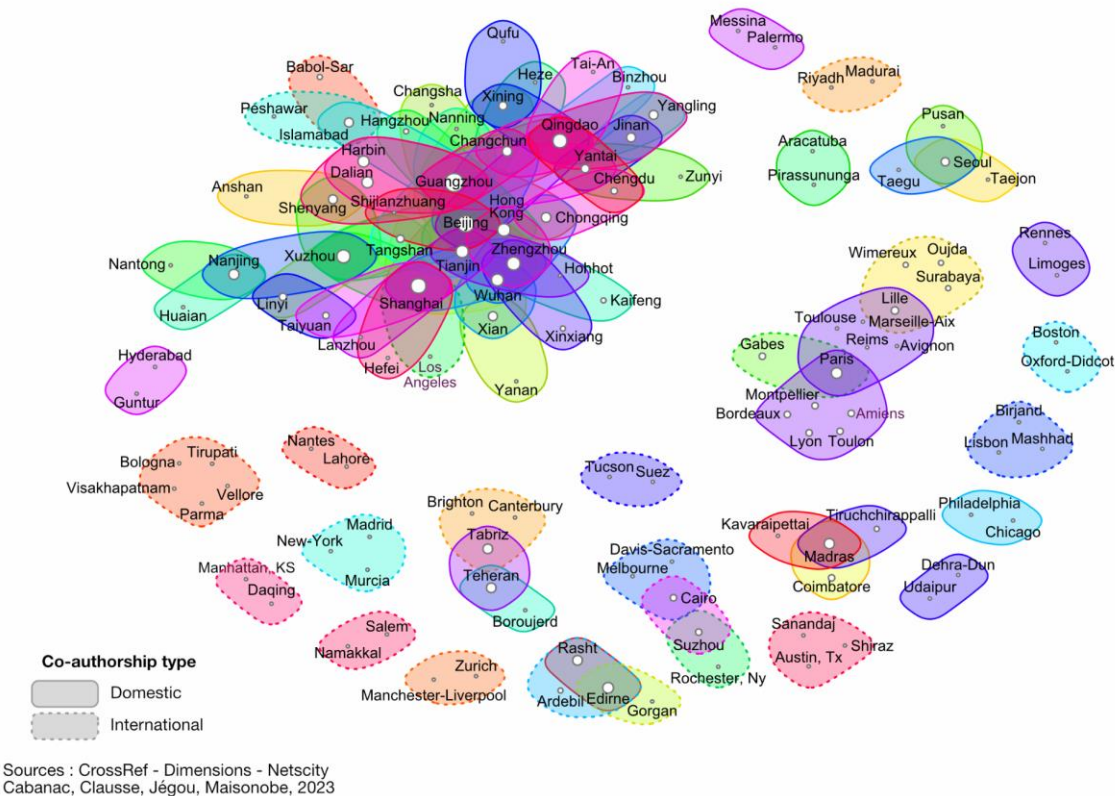


Figure 4: Groups of urban areas between which at least 2 retracted articles were co-authored in 2017-2019



## Groups of urban areas between which at least 2 retracted articles were co-authored in 2017-2019



## 4. Limitations

This section discusses two limitations of our geographic exploration of retractions. First, we opted to source retractions from Crossref, as publishers are asked to mark their records accordingly. However, less retractions are recorded in Crossref compared to the Retraction Database (12k vs. 30k). This discrepancy suggests that some publishers failed to abide with their duties as Crossref members. It also suggests that the present research reveals *part of* the geography of science *only*. There is no evidence that the sampling indirectly performed when drawing data from Crossref vs. from the Retraction Database is random. Readers should bear in mind that retractions issued by publishers that don't update their Crossref records are absent from this analysis.

Second, the rationale behind a retraction runs the whole gamut from involuntary errors (sometimes called 'honest mistakes') to proven scientific misconduct. Consequently, counting retractions is an indirect and imprecise indicator of scientific misconduct.

## 5. Conclusion

This research suggests that the geography of retracted papers has evolved between the late 2000s and the late 2010s. More countries and urban areas are concerned, and we observe a growing number of groups of affiliations that are often co-occurring, which contributes to drawing a specific geography that differs from the usual geography of co-authorship of articles.

Considering the scientific fields and specific journals (or publishing groups) in which these retracted articles were published is part of our future work. We also plan to compare this

geography with that of scientific production in general in order to identify the places (and groups of places) featuring an over-representation of retracted publications.

### **Open science practices**

Crossref releases metadata to the public for free. We queried the Crossref API with the following URL, <https://api.crossref.org/works?filter=update-type:retraction>.

The affiliation of authors was found by querying the Dimensions database with the DOIs, to get the affiliation metadata of the retracted publications retrieved from Crossref.

The processing of geographic information conveyed in the authors' affiliation was performed with [NETSCITY](#), a web application that geocodes and enriches geographic metadata at the level of urban areas, countries and world regions. NETSCITY is free to use, the source code will be released in open access when the software will be considered as stable. To geocode the paper's affiliations, it mainly uses the [GeoNames](#) geographical open database.

The data from our analysis will be made available on a public repository during the review period of the publication.

### **Author contributions**

Guillaume Cabanac: Conceptualization, Data curation, Funding acquisition, Software, Writing - review & editing.

Alexandre Clause: Software, Data curation, Writing - review & editing

Laurent Jégou: Visualization, Software, Writing - review & editing

Marion Maisonobe: Formal Analysis, Methodology, Visualization, Writing – original draft

### **Competing interests**

We declare no competing interests.

### **Funding information**

For this research, we were supported by two research grants:

- Knowledge and Innovation on Grain-legumes in food sciences & technology (KING), Région Occitanie grant.
- Royal Society of Edinburgh Saltire Award 'Geography of Collaboration'

### **References**

Bhattacharjee, Y. (2011). Saudi Universities Offer Cash in Exchange for Academic Prestige. In *Science* (Vol. 334, Issue 6061, pp. 1344–1345). <https://doi.org/10.1126/science.334.6061.1344>

Barbour, B., & Stell, B. M. (2020). PubPeer: Scientific assessment without metrics. In M. Biagioli & A. Lippman (Eds.), *Gaming the metrics: Misconduct and manipulation in academic research* (pp. 149–155). Cambridge, MA, USA: MIT Press. <https://doi.org/10.7551/mitpress/11087.003.0015>

Cabanac, G., & Labbé, C. (2021). Prevalence of nonsensical algorithmically generated papers in the scientific literature. In *Journal of the Association for Information Science and Technology* (Vol. 72, Issue 12, pp. 1461–1476). <https://doi.org/10.1002/asi.24495>

Else, H., & Van Noorden, R. (2021). The fight against fake-paper factories that churn out sham science. In *Nature* (Vol. 591, Issue 7851, pp. 516–519). <https://doi.org/10.1038/d41586-021-00733-5>

Fanelli D, Costas R, Larivière V (2015) Misconduct Policies, Academic Culture and Career Stage, Not Gender or Pressures to Publish, Affect Scientific Integrity. *PLOS ONE* 10(6): e0127556. <https://doi.org/10.1371/journal.pone.0127556>

Fanelli D, Larivière V (2016) Researchers' Individual Publication Rate Has Not Increased in a Century. *PLOS ONE* 11(3): e0149504. <https://doi.org/10.1371/journal.pone.0149504>

Grossetti, M., Eckert, D., Gingras, Y., Jégou, L., Larivière, V., & Milard, B. (2014). Cities and the geographical deconcentration of scientific activity : A multilevel analysis of publications (1987-2007). *Urban Studies*, 51(10), 2219-2234. <https://doi.org/10.1177/0042098013506047>

Hvistendahl, M. (2013). China's Publication Bazaar. In *Science* (Vol. 342, Issue 6162, pp. 1035–1039). <https://doi.org/10.1126/science.342.6162.1035>

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. In *Quantitative Science Studies* (Vol. 1, Issue 1, pp. 414–427). [https://doi.org/10.1162/qss\\_a\\_00022](https://doi.org/10.1162/qss_a_00022)

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. In *Quantitative Science Studies* (Vol. 1, Issue 1, pp. 387–395). [https://doi.org/10.1162/qss\\_a\\_00020](https://doi.org/10.1162/qss_a_00020)

Maisonobe, M., Jégou, L., & Cabanac, G. (2018). Peripheral forces. In *Nature* (Vol. 563, Issue 7729, pp. S18–S19). <https://doi.org/10.1038/d41586-018-07210-6>

Mallapaty, S. (2020). China's research-misconduct rules target 'paper mills' that churn out fake studies [News]. In *Nature*. <https://doi.org/10.1038/d41586-020-02445-8>

Oransky, I. (2021). Springer Nature slaps more than 400 papers with expressions of concern all at once, *Retraction Watch*, <https://retractionwatch.com/?p=123181>

Oransky, I. (2022a). Retractions are increasing, but not enough. In *Nature* (Vol. 608, Issue 7921, p. 9). <https://doi.org/10.1038/d41586-022-02071-6>

Oransky, I. (2022b). Physics publisher retracting nearly 500 likely paper mill papers, *Retraction Watch*, <https://retractionwatch.com/?p=125630>