Delineating urban agglomerations across the world: a dataset for studying the spatial distribution of academic research at city level

Marion Maisonobe, Laurent Jégou, Denis Eckert

Cybergeo : European Journal of Geography
Data papers | 2018

Delineating urban agglomerations across the world: a dataset for studying the spatial distribution of academic research at city level

Définir des agglomérations à l'échelle mondiale : un ensemble de périmètres urbains pour étudier la répartition des activités de recherches

Marion Maisonobe, Laurent Jégou and Denis Eckert

Electronic version
URL: http://journals.openedition.org/cybergeo/29637
DOI: 10.4000/cybergeo.29637
ISSN: 1278-3366

Publisher
UMR 8504 Géographie-cités

Electronic reference
Marion Maisonobe, Laurent Jégou and Denis Eckert, « Delineating urban agglomerations across the world: a dataset for studying the spatial distribution of academic research at city level », *Cybergeo : European Journal of Geography* [Online], Data Papers, document 871, Online since 12 November 2018, connection on 20 April 2019. URL : http://journals.openedition.org/cybergeo/29637 ; DOI : 10.4000/cybergeo.29637

Delineating urban agglomerations across the world: a dataset for studying the spatial distribution of academic research at city level

Définir des agglomérations à l'échelle mondiale : un ensemble de périmètres urbains pour étudier la répartition des activités de recherches

Marion Maisonobe, Laurent Jégou and Denis Eckert

Introduction: Defining a set of comparable agglomerations at the world scale

1 Attempting to delineate urban areas at the world scale is a challenge because of the lack of homogeneous datasets at this scale. This problem is well known in comparative urban research (Moriconi-Ebrard, 1991; Pumain et al., 2015). In order to build comparable international datasets at city level, geographers frequently aggregate small spatial units which can be grouped into bigger entities (agglomeration, functional urban region, etc.) that make more sense for data interpretation. At the European scale, an important work has been done to delineate functional urban regions using data on home-work commutes (Guérois *et al.*, 2014). Unfortunately, information required to delineate functional urban areas are not available with a homogeneous quality for all countries in the world and there are few datasets with a high resolution at the scale of the entire world, except for simplified land use or population density.

2 To study the distribution of scientific activity in the world we decided to develop a new methodology. Provided some adjustments, we believe this methodology could now serve for various purposes. Indeed, our aim was to produce some universal delimitation criteria, and not divisions corresponding to a juxtaposition of national criteria (for example, using SMSAs for the United States, *Aires Urbaines* in France, etc.). To do so, we

used both the spatial distribution of scientific affiliations indexed in a bibliographic database (localities from which scientific publications are authored) and the distribution of population density (fine-tuned raster data). For the world's 500 most "publishing" localities, the delineations have been double-checked by specialists of the regions or countries needing verifications and by members of our team.

3 In this data paper, we give access to the cartographic information delineating agglomerations that are participating in more than 80% of the world scientific production between 1999 and 2014. First, we explain the specificity of the distribution of scientific production and detail the source and the geocoding stage, then we detail the methods used to delineate a set of comparable agglomerations. After a short discussion, we show how these spatial delineations can be used to analyse the spatial distribution of Cybergeo's publications between 2015 and 2017.

The specificity of the spatial distribution of scientific activities and its geocoding

4 Science is done in different types of settings: universities, research centres, R&D departments of enterprises, hospitals, academies, observatories, NGO, etc. The places of science are diverse and some of them have been settled down centuries ago (Livingstone, 2003). As a result, the spatial logics explaining the distribution of scientific activities are diverse and this distribution does not fit perfectly that of the population (Maisonobe, 2015). For instance, university cities such as Oxford, Cambridge, College Station etc. are mostly populated by students and professors. These locations are often non-metropolitan. Further, when scientific venues are located in metropolitan areas, they can be found both in old city centers and in suburban areas. Sometimes, the remoteness is justified by scientific purposes: for instance, observatories should be localised in elevation and marine observatories nearby the sea. In other cases, it is economic or urban planning rationales that are driving the decisions, such as the project of delocalizing Parisian universities in the Greater Paris suburbs. Thus, in a single metropolitan area, the scientific production can be authored from a high diversity of postal addresses. In countries where the administrative fragmentation of the territory is very developed such as in France (almost 37 000 municipalities), metropolitan areas can encompass dozens of "publishing" localities.

5 To study this phenomenon in an international comparison, we used the *Web of Science Core Collection*, which is one of the most comprehensive bibliographic data sources. This database is indexing more than 1 million scientific publications (articles, reviews and letters) annually. The base element of this source is the bibliographic record, listing the authors, their institutions and the postal addresses of their institution. In a frame of a partnership with the French organism of statistics OST-HCERES (*Observatoire des Sciences et Techniques*), we retrieved all the authors' addresses indexed from 1999 to 2014 and geocoded them.

6 The bulk of the geocoding work was done by supervised automatic matching, using a twofold process. First, we used available geographical databases like GeoNames[1] and Nominatim[2] (the gazetteer of the OpenStreetMap project) to assign geographical coordinates to authors addresses. These digital resources allowed us to geocode approximately 80% of the total number of addresses, those with easily recognizable city-

province-country triplets. The problem was to localize the remaining 20% addresses, about 40,000 items.

7 We faced several geocoding challenges:

- Somewhat small localities not mentioned in open gazetteers;
- Confusions between street names, neighbourhoods and cities in the source data;
- Confusion between institutions' name and cities in the source data;
- General lack of province or sub-country level information;
- Homonyms between city names and province of the same country.

8 The second phase of the geocoding process used automatic services such as Google Maps API and Microsoft's Bing Maps API, to further improve the geocoding. These web services have access to a much larger toponyms database than free gazetteers and can automatically resolve ambiguities between several alternative homonym locations, helped by spatial hints like the country or the province. A human operator supervised this procedure: he could input significant parameters like the country or geographical zone of search, but most especially examine and correct the results, if needed. To help working with these web services and embed their use into a more integrated and fluid geocoding procedure, we developed a range of web applications, from data correction to geocoding and the evaluation of results (Jégou, 2014).

9 After more than one year of work, with the help of specialists in the fields of sociology and geography of science, geospatial analysts and cartographers, we obtained a fine-grained spatial database of the scientific production over the last decades. Author's addresses were geocoded, by using the "city" string in the addresses. The geographical points thus obtained are the elementary spatial units used in our research; we call them "localities".

10 In 2016, this work has been updated for the most recent publishing years (2009-2014) (Table 1)[3].

Table 1. Number of geocoded publications, localities and resulting agglomerations by year series

	Publications	Localities	Agglomerations
2000*	832 254	14 422	4 483
2003*	950 615	15 620	4 802
2006*	1 129 912	16 987	5 196
2009*	1 340 916	19 091	5 729
2013*	1 614 400	20 808	6 123

*mobile average over 3-years

Source: Web of Science Core Collection, Clarivate Analytics/OST-HCERES

11 The next phase of the work is the spatial clustering of these "localities" in order to build scientific "agglomerations".

Construction of a set of urban agglomerations for international comparison

12 Given the worldwide comparative scope of our work, we consider the locality level as inadequate for international comparison. The characteristics of the mail address, originally designed for postal use, the geographical variability of the postal reference systems and the great diversity of administrative geographical segmentation, prevent any direct comparison between the "scientific localities" (postal addresses from which publications are authored). Our team addressed this problem by building globally comparable geographical entities at an agglomerated level.

13 Surprisingly, most authors that deal with the spatialisation of scientific activity by using publication data do not address this issue. Following the spatial turn in scientometrics studies (Frenken et al., 2009), many articles in scientometrics present results at the level of geocoded addresses without clustering them into urban areas (eg. Waltman et al., 2011; Pan et al., 2012; Masselot, 2016; Csomós, 2018). These articles do not consider the issue of the statistical heterogeneity of the geographical entities they compare.

14 Yet some exceptions can be mentioned:

- Matthiessen et al. (2002, 2010) address the issue but only focus on few metropolitan areas ("world cities of scientific knowledge"), mainly in the US and Europe;
- Comin (2009) deals with the issue and uses the FUR delineations (functional urban areas) for a comparison at European scale;
- Catini *et al.* (2015) focus on small clusters within metropolitan areas, a method rather similar to the one presented here, by using also population density data.
- Several scholars in regional economy use the European nomenclature NUTS (level 2 or 3).

15 In order to cluster scientific localities, we exploit global data sets that are highly fine-tuned and of comparable quality for the whole world.

16 Different global data sets met these conditions at the time we began the study: there was on one side land occupation data, such as ESA Iona GlobCover (ESA, 2005) or Global UrbanExtent (Schneider *et al.*, 2009); and datasets focussing on population densities on the other. Comparing urban areas obtained by using data from land artificialization and data on population densities, we concluded that land artificialization was not the best criterion for our purpose (Eckert *et al.*, 2013). This holds particularly true in continuously built coastal areas (especially in touristic places), which do not necessarily match continuous human occupation, and are even less likely to harbour areas of scientific activity. On the contrary, data on population density allowed us to delineate urban spots that better correspond to the limits of "local innovation systems" (Bathelt *et al.*, 2004).

17 To delineate urban zones by taking into account the distribution of population density we used a dataset produced by the SEDAC[4] (the Socioeconomic Data and Applications Center of the NASA) for the year 2005[5] (SEDAC, 2005).

18 Due to the extreme variability of density, it was impossible to define a single and universal threshold value enabling to differentiate urban areas, in particular in the most densely urbanised areas of the world. To mitigate this problem, we decided to reason relatively. The solution was to use an indicator to identify gradient slope changes in the spatial distribution of population density. In spatial analysis, one could use several indicators for this purpose: Local Indicators of Spatial Association (LISA), including the

local I Moran. It delimits spatially significant areas called "density nuclei" (Anselin, 1995) using the population density distribution in a homogeneous way over the territory. The obtained urban areas, automatically delineated by applying this indicator, were combined with the spatial distribution of scientific localities.

19 For the denser urban spaces both in terms of population and scientific production, we checked the delimitation on a case-by-case basis, sometimes questioning local experts. For instance, we took into account the spatial distribution of close scientific localities (e.g. including a small locality near an existing agglomeration). We also sometimes considered the presence of key transport infrastructures connecting close urban zones (highways, bridges, or ferries). The relevance of such a meticulous work to define the boundaries of the most populated agglomerations is no doubt. In 2012, this procedure allowed us to delineate 376 agglomerations encompassing the 500 most publishing localities (in decreasing order of the total number of scientific publications in 2008). Among the list of these 500 localities, some were affected to the same agglomeration, like Manchester and Liverpool. Thus, the number of manually delineated agglomerations actually in the final dataset is inferior to 500.

20 For the publishing localities associated with smaller volumes of publications, a fully automated procedure was chosen. The localities located outside densely populated areas have been grouped together to form agglomerations when their distance to each other was less than 40 km. This criterion was applied following a descending order, that is to say that the centres from which to apply the distance criterion were selected in the order of the most publishing to the least publishing. In order to apply this criterion, our team used spatial queries using the PostGIS extension of the PostgreSQL database system.

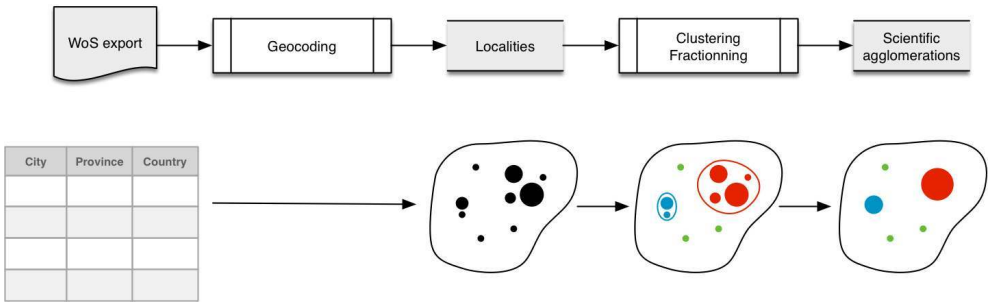
21 For each science center, this query returned the list of localities fulfilling the following conditions:

- not being already agglomerated to another center;
- being less than 40 km from this center;
- being associated with fewer publications than this center.

22 As a result, two types of spatial objects were obtained: polygons grouping at least three localities and lines grouping only two localities. Thus, this automatic procedure enabled us to group St. Andrews and Dundee, which are nearby places, too far away from Edinburgh to be integrated into its urban agglomeration.

23 It is important to specify that our objective was to draw agglomerations aiming at capturing and clustering “publishing” localities. The course of the boundaries of the lines or polygons thus obtained have no significance *per se*. Their function is solely to gather the punctual information of the localities pertaining to a given agglomeration (Figure 1).

Figure 1. From "publishing" localities to "scientific agglomerations"

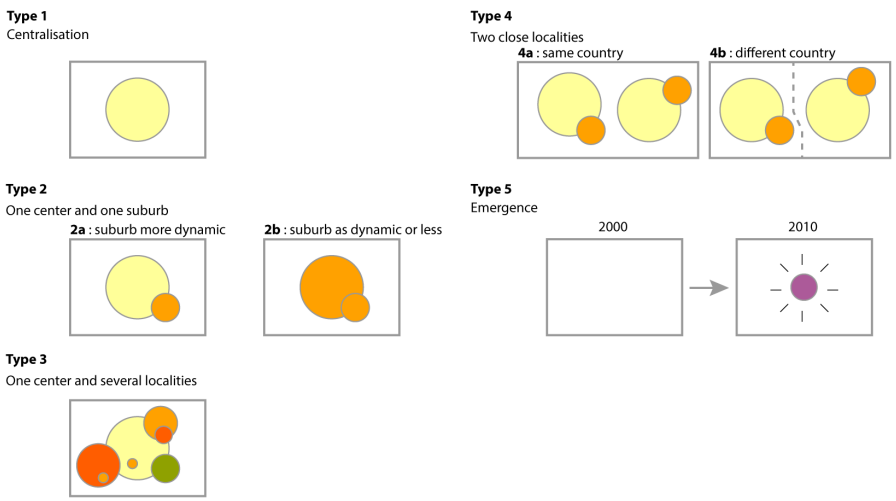


Design: Laurent Jégou

24 The most important justification for our approach is that it helps homogenizing the spatial entities compared in our research. Before clustering, we could distinguish between different configurations (Figure 2). Although these configurations could be interesting to study *per se*, they are too often the result of the varying levels of cities' administrative fragmentation in the world. Sometimes, the spatial distribution is very simple: a single center of publication, stable in time (no new centers of production in the agglomeration during the period under study). Corresponding to our type 1, we can think of Beijing in China or Kiev in Ukraine. The simplicity of this pattern can be easily explained by the administrative structure: one single huge municipality.

25 On the contrary, when the administrative structure is more fragmented, we often observe the type 3 where many smaller but significant scientific localities are adjacent to the main center (typically Paris urban region). It is especially important to take this configuration into account when the main urban center accounts for a smaller part of the total scientific output of the agglomeration (typically Washington DC and surroundings). Consequently, taking only into account this center locality instead of the whole multi-center agglomeration can lead to important data misrepresentation.

Figure 2. A typology of scientific agglomerations' varieties



Design: Laurent Jégou

Discussion and limits

26 Our dataset was built to be used only at the global level, for urban comparisons and in the domain of scientific production. It will require specific data verification if used for other purposes, especially at larger geographical scales or smaller areas (i.e. a single state, a single region or a single metropolitan area).

27 The delineations produced and used so far are likely to evolve with the update of publication data as well as population density data. Nevertheless, we consider the delineations of the world agglomerations from which are authored more than 80% of the publications between 1999 and 2014 to be robust enough to be shared and reused both for spatializing a corpus of scientific publications (next section). These delineations can also serve for exploring other types of academic activities, or relating scientific publication to other indicators (students, academic staff, funding...). Besides, these perimeters could be tested for their potential adequacy to the display of other human activities (like tourism, road traffic, pollution...). These perimeters would not be adequate *per se*, but their delineations might still be adapted to the specific distribution of these activities. More generally, we believe that the overall methodology could inspire geographers aiming at studying several types of world distributions.

28 Given the fact that the publication activity is increasingly distributed at the world level and that a growing number of places are contributing to this activity, the global share of these top 495 agglomerations is diminishing. Nevertheless, it still contributes to more than 80% of the world production in 2013 (Table 2).

Table 2. The total share of scientific production of the set of 495 agglomerations

	2000*	2003*	2007*	2010*	2013*
World share of scientific production of the top 495 publishing agglomerations (%)	88.4	87.5	86.3	83.9	82.6
Total number of publications in the Web of Science	829 309	975 245	1 203 035	1 411 756	1 651 684

*mobile average over 3-years
Source: Web of Science Core Collection, Clarivate Analytics/OST-HCERES

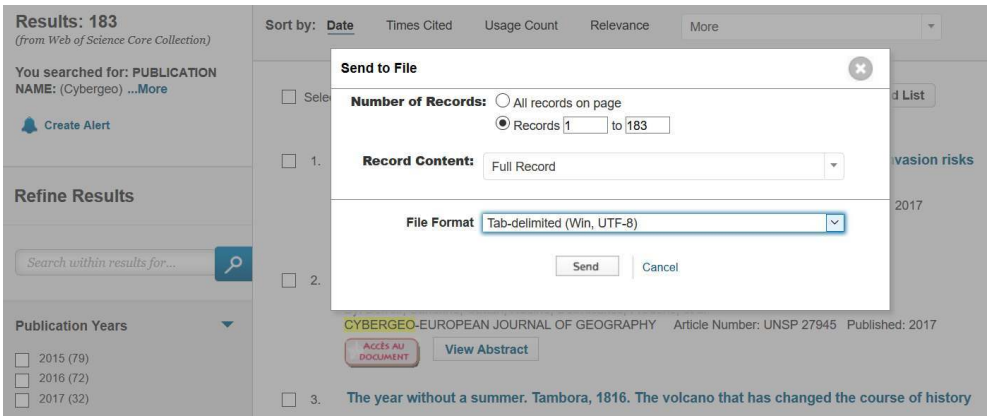
29 As a whole, the dataset shared in this data paper encompasses 495 agglomerations among which 376 were carefully designed by experts's hand (their ID name begins by "AD" as in "drawn") and the 119 others are automatic clusters of close publishing localities (their ID name begins by "AA" as in "automatic").

30 As an example of the use of these perimeters, we propose a spatial bibliometrics' analysis of the Cybergeo journal.

Application to analyse the spatial distribution of Cybergeo publications (2015-2017)

31 To study the spatial distribution of the authorship of Cybergeo papers, we need a list of institutional addresses. To obtain it, we can use the *Web of Science* since this database includes authors' addresses. Cybergeo entered the *Web of Science* in 2015 so that it is possible to extract the bibliographic records of this journal from 2015 to 2017.

Figure 3. Extracting a set of bibliographic records from the Web of Science



Source: Online version of the Web of Science Core Collection, Clarivate Analytics

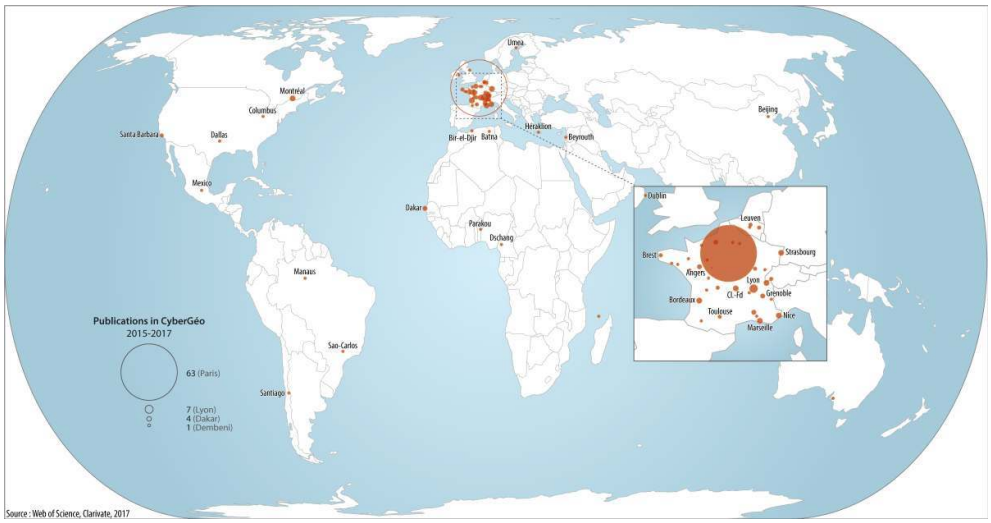
32 By searching for "Cybergeo" in the "Publication Name" entry, we get 183 results among which 79 publications indexed in 2015; 72 publications indexed in 2016 and 32 publications indexed in 2017 (the database administrators have not yet completed the indexing of all the literature published in 2017). We can download all these 183 records (Record content: "full record") by using the option "save to other file formats" and by choosing "Tab-delimited" with "UTF-8" as the text encoding option (Figure 3).

33 As a result, we obtain a table of 68 columns among which columns specifying, for each record, the authors' name, the publication title, the publication issue, the publication year, the number of references as well as the authors' addresses. To retrieve the spatial information of this dataset, we decide to focus only on one address by publication, which is the address of the corresponding author. Among the 183 records, 16 do not have any associated addresses (8 editorial materials, 1 book review and 7 articles). All the 167 remaining publications can be geocoded. To do so, we start by selecting only the ending part of the addresses (city name, province name and country name) which makes much easier for a geocoding tool to get relevant results. To geocode our list of 88 distinct location names (here "Paris, France" is considered to be distinct from "F-75005 Paris"), we use the "batch geocode tool" of the "Map Developers" website[6]. We retrieve all the distinct locations and we obtain a list of geographical coordinates. We enter these data in the GIS software QGIS and it allows us to derive a shapefile of all the publishing localities from which Cybergeo publications have been signed. Finally, we cluster the publishing localities (notably the one from "Paris, France" and from "F-75005 Paris") into agglomerations by using our dataset of 495 agglomerations' shapes.

34 Following this step, we found that 28 publications (among which 18 publications from France) are not located in one of the 495 most publishing agglomerations of the world. Nevertheless, all the publications that have been authored in dense urban areas are clustered thanks to our agglomerations' dataset. The remaining publications are authored from localities that are not in dense urban areas and therefore can be counted as separate punctual agglomerations. The only remaining problem occurring is with "Avignon" since three publications have been signed from Avignon but one specifying "F-84029 Avignon 1_ France" and the two others "Avignon_ France". As a result, the geocoding tool returns two different pairs of geographical coordinates, albeit very close to each other. To cluster these two points into one agglomeration, we construct a buffer around each isolated point and only keep one of the two buffers created around "Avignon".

35 Resulting from our analysis, we found that Cybergeo publications came from 58 different agglomerations during 2015 and 2017 among which 32 French agglomerations. Paris is the city from which the most important number of Cybergeo publications have been signed (63 publications) followed by Lyon (7 publications) and Bordeaux, Strasbourg, Geneva, Nice, Clermont-Ferrand, Montréal, and Marseille-Aix (with 4 publications each). The first non-francophone place from which Cybergeo articles are signed is Santa Barbara in California. Two publications are from Santa Barbara: one by Michael F. Goodchild and the other by Helen Couclelis. The importance of French publications can be explained both by the bilinguism of the journal and by the origin of the journal which was founded in Paris and is headquartered in the Géographie-Cités laboratory. Nevertheless, scientists from an interesting diversity of countries are contributing to Cybergeo: 19 different countries and notably countries with an under-represented level of production in the *Web of Science* (Algeria, Benin, Brazil, Cameroon, Chile, Greece, Lebanon, Mexico, and Senegal). The open data and open access characteristics of this journal might enhance this diversity (Figure 4). This distribution can also be explained by the important involvement of several French laboratories in the European field of theoretical and quantitative geography (Cuyala, 2013).

Figure 4. Map of urban agglomerations from which Cybergeo publications have been signed (corresponding author) between 2015 and 2017



Source: Web of Science Core Collection, Clarivate Analytics. Map: Marion Maisonobe and Laurent Jégou

Dataset Description

Spatial coverage:
World
Temporal coverage:
Publication data used to delineate agglomerations: 1999-2014
Format name and version:

- One geojson file of 495 agglomerations' shape
- One geojson file of the spatial distribution of Cybergeo publications indexed in the Web of Science database (Clarivate Analytics)

Creation dates:
The data set of 495 agglomerations was created between 2010 and 2016
The Cybergeo shape file has been created for the purpose of this data paper
Data creators and data cost:
The geocoding and clustering stages took 12-months for the publication years 1999-2001 plus 2006-2008 and 6 months of additional work have been necessary to take into account the publication years 2002-2005 and 2009-2014.

- The geocoding stage has been mainly handled by Laurent Jégou with the helpful contribution of Fabien Goblet and Marion Maisonobe as well as the verifications of all the members of the concerned Géoscience ANR programme.
- The clustering stage has been handled together by Denis Eckert and Laurent Jégou with the helpful contribution of Myriam Baron and Marion Maisonobe.
- The analyses and publications derived from this work have involved a larger team of researchers and most particularly Michel Grossetti and Béatrice Milard.

Data acquisition:
Data retrieved in the frame of a partnership between OST-HCERES and UMR LISST, which are based on the content of the Web of Science Core Collection (Clarivate Analytics).
Language:
English
Repository location:

- The urban agglomerations' shape file:
  https://www.nakala.fr/nakala/data/11280/3660ebc0
- The spatial distribution of Cybergeo publications (2015-2017):
  https://www.nakala.fr/nakala/data/11280/95f17b5d

License:

Reuse potential:
To our knowledge this is the only comprehensive dataset delineating agglomerations involved in the world scientific production in a systematic way and enabling their comparison at different dates and with other countries. It has been built to be only used at the world level. It needs additional data verification to be used at larger scales (i.e. smaller surfaces).
Acknowledgements:
This dataset has been tested, improved and used in the frame of the doctoral research of Marion Maisonobe with a support from CNRS and ANR grant Géoscience (PI Denis Eckert and Michel Grossetti). It is has led to several collective publications and communications

about the world geography of research (cf. the list on the website describing our research: http://geoscimo.univ-tlse2.fr/results-and-analyses/).

BIBLIOGRAPHY

Anselin, L., 1995. "Local Indicators of Spatial Association—LISA", *Geographical Analysis*, Vol.27, No.2, 93–115.

Bathelt, H., Malmberg, A., & Maskell, P., 2004, "Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation", *Progress in Human Geography*, Vol.28, No.1, 31–56.

Catini, R., Karamshuk, D., Penner, O. & Riccaboni, M., 2015, "Identifying geographic clusters: A network analytic approach", *Research Policy* 44, 1749-1762.

Comin, M.-N., 2009, *Réseaux de villes et réseaux d'innovation en Europe: structuration du système des villes européennes par les réseaux de recherché sur les technologies convergentes* (Thèse de géographie sous la direction de Denise Pumain). Université Paris I-Panthéon-Sorbonne, Paris.

Cuyala, S., 2013, "La diffusion de la géographie théorique et quantitative européenne francophone d'après les réseaux de communications aux colloques européens (1978-2011)", *Cybergeo: European Journal of Geography*, No.657

Csomós, G., 2018, "A spatial scientometric analysis of the publication output of cities worldwide", *Journal of Informetrics*, Vol.12, No2, 547–566.

Eckert, D., Baron, M., & Jégou, L., 2013, "Les villes et la science : apports de la spatialisation des données bibliométriques mondiales", *M@ppemonde*, Vol.110, No.2. Retrieved from http://mappemonde.mgm.fr/num38/articles/art13201.pdf

ESA 2005, 2009, Iona GlobCover project, led by MEDIAS-France/POSTEL, http://due.esrin.esa.int/page_globcover.php

Frenken, K., Hardeman, S., & Hoekman, J., 2009, "Spatial scientometrics: Towards a cumulative research program", *Journal of Informetrics*, Vol.3, No.3, 222–232.

Guérois, M., Bretagnolle, A., Mathian, H., & Pavard, A., 2014, Functional Urban Areas (FUA) and European harmonization. *A feasibility study from the comparison of two approaches: commuting flows and accessibility isochrones (Technical Report, Espon 2013 Database)* (p. 35). Paris: Union Européenne.

Jégou, L., 2014, *Toward spatially referenced academic data at global scale: the full geocoding of Wos-Datasets, methods and results*. Presented at the 2nd Geography of Innovation International Conference, Utrecht.

Livingstone, D., 2003, *Putting science in its place: Geographies of scientific knowledge,* Chicago: The University of Chicago Press.

Maisonobe, M., 2015, *Étudier la géographie des activités et des collectifs scientifiques dans le monde. De la croissance du système de production contemporain aux dynamiques d'une spécialité : la réparation de l'ADN*. (Dir. Denis Eckert). PhD Thesis, Université de Toulouse Jean-Jaurès, Toulouse. September 17. https://tel.archives-ouvertes.fr/tel-01235015/

Masselot, A., 2016, *Where are the scientific publications coming from? Geolocalizing Medline citations*. January 12. Retrieved from https://blog.octo.com/en/geo-localizing-medline-citations/

Matthiessen, C. W., Schwarz, A. W., & Find, S., 2002, "The Top-level Global Research System, 1997-99: Centres, Networks and Nodality. An Analysis Based on Bibliometric Indicators", *Urban Studies*, Vol.39, No5–6, 903–927.

Matthiessen, C. W., Schwarz, A. W., & Find, S., 2010, "World Cities of Scientific Knowledge: Systems, Networks and Potential Dynamics. An Analysis Based on Bibliometric Indicators", *Urban Studies*, Vol.9, No.47, 1879–1897.

Moriconi-Ebrard, F., 1991, "Les 100 plus grandes villes du monde", *Économie et Statistique*, No.254, 7–18.

Pan, R. K., Kaski, K., & Fortunato, S., 2012, "World citation and collaboration networks: uncovering the role of geography in science", *Scientific Reports*, 2.

Pumain, D., Swerts, E., Cottineau, C., Vacchiani-Marcuzzo, C., Ignazzi, A., Bretagnolle, A., Delisle, F., Cura, R., Lizzi, L., & Baffi, S., 2015, "Multilevel comparison of large urban systems", *Cybergeo: Revue européenne de géographie*, No.706, https://cybergeo.revues.org/26730

SEDAC – NASA, Columbia University, *United Nations Food and Agriculture Programme* - FAO, & Centro Internacional de Agricultura Tropical - CIAT. (2005). Gridded Population of the World, Version 3 (GPWv3): Population Count Grid. NASA Socioeconomic Data and Applications Center (SEDAC)

Schneider, A., Friedl, M., Potere, D., 2009, "A new map of global urban extent from MODIS data", *Environmental Research Letters*, Vol.4, article 044003.

Schneider, A., et al., 2010, "Monitoring urban areas globally using MODIS 500m data: New methods and datasets based on urban ecoregions", *Remote Sensing of Environment*, in review.

Waltman, L., Tijssen, R. J. W., & Eck, N. J. van., 2011, "Globalisation of science in kilometres", *Journal of Informetrics*, Vol.5, No.4, 574–582.

NOTES

1. http://www.geonames.org/
2. http://nominatim.openstreetmap.org/
3. For the last batch of data (2012 to 2014), about 6000 publications per year could not be geocoded, thus accounting for 0.33 % of the total.
4. http://www.ciesin.org/
5. At the time we are writing this article, a more recent release is available. It is the 10th revision of the Gridded Population of the World project by NASA SEDAC which is based on the world censuses from 2005 to 2014: http://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-rev10. In future years, it will be necessary to address the question of updating our agglomerations' perimeters hosting scientific activities using this updated dataset on population density.
6. https://www.mapdevelopers.com/batch_geocode_tool.php

ABSTRACTS

In this data paper, we provide a global dataset of urban perimeters that has been built in order to study the evolving geography of scientific activity at city level across the world. The method developed for building these agglomerations associates the distribution of population density and the distance between geolocalized scientific publications (issued between 1999 and 2014), whose author's addresses have been systematically geocoded. The location of scientific production is obtained by processing bibliographic data retrieved in the Web of Science Core Collection (Clarivate Analytics). In the first part of the article, we detail the geocoding stage of our methodology. Next, we discuss the importance of delineating homogeneous urban perimeters to study the world geography of science production and we detail the methodology used to build these perimeters (the clustering stage). After discussing the extent to which our work can be re-used and enriched, we propose to use this dataset in order to capture the worldwide distribution of Cybergeo publications at the city level between 2015 and 2017.

Dans ce *data paper*, nous mettons à disposition un nouveau jeu de données de périmètres urbains que nous avons construit afin d'étudier la géographie mondiale des activités scientifiques et son évolution récente. Ces délimitations tiennent compte à la fois de la répartition de la densité de population et de la distance kilométrique entre les adresses d'où ont été signées les publications scientifiques parues entre 1999 et 2014. La géographie de la production scientifique est obtenue à partir des données bibliographiques indexées dans le *Web of Science Core Collection* (Clarivate Analytics). Dans la première partie de l'article, nous détaillons l'étape du repérage ou géocodage des localités « publiantes ». Ensuite, nous discutons de la nécessité de délimiter des périmètres urbains à l'aide de critères homogènes pour pouvoir étudier la géographie mondiale de la science et nous décrivons la méthodologie mise en place pour construire ces périmètres (étape du passage aux agglomérations). Après avoir discuté des conditions auxquelles ces périmètres pourront être réutilisés et enrichis, nous proposons une exploitation de ces périmètres au cas de la répartition mondiale des publications parues dans la revue Cybergéo entre 2015 et 2017.

INDEX

Mots-clés: agglomération, agrégation, bibliométrie, données ouvertes, monde, science
Keywords: agglomeration, aggregation, bibliometrics, open data, science, world

AUTHORS

MARION MAISONOBE

Université Paris-Est, UMR LATTS, ENPC, UPEM, CNRS, 6 Avenue Blaise Pascal, 77455, Marne-la-Vallée, France
Post-doctorante LABEX Futurs Urbains
marion.maisonobe@enpc.fr; marion.maisonobe@univ-tlse2.fr

LAURENT JÉGOU

Université de Toulouse – Jean Jaurès, UMR LISST, CNRS, 5 allées Antonio Machado, 31058, Toulouse, France
Maitre de Conférence en cartographie
jegou@univ-tlse2.fr

DENIS ECKERT

Centre Marc Bloch e.V., CNRS, Friedrichstraße 191, 10017 Berlin, Germany
Directeur de Recherche
eckert@cmb.hu-berlin.de